Metodología de análisis tiempo-frecuencia para la evaluación automática de la voz de pacientes con enfermedad de Parkinson

Tatiana Villa Cañas

Universidad de Antioquia
Facultad de Ingeniería, Departamento de Electrónica y Telecomunicaciones
Medellín, Colombia
2015

Metodología de análisis tiempo-frecuencia para la evaluación automática de la voz de pacientes con enfermedad de Parkinson

Tatiana Villa Cañas

Trabajo de investigación para optar al título de:
Magíster en Ingeniería de Telecomunicaciones

Asesor:
PhD. Julián David Arias Londoño

Lineas de Investigación:
Procesamiento de señales de voz
Análisis de Datos, Aprendizaje de maquina e Inteligencia Computacional

Grupo de Investigación:
Grupo de Investigación en Simulación y Comportamiento de Sistemas (SICOSIS)

Universidad de Antioquia
Facultad de Ingeniería, Departamento de Electrónica y Telecomunicaciones
Medellín, Colombia
2015

Agradecimientos

Quiero dar un agradecimiento especial a mí mama, que con su cariño y apoyo incondicional, pude lograr el objetivo de estudiar una maestría y culminarla pese a todos los obstáculos. Además, por siempre comprenderme y no dejarme derrotar frente a las dificultades, no sólo en el estudio sino a nivel personal. Igualmente, quiero darle las gracias a las personas más importantes de vida como son la familia, los amigos, y compañeros de estudio y trabajo, que han comprendido mis ausencias y motivado para seguir adelante, han sido incondicionales tanto en los buenos como malos momentos.

También quiero agradecerle a mi asesor, el profesor Ph.D Julián David Arias Londoño, quien me ha orientado y brindado el conocimiento necesario para lograr los objetivos propuestos a lo largo de esta maestría. A los profesores Ph.D Jesús Francisco Vargas y Ph.D Juan Rafael Orozco por brindarme su valioso conocimiento y su gran experiencia en el campo de la investigación.

Este trabajo se encuentra enmarcado dentro del proyecto "Análisis de la capacidad discriminante de características de fonación, articulación y prosodia en pacientes con enfermedad de Parkinson en etapa preclínica y avanzada, para el desarrollo de herramientas computacionales de apoyo al diagnóstico y tratamiento de la enfermedad'", proyecto No. 111556933858 financiado por COLCIENCIAS. Se da un especial agradecimiento al proyecto por la financiación de este trabajo de maestría, al financiador COLCIENCIAS, a la Universidad de Antioquia y a la Universidad Nacional.

Resumen

En los últimos años un gran número de trabajos se han centrado en la detección automática de patologías de la voz, con el fin de realizar evaluaciones objetivas de la voz de una manera no invasiva. En los casos en que la patología afecta principalmente a los patrones vibratorios de los pliegues vocales, los análisis que se llevan a cabo típicamente son la pronunciación de las vocales sostenidas. Sin embargo, si la señal de voz pierde parte de su comportamiento cuasi-periódico debido a la presencia de la patología, y se afectan no sólo los procesos de fonación, sino también las dimensiones de articulación y prosodia, el análisis de señales de voz debe incluir ejercicios adicionales (palabras, frases y/o monólogos), ya que las vocales sostenidas por sí solas no son suficientes para evaluar las habilidades de calidad de voz y comunicación de un paciente, puesto que no incorporan aspectos dinámicos de voz continua (por ejemplo, la coarticulación, las características suprasegmentales, los efectos de inicio y compensación de la señal de voz, etc.) [1].

En general las patologías que afectan el habla pueden ser de tipo orgánico, funcional o neurológico. Dentro de las patologías de tipo neurológico, el desorden de Parkinson es actualmente una de las de mayor interés en el campo del procesamiento de voz patológica, debido a que se considera un desorden neurológico irreversible que afecta de forma exclusiva las áreas cerebrales encargadas del control motor del organismo, viéndose afectada la facultad del habla de múltiples maneras: la reducción en el nivel de la presión sonora, la inestabilidad en la fonación, la disminución de la entonación, el incremento en el ruido, y variabilidad durante la fonación ciclo a ciclo, produciendo múltiples cambios e introduciendo componentes de baja frecuencia [2]. El análisis automático de la voz de pacientes con Parkinson ha tomado reciente interés, con el objetivo de encontrar marcadores que ayuden al diagnóstico de la enfermedad, y para desarrollar herramientas que permitan medir objetivamente el grado de afectación de los pacientes y ayuden a evaluar terapias de lenguaje que propendan por mantener la capacidad comunicativa de los pacientes, por más tiempo.

Uno de los problemas principales que surgen en este escenario es el hecho de que varias de las medidas acústicas más utilizadas, se basan en el análisis en tiempo corto, bajo supuestos de estacionariedad que no son apropiados para las señales de voz continua, debido a las características suprasegmentales (variaciones de *pitch*, el ritmo, la entonación, entre otras) [1]. En general, las técnicas basadas en el análisis de muestras de voz continua se basan en algún procedimiento de segmentación para identificar los periodos sonoros y sordos, debido al hecho de que las medidas de periodicidad y regularidad (por ejemplo, relación armónica a ruido, primer pico cepstral y amplitud del tono) son válidas sólo en las regiones sonoras [3]. Sin embargo, la segmentación sigue presentando problemas para la determinación del punto de inicio y final de los segmentos (sonoros y sordos), a causa del enlace natural que se produce entre un fonema y el siguiente, fenómeno conocido como coarticulación.

Un enfoque alternativo en la manera de extraer características a partir de señales de voz continua, es utilizar directamente una técnica no estacionaria, evitando así la necesidad de usar un procedimiento de segmentación y permitiendo el análisis de segmentos más largos, posibilitando la caracterización de fenómenos que afectan varias unidades fonéticas y permitiendo la detección de cambios en baja frecuencia. Las técnicas de análisis tiempo-frecuencia, brindan la capacidad de operar en condiciones no estacionarias y proporcionan una perspectiva mejorada de las características temporales y espectrales de la señal, lo cual hace que sea un método adecuado para el análisis de señales cuyas características espectrales varían en el tiempo, como las señales biomédicas: fonocardiogramas, electrocardiogramas y señales de voz; además de otras señales de tipo no biológico como las ondas sísmicas, las vibraciones de motores, la propagación de la onda electromagnética, los parámetros atmosféricos, entre otras [4].

En el presente trabajo se desarrolla una metodología basada en el análisis tiempo-frecuencia y en la extracción de características dinámicas de las señales de voz de pacientes con enfermedad de Parkinson, que permite analizar el contenido espectral en registros de voz continua, y usar dicha información para la construcción de sistemas automáticos de detección de patologías, basados en técnicas de reconocimiento de patrones.

Abstract

In the last years a large number of works have been focused on the automatic detection of voice pathologies, in order to be able of performing voice evaluation in a non-invasive framework. Whenever the pathology affects mainly the vibratory patterns of the vocal folds, the analyses are typically performed using pronunciation of sustained vowels. However, if the speech signal loses some of its quasi-periodic behavior due to the presence of pathology, and are affected not only phonatory processes, but also the dimensions of articulation and prosody, the analysis of speech signals should include additional exercise (words phrases and/or monologues), since sustained vowels alone are not sufficient for assessing voice quality and communication skills, because do not incorporate dynamic aspects of continuous speech (e.g. coarticulations, suprasegmental characteristics, onset and offset effects etc.) [1].

The diseases that affect speech can be organic, functional or neurological. Among the neurological diseases, Parkinson's disorder is currently one of the most interesting in the area of pathological voice processing, because it is considered an irreversible neurological disorder that affects exclusively responsible motor brain areas controlling body. The faculty of speech is affected in multiple ways: reduced loudness, monopitch, monoloudness, reduced stress, breathy, hoarse voice quality, and imprecise articulation, producing multiple changes and introducing low frequency components [2]. Automatic speech analysis of patients with Parkinson has taken recent interest, with the aim of finding markers that aid in the diagnosis of disease, and to develop tools to objectively measure the degree of patients' affectation, and help assess language therapies that tend to keep the communication skills of patients for longer.

One of the main problem arising in this scenario is the fact that several of the most used acoustic measures are based on short-time analysis, under assumptions of stationarity which are not appropriate for continuous speech signals due to suprasegmental features [1]. In general, techniques based on the analysis of continuous speech samples rely on some segmentation procedure to identify the voiced and unvoiced periods, due to the fact that measures that quantify periodicity and regularity (e.g. Harmonic-to-Noise ratio, cepstral peak prominence, and pitch amplitude) are valid only in the voiced regions [3]. However, the segmentation still presents problems for determining the start and the end of the segments (voiced and unvoiced), because of the natural binding that occurs between one phoneme and the next, phenomenon known as coarticulation.

A different way to extract features from continuous speech signals is to use a non-stationary technique directly, thereby avoiding the need for any segmentation procedures and allowing the analysis of longer frames, making feasible the characterization of phenomena which affect several phonetic units and allowing the detection of low frequency changes. The techniques of time-frequency analysis, provide the ability to operate in non-stationary conditions and offer improved perspective of the temporal and spectral characteristics of

the signal, which makes them suitable methods for the analysis of signals whose spectral characteristics vary over time, such as biomedical signals: phonocardiograms, electrocardiograms and voice signals; as well as others non-biological signals such as seismic waves, the vibrations of engines, the propagation of the electromagnetic wave, atmospheric parameters, and other ones [4].

In this work a methodology based on the time-frequency analysis and extraction of dynamic characteristics of voice signals of patients with Parkinson's disease is developed, in order to analyze the spectral content in continuous speech records, and use such information in the development of systems for the automatic detection of pathological voices.

Índice general

Índice de figuras

Índice de tablas

Lista de acrónimos

AE	Entropía aproximada (*Approximate Entropy*)
APQ	Cociente de perturbación de amplitud (*Amplitude Perturbation Quotient*)
AUC	Área bajo la curva (*area under the curve*)
BJD	Distribución de Born-Jordan (*Born-Jordan Distribution*)
CD	Dimensión de correlación (*Correlation Dimension*)
CPL	(*Component Linking*)
CSKD	Distribución Cone-Shaped Kernel (*Cone-Shaped Kernel Distribution*)
CWD	Distribución de Choi-Williams (*Choi-Williams Distribution*)
CWT	Transformada Wavelet continua (*Continuous wavelet transform*)
CWT	Transformada Wavelet discreta (*Discrete wavelet transform*)
EBW	Anchos de banda equivalentes (*Equivalent Bandwidth*)
ECG	Electrocardiograma
ECoG	Electrocorticográfico
EEG	Electroencefalograma
EM	Máxima-Esperanza (*Expectation Maximization*)
EMD	Descomposición en modo empírico (*Empirical Mode Decomposition*)
EP	Enfermedad de Parkinson
FCR	Formant Centralization Ratio
FDR	Coeficiente discriminante de Fisher (*Fishers' Discriminant Ratio*)

FFT Transformada rápida de Fourier (*Fast Fourier Transform*)

FI Frecuencia instantánea

GMM Modelo de mezclas gaussianas (*Gaussian Mixture Models*)

GMM-UBM Modelo de mezclas gaussianas con modelos universales (*GMM-Universal Background Model*)

GNE Relación de excitación glotal a ruido (*Glottal to Noise Excitation Ratio*)

HMM Modelo oculto de Markov (*Hidden Markov Models*)

HNR Relación de armónicos a ruido (*Harmonics to Noise Ratio*)

HOS Estadísticas de alto orden (*High Order Statistics*)

HOSVD Descomposición en valores singulares de alto Orden (*Higher Order Singular Value Decomposition*)

HRV Variabilidad de la frecuencia cardiaca

kNN k-vecinos más cercanos (k-*Nearest Neighbor*)

LDA Análisis discriminante lineal (*Linear Discriminant Analysis*)

LFCC Coeficientes cepstrales de frecuencia lineal (*Linear Frequency Cepstral Coefficients*)

LFCC Coeficientes de predicción lineal (*Linear Prediction Coefficients*)

LLE Máximo exponente de Lyapunov (*Largest Lyapunov Exponent*)

LP patologías laríngeas (*Laryngeal pathologies*)

LPH Labio y/o paladar hendido

LVQ Aprendizaje de cuantificación vectorial (*Learning Vector Quantization*)

MAP (*Maximum A Posteriori*)

MBD Distribución B-modificada (*Modified-B distributions*)

MEG Magnetoencefalografía

MFCC Coeficientes cesptrales en la escala de frecuencia Mel (*Mel Frequency Cepstrum Coefficients*)

MHD Distribución de Margenau-Hill (*Margenau-Hill Distribution*)

MKL Aprendizaje con múltiples kernels (*Multiple Kernel Learning*)

ML Estimación de la máxima verosimilitud (*Maximum Likelihood*)

MLP Perceptron multicapa (*Multilayer Perceptron*)

MS Espectros de modulación(*Modulation Spectra*)

NNE Energía de ruido normalizada (*Normalized Noise Energy*)

NSR Formant Centralization Ratio

PCA Análisis de componentes principales (*Principal Component Analysis*)

PPE Pitch Period Entropy

PPQ Cociente de perturbación de pitch (*Pitch Perturbation Quotient*)

PR Pause Ratio

PWVD Distribución Pseudo Wigner-Ville (*Pseudo Wigner-Ville Distribution*)

RAP Perturbación relativa promedio (*Relative Average Perturbation*)

RD Distribución de Rihaczek (*Rihaczek Distribution*)

RID Distribución de interferencias reducidas (*Reduced Interferences Distribution*)

ROC Característica de Operación del Receptor (*Receiver Operating Characteristic*)

SFR Porcentaje de planitud espectral (*Spectral Flatness Ratio*)

SJE Spectral Jitter Estimator

SNR Relación señal a ruido (*Signal to Noise Ratio*)

SPWVD Distribución Pseudo Wigner-Ville suavizada (*Smoothed Pseudo Wigner-Ville Distribution*)

STFT Transformada de Fourier en tiempo corto (*Short Time Fourier Transform*)

SVD Descomposición en valores singulares (*Higher Order Singular Value Decomposition*)

SVM Máquina de soporte vectorial (*Support Vector Machine*)

TEO Operador de energía de Teager (*Teager Energy Operator*)

TF Tiempo-Frecuencia

TFD Distribuciones tiempo-frecuencia (*Time-Frequency Distribution*)

TNI Índice de ruido turbulento (*Turbulent Noise Index*)

UBM Modelo universal (*Universal Background Model*)

UPDRS (*Unified Parkinson Rating Scale*)

VAI Índice de articulación vocal (*Vowel Articulation Index*)

VSA Area del espacio vocal (*Vowel Space Area*)

WD Distribución Wigner (*Wigner Distribution*)

WPT Transformada Wavelet de paquetes (*Wavelet Packet Transform*)

WT Transformada Wavelet (*Wavelet Transform*)

WVD Distribución Wigner-Ville (*Wigner-Ville Distribution*)

Introducción

La enfermedad de Parkinson (EP) es una enfermedad neurodegenerativa caracterizada por la pérdida progresiva de las células dopaminérgicas en la sustancia negra del mesencéfalo [6], afectando diferentes funciones realizadas en los ganglios basales, que incluyen el control de los movimientos voluntarios, el procedimiento de aprendizaje, y los movimientos de la mandíbula y los ojos [7]. La EP es la segunda enfermedad neurodegenerativa más frecuente en el mundo después del Alzheimer [8]. En el caso de Colombia, la prevalencia de la EP se encuentra alrededor de 172,4 casos por cada 100.000 habitantes [9].

Debido a la influencia en la reducción del control motor de los pacientes con EP, el 90 % desarrollan problemas en la voz y el habla, siendo los problemas de fonación los primeros en manifestarse. Los desordenes en la voz de las personas con EP han sido asociados tradicionalmente a problemas motores, tales como: rigidez, tremor (temblor), bradicinesia (lentitud de los movimientos) e inestabilidad de la postura. Perceptualmente, la voz de las personas con EP se caracteriza por tener una intensidad reducida o monótona, áspera, con articulación imprecisa, no fluida, entre otros problemas asociados [10].

Estudios recientes [11–13] sugieren que la EP puede afectar a diferentes dimensiones de la producción del habla, como son la respiración, fonación, articulación y prosodia; lo que ha logrado que en los últimos años la comunidad científica haya puesto un particular interés en estudiar la influencia de la EP en la voz, tanto en etapa pre-clínica como en etapas posteriores, debido a que los problemas en la voz de pacientes con EP, son en algunos casos uno de los primeros síntomas de la enfermedad y pueden avanzar incluso hasta el punto de hacer que el paciente pierda completamente la capacidad de comunicación mediante el habla; a pesar de esto, sólo entre el 3 % y el 4 % de los pacientes reciben tratamiento o terapia del habla [10].

En este sentido, se ha demostrado que la mejora en la calidad del habla y la voz de las personas con EP, no son exclusivas de los tratamientos farmacológicos, sino que también se requiere de una terapia exhaustiva del habla [14]. El hallazgo de identificadores tempranos de esta enfermedad tendría efectos positivos para los afectados, tanto desde el punto de vista social como económico, así como para sus familias y la sociedad en general. Siendo de especial interés clínico y social desarrollar metodologías para el seguimiento de problemas específicos en la voz de las personas con EP.

Por otra parte, debido a los desordenes en la comunicación que padecen las personas con EP, uno de los efectos más comunes en todas las dimensiones de la producción del habla, es la introducción de componentes de baja frecuencia conocidos como tremor [2, 15]. El tremor parkinsoniano es ante todo la presencia de temblor en estado de reposo con una frecuencia que puede variar en el rango de 4 a 9 Hz [16] que, si se detecta en una etapa temprana de la enfermedad, podría convertirse en un bio-marcador relevante para el diagnóstico y evaluación del tratamiento.

Tradicionalmente, la detección automática de patologías de la voz está basada en métodos de análisis de voz en tiempo corto, mediante pequeños tramos de voz llamados ventanas [17], bajo la suposición de que en dichos tramos la señal se comporta como un proceso estacionario [18, 19]. En la mayoría de trabajos enfocados al análisis de voces patológicas, gran parte de los esfuerzos se centran en la evaluación de fonaciones sostenidas [17, 20], debido a que las patologías más comunes son de origen orgánico y/o funcional, y afectan directamente el movimiento de los pliegues vocales durante los periodos fonatorios [2].

Las señales producidas a partir de fonaciones de vocales sostenidas tienen un comportamiento cuasi-periódico y durante segmentos cortos puede asumirse que son observaciones de procesos estacionarios [18]. Sin embargo, cuando la afectación de la voz no se da a nivel de las cuerdas vocales, sino de movimientos articulatorios o a nivel prosódico, el análisis de las señales de voz debe incluir ejercicios adicionales como la pronunciación de palabras, frases y/o monólogos. Esto se debe a que las vocales sostenidas por si solas no son adecuadas para la evaluación de las habilidades de calidad de la voz y comunicación, y también porque no incorporan aspectos dinámicos de la voz continua (las variaciones de periodos de *pitch*, el ritmo, la entonación y otras características suprasegmentales) [1]. Este es el caso de la evaluación de la voz de pacientes con EP, la cual se ve afectada de múltiples formas: reducción en el nivel de la presión sonora, reducción en la estabilidad del *pitch*, inestabilidad en la fonación, incremento en el ruido y variabilidad durante la fonación ciclo a ciclo, produciendo múltiples cambios e introduciendo componentes de baja frecuencia, conocidos como tremor, que por el simple hecho de la resolución en el tiempo, no pueden ser analizados con técnicas que requieran definir ventanas de análisis muy pequeñas [2].

En este escenario los métodos convencionales de análisis basados en suposiciones de estacionariedad pierden validez o deben limitarse al análisis de los periodos fonatorios, conocidos como sonoros (en inglés, voiced), dejando de lado los segmentos sordos (en inglés, unvoiced) o perdiendo la información prosódica contenida en todo el registro, ya que la mezcla de los periodos sonoros, sordos y de silencio conduce a condiciones no estacionarias [3]. Es importante resaltar que en el caso particular de los pacientes con EP, la evolución de la enfermedad repercute de manera significativa en su capacidad comunicativa, en la cual intervienen no sólo los procesos fonatorios, sino también las dimensiones de articulación y prosodia [2].

El análisis a partir de la segmentación sonora-sorda ha permitido en muchos casos obtener resultados satisfactorios [21, 22], sin embargo sigue presentando inconvenientes debido a que la determinación del punto de inicio y final de los segmentos sonoros (y sordos) no es fácil de encontrar debido a un fenómeno que se presenta en la producción de voz, conocido como coarticulación y que define el enlace natural que se produce entre un fonema y el siguiente.

Una forma alternativa de realizar el análisis de registros de voz continua, es utilizar técnicas de análisis tiempo-frecuencia (TF), las cuales permiten, en muchos casos, caracterizar el proceso sin hacer suposiciones de estacionariedad o al menos permiten analizar ventanas de la señal mucho mayores que las convencionales (que oscilan entre los 20 y 40 ms) [3]. El análisis TF tiene como objetivo determinar la concentración de energía en un espacio de dos dimensiones, tiempo y frecuencia, brindando una perspectiva mejorada de las características temporales y espectrales de una señal, lo cual hace que sea un método adecuado para el análisis de señales cuyas características espectrales varían en el tiempo, conocidas como señales no estacionarias. Este tipo de representaciones son de gran utilidad en escenarios donde se pretende extraer información específica de una señal, la cual no puede ser obtenida fácilmente en el dominio del tiempo.

Este hecho beneficia la posibilidad de caracterizar los cambios espectrales introducidos en la señal de voz a causa de la EP, entre ellos el tremor [2, 15], ya que al ser una perturbación en baja frecuencia no puede ser detectada mediante segmentos cortos, como los obtenidos al realizar segmentación en periodos sonoros y sordos. Analizar el efecto del temblor de baja frecuencia, implica analizar segmentos que oscilan entre los 100 y los 500 ms. De manera que, el uso de técnicas de caracterización no estacionarias permite la extracción de información en segmentos apropiados para una evaluación de capacidades comunicativas en un sentido más global, siendo de gran interés para el desarrollo de herramientas de análisis que pueden ser usadas en este contexto y que además permiten el desarrollo de sistemas computacionales de apoyo al diagnóstico y tratamiento fonoaudiológico en este tipo de pacientes.

Con este trabajo se busca explorar escenarios basados en procesamiento digital de señales de voz en pacientes con EP, con el fin de analizar el contenido espectral, específicamente, en baja frecuencia, sin tener en cuenta supuestos de estacionariedad, mediante el estudio de distribuciones TF, permitiendo establecer cuáles técnicas o características pueden ofrecer información relevante en la detección automática de la EP. La metodología desarrollada comprende la evaluación teórica y experimental de diferentes métodos de análisis TF y la extracción de características dinámicas sobre los espectros obtenidos de cada una de las distribuciones, además de la evaluación de métodos de reducción de dimensión en los casos en los que sea necesario. Teniendo en cuenta que el objetivo final de este análisis es determinar la capacidad discriminante de las técnicas TF para su uso en sistemas automáticos de apoyo al diagnóstico, la información obtenida a partir de los métodos y técnicas evaluadas, se usa como parte de un sistema de clasificación automática, teniendo en cuenta técnicas de reconocimiento de patrones y aprendizaje de máquina aplicados en sistemas de procesamiento

de voz. Finalmente, con el ánimo de explotar al máximo la capacidad discriminante de las diferentes representaciones TF, el trabajo explora diferentes metodologías y/o estrategias para combinar la información extraída.

Objetivos

General:
Proponer una metodología para la caracterización de señales a partir de representaciones tiempo-frecuencia, que permitan evaluar los cambios en la riqueza espectral de las señales de voz en la evaluación automática de habla continua de pacientes con enfermedad de Parkinson.

Específicos:

1. Analizar e identificar las diferentes distribuciones basadas en el análisis tiempo-frecuencia que permitan representar correctamente señales no estacionarias y variantes en el tiempo.
2. Determinar a partir de las representaciones tiempo-frecuencia las medidas que permitan capturar los cambios en la dinámica de la señal debidos a la presencia de la enfermedad de Parkinson.
3. Evaluar la pertinencia en el uso de técnicas de selección y/o extracción de características para reducir la dimensionalidad del espacio de representación de las medidas derivadas del análisis tiempo frecuencia, para mejorar la capacidad discriminante y el costo computacional del sistema.
4. Validar la metodología desarrollada utilizando un conjunto de muestras apropiado y a partir de técnicas de clasificación aceptadas en el campo del procesamiento de voz.

Capítulo 1

Estado del arte

En los últimos años, gracias a los adelantos tecnológicos, la comunidad científica ha desarrollado una gran cantidad de trabajos en el campo de procesamiento digital de señales de voz, en particular relacionados con el estudio y la detección automática de patologías del habla. Los estudios sugieren, que en el caso de la enfermedad de Parkinson (EP) los problemas en la voz son uno de los primeros síntomas en aparecer, afectando las diferentes dimensiones del habla (respiración, fonación, articulación y prosodia); lo que ha creado un interés especial en estudiar la influencia de esta enfermedad en la voz, tanto en etapa preclínica como en etapas posteriores, para la implementación de herramientas automáticas como apoyo al diagnóstico médico, debido a su objetividad y naturaleza no invasiva.

En este contexto, inicialmente se hace una revisión de los trabajos de investigación más representativos en reconocimiento automático de patologías de la voz, en cuanto a las técnicas clásicas usadas en la etapa de caracterización y clasificación, así como el conjunto de datos utilizados; siendo el foco principal la EP.

Para finalizar, son explorados los trabajos orientados al uso de técnicas tiempo-frecuencia (TF) que determinan la concentración de energía en un espacio de dos dimensiones, tiempo y frecuencia, brindando una perspectiva mejorada de las características temporales y espectrales de la señal. Las técnicas TF presentan especial atención a estudios dedicados a la clasificación automática de patologías, y principalmente aquellos enfocados a la detección automática de patologías del habla.

1.1. Análisis de voz patológica

Clásicamente, la detección automática de patologías de la voz se ha realizado mediante el análisis de vocales sostenidas y habla continua [1, 17]. Los estudios encontrados consideran varios fenómenos físicos que intervienen en el proceso de producción de la voz, y son susceptibles a ser afectados por la presencia de la patología, entre ellos se encuentran: la periodicidad o estabilidad de la voz, la presencia de ruido, la riqueza espectral y el comportamiento no lineal [20].

En la literatura, en cuanto al análisis de vocales sostenidas, se han encontrado diferentes parámetros acústicos. Considerando que la mayoría de voces manifiestan cierto grado de

ruido en presencia de patología, las medidas de ruido son unos de los principales parámetros acústicos con demostrada fiabilidad para detectar la presencia de trastornos de la voz. En [23], los autores utilizaron diferentes parámetros de ruido incluyendo la relación señal a ruido (*SNR − Signal to Noise Ratio*), la relación de armónicos a ruido (*HNR - Harmonics to Noise Ratio*), la energía de ruido normalizada (*NNE − Normalized Noise Energy*), un parámetro relacionado con la amplitud de paso y uno denominado porcentaje de planitud espectral (*SFR − Spectral Flatness Ratio*), para detectar voces patológicas. Los experimentos se realizaron a través de un subconjunto de la base de datos desarrollada por el laboratorio *Massachusetts Eye and Ear Infarmary* (*MEEI* [24]), con 175 voces patológicas y 53 voces sanas. Los resultados fueron obtenidos por medio de umbrales, en donde el resultado más bajo se obtuvo para NNE con 63.2 % y el mejor para el SFR con 96.5 %. Así mismo, los autores en [25] utilizan varios de los parámetros anteriores junto con una nueva medida denominada índice de ruido turbulento (*TNI − Turbulent Noise Index*), para el reconocimiento automático de voces patológicas de la base de datos MEEI, alcanzando una precisión de 96.1 % por el método de clasificación k-vecinos más cercanos (*kNN − k-Nearest Neighbor*). En [26] es estimado el HNR en cuatro bandas de frecuencias diferentes, los experimentos se llevaron a cabo con 53 grabaciones de personas con voces sanas y 163 patológicas de la base de datos MEEI, alcanzando una tasa de acierto de 94.3 % con el método kNN. Además, los autores en [27] evalúan la capacidad de la relación de excitación glotal a ruido (*GNE − Glottal to Noise Excitation Ratio*) para la detección de trastornos de la voz, utilizado un conjunto de 226 voces (53 sanas y 173 patológicas) de la misma base de datos. Para evaluar el parámetro, se analizó el efecto del ancho de banda de las envolventes de Hilbert y el desplazamiento de frecuencia, concluyendo que una buena discriminación se obtiene con un ancho de banda de $1000\,Hz$ y un desplazamiento de frecuencia de $300\,Hz$. Los resultados alcanzan una precisión de clasificación de hasta el 89.9 %, comparable con otras medidas clásicas de ruido en la literatura, como NNE y HNR que han alcanzado tasas de 89.7 % y 84.5 %, respectivamente.

Con relación al grado de periodicidad o estabilidad de la voz, se encuentran dos parámetros muy utilizados, la perturbación de la frecuencia (*Jitter*) [28] y la perturbación de amplitud (*Shimmer*) [29]. En [30] se utilizaron los parámetros *jitter* y *shimmer*, así como dos medidas de ruido para la detección remota de patologías de la voz a partir de grabaciones realizadas en un ambiente controlado y de conversaciones telefónicas. Los experimentos se realizaron mediante el uso de la base de datos MEEI, que contiene 573 registros de pacientes de control y 58 patológicos de la fonación sostenida de la vocal /a/. Los resultados fueron obtenidos por medio de un clasificador lineal conocido como LDA (*Linear Discriminant Analysis*) y muestran que las fonaciones sostenidas grabadas en un ambiente controlado pueden ser reconocidas como sanas o patológicas con una precisión de 89.1 %, mientras que las fonaciones que fueron obtenidas a través de conversaciones telefónicas alcanzan una precisión sólo de 74.2 %. Los autores concluyen que el principal inconveniente de los parámetros de perturbación, es la dependencia de una correcta estimación del periodo del *pitch* (frecuencia fundamental) y, para muchas voces patológicas, tal estimación es bastante difícil. También en [31], son usadas las dos medidas de perturbación antes mencionadas junto con una medida de ruido en voces patológicas, a través de un conjunto de registros de voz previamente identificados como señales periódicas o cuasi-periódicas. De donde se

infiere, que ninguna de las medidas de perturbación fue capaz de diferenciar entre personas disfónicas y no disfónicas. El conjunto de datos utilizado fue compuesto por 112 registros de voces disfónicas y 41 registros de voces sanas, realizando el análisis a partir de la evaluación de una prueba estadística, para establecer los parámetros que presentan diferencias reales entre las clases de control y patológicas, en lugar de un procedimiento basado en métodos de reconocimiento de patrones. Similarmente en [32], se investiga la capacidad que tienen diferentes tipos de características para la identificación del edema de Reinke, mediante el análisis de las vocales /a/, /e/ y /i/. Para los experimentos fueron usados, el ancho de banda del primer pico, el *pitch* y el *jitter*; utilizando un simple árbol de decisión, se obtuvo que el 94 % de todos los sujetos de la base de datos son diagnosticados correctamente.

Por otra parte, también ha sido estudiada la riqueza espectral de las voces patológicas, ya que se encuentra relacionada con las lesiones presentes en los pliegues vocales, y cuando se produce un cierre incompleto de la glotis da lugar a un flujo de aire turbulento. Debido a esto, la voz se percibe entrecortada y la señal tiende a ser menos periódica e impredecible en su variación ciclo a ciclo [33], lo que produce alteraciones en la estructura armónica de la señal de voz [34]. Dicho lo anterior, varios enfoques han utilizado los coeficientes cepstrales en la escala de frecuencia Mel (*MFCC – Mel Frequency Cepstrum Coefficients*) para caracterizar las voces patológicas. Los MFCC son una representación definida como el cepstrum real de una ventana corta de la señal, derivada del espectro de la transformada rápida de Fourier (*FFT – Fast Fourier Transform*) y no depende de la estimación del *pitch*, ya que su estimación es un problema común en la mayoría de los parámetros acústicos que se encuentran en el estado del arte [28, 34]. En [35] se realiza la implementación de 12 MFCC y el *pitch*, características que fueron evaluadas por medio de un modelo oculto de Markov (*HMM – Hidden Markov Models*). Para este experimento fueron utilizadas 657 voces patológicas y 53 voces sanas de la base de datos MEEI, obteniendo tasa de acierto de 98.3 %. Del mismo modo, en [36] se implementaron los MFCC, además de su primera y segunda derivada, para caracterizar 53 voces sanas y 82 patológicas de la base de datos MEEI. En donde fueron usados dos enfoques de clasificación basados en redes neuronales, conocidos como MLP (*Multilayer Perceptron*) y aprendizaje de cuantificación vectorial (*LVQ – Learning Vector Quantization*), la mejor precisión obtenida fue de 96 % mediante el uso de 24 MFCC y la energía de la ventana. Al igual que en el trabajo anterior, en [37] se estiman los MFCC junto con sus primeras y segundas derivadas, para caracterizar 53 voces sanas y 147 voces patológicas de la misma base de datos anterior. En este caso se utilizó el coeficiente discriminante de Fisher (*FDR – Fishers' Discriminant Ratio*) para evaluar la capacidad de discriminación de cada una de las características, alcanzando una precisión de 94.6 %. Además, los autores concluyen que las segundas derivadas no aportan información relevante durante el proceso de discriminación. Una vez más, la parametrización de voces sanas y patológicas se realizó utilizando MFCC [38], para los experimentos se usó un subconjunto de la base de datos MEEI, pero en este caso los registros se dividieron diferenciando por sexo. El mejor resultado obtenido fue de 88.3 % para una estrategia de sexo-específico en comparación con 87.2 %, para una estrategia de sexo-independiente. En contraste con lo anterior, en [39] se utilizan medidas de perturbación estimadas a partir de los dominios espectrales y cepstrales, junto con un conjunto de siete estadísticas de alto orden (*HOS – High Order Statistics*), que se pueden considerar desde

un punto de vista estadístico como características no lineales, produciendo una precisión de hasta el 98.3 %, combinando la información extraída de las cinco vocales en español. También en [40], los autores han propuesto una técnica no lineal que emplea el operador de energía Teager, para obtener la amplitud y la modulación de las frecuencias de los formantes. Las señales de voz fueron parametrizadas mediante una interpolación polinómica de tercer orden, obteniendo muy buenos resultados en patologías del habla, aunque la prueba se llevó a cabo utilizando un conjunto de datos pequeño (11 grabaciones) extraídas de una base de datos privada.

En relación con el comportamiento no lineal, diversos estudios han demostrado que en el proceso de producción de la voz se presentan diferentes fenómenos físicos con características no lineales, que no pueden ser caracterizados por métodos convencionales basados en técnicas lineales. Este análisis se deriva de la teoría de los sistemas dinámicos, y en la mayoría de los casos, se basa en la reconstrucción del espacio de estado por medio de alguna técnica de embebimiento, así como algunas medidas de complejidad basadas en teoría de la información. En este sentido, algunos investigadores han estado interesados en la aplicación del análisis no lineal para señales de voz con trastornos del habla, uno de estos trabajos se presenta en [41], donde se utiliza la dimensión de correlación (*CD − Correlation Dimension*), la cual evalúa la dimensionalidad intrínseca de la señal de voz, permitiendo describir la complejidad de las vocales sostenidas. La base de datos contiene 79 voces sanas y 68 voces de pacientes con pólipos en los pliegues vocales. La estimación de CD se llevó a cabo utilizando ventanas de $200ms$, que pertenecen a la parte más estacionaria de la expresión. Los autores demostraron que los valores de CD de personas sanas y patológicas tienen diferencias estadísticamente significativas, y concluyen que el análisis no lineal se puede utilizar como método complementario para evaluar y detectar patologías laríngeas. En [42], se estudió el CD y el máximo exponente de Lyapunov (*LLE − Largest Lyapunov Exponent*) junto con otras medidas de complejidad, para caracterizar 51 registros de control y 112 registros patológicos, extraídos de la base de datos MEEI. Los resultados fueron obtenidos por medio de una máquina de soporte vectorial (*SVM − Support Vector Machine*), con una precisión de hasta 94.4 % usando solamente CD. Hay que mencionar, además, que también han sido usadas características basadas en teoría de la información, las cuales intentan cuantificar la complejidad de la señal como una forma alternativa para evaluar el comportamiento no lineal, sin hacer suposiciones sobre la naturaleza de la señal, es decir, determinística o estocástica. En este contexto, la medida más común es la entropía aproximada (*AE − Approximate Entropy*), que cuantifica la regularidad de las fluctuaciones de una serie de tiempo y refleja la probabilidad de que patrones similares de las observaciones no serán seguidos por otras observaciones similares, como se presenta en [43], AE fue usado junto con un parámetro de escala para diferenciar entre voces sanas y patológicas. El número de muestras utilizadas para el entrenamiento fueron 17 voces sanas y 12 patológicas, la prueba se llevó a cabo con 5 muestras adicionales. Los autores concluyen que la AE es una herramienta eficaz para la detección de los trastornos de la voz, pero los resultados no fueron proporcionados en términos de tasas de rendimiento.

Para finalizar el análisis de vocales sostenidas, se presenta uno de los trabajos más actuales y representativos en detección de patologías de la voz [44], en donde los autores evalúan la precisión de diferentes métodos de caracterización para la detección automática de múlti-

ples trastornos del habla. Los trastornos considerados incluyen: disfonía en personas con EP, disfonía diagnosticada en pacientes con diferentes patologías laríngeas (LP) e hipernasalidad en niños con labio y/o paladar hendido (LPH). Se aplicaron cuatro métodos diferentes para analizar las señales de voz: medidas de estabilidad o periodicidad, riqueza espectral, medidas de ruido y comportamiento no lineal. Estas medidas se ponen a prueba en seis bases de datos, tres con grabaciones de pacientes con EP, dos con pacientes con LP y una con los niños con LPH. Las características espectrales-cepstrales se utilizan para modelar el espectro de voz con un énfasis especial en torno a los dos primeros formantes, las precisiones van desde 95 % hasta 99 % en voces hipernasales, lo que confirma la presencia de cambios en el espectro de voz debido a la enfermedad. Las medidas de ruido discriminan adecuadamente entre disfonías y voces sanas en las bases de datos de pacientes que sufren de LP. Mientras que las medidas de estabilidad permiten modelar la vibración anormal de las cuerdas vocales que se observa en pacientes con EP y en personas con LP, alcanzando precisiones que van desde 81 % hasta 99 %. Los resultados obtenidos en este estudio sugieren que no es adecuado utilizar todas las características para modelar las patologías de la voz, por el contrario, es necesario estudiar la fisiología de cada deterioro para elegir el conjunto más apropiado de características.

Otro enfoque estudiado, ha sido el análisis de habla continua, debido a que el análisis de fonación de vocales sostenidas, por si solo, no es suficiente para evaluar la calidad de la voz patológica, ya que no incorpora aspectos dinámicos de la voz como son las variaciones de periodos de tono, el ritmo, la entonación, entre otras [1]. Sin embargo, existen pocos estudios que abordan la detección de patologías de la voz en habla continua. En [29] se encuentran definiciones alternativas a *jitter* y *shimmer*, como son el cociente de perturbación de amplitud (*APQ – Amplitude Perturbation Quotient*), el cociente de perturbación de *pitch* (*PPQ –Pitch Perturbation Quotient*) y la perturbación relativa promedio (*RAP – Relative Average Perturbation*) [45], que se consideran más apropiadas para ser utilizadas en habla continua, ya que tienen en cuenta los cambios presentes en el *pitch*. También han sido utilizadas las medidas de perturbación junto con aquellas que dan información de la riqueza espectral, como es presentado en [46], donde se utilizaron *jitter* y *shimmer* junto con una característica cepstral para detectar patologías, mediante el uso de un conjunto de datos compuesto por 9 hablantes de control y 20 patológicos. Los parámetros fueron estimados en siete palabras aisladas y cinco frases, con una precisión de hasta 85.5 % obtenida por medio de un HMM. Así mismo, en [28] a través de diferentes implementaciones de *jitter*, se realiza la detección de voz patológica en habla continua (texto leído), mediante el método SJE (*Spectral Jitter Estimator*). SJE se basa en una descripción matemática del fenómeno de fluctuación de fase y ha demostrado ser robusto frente a errores de la estimación del pitch, lo que lo convierte en un buen candidato para la medición de *jitter* en habla continua. El rendimiento de la detección fue evaluado por medio de umbrales, obteniendo tasas de alrededor del 95 % para el primer umbral y 87,8 % para el segundo, en términos del área bajo la curva. Además, se encontraron valores de *jitter* estimados en habla continua, que confirman los estudios que muestran una disminución de *jitter* con el aumento de las frecuencias fundamentales, y la presencia más frecuente de altos valores de *jitter* en el caso de voces patológicas conforme

aumenta el tiempo.

Por otra parte, fue extendido el concepto de HNR a las señales de habla continua en [47], donde se obtuvo una tasa de error de 22.5 % en la detección de patologías de la voz. A diferencia de los trabajos anteriores, en [48] se realiza una segmentación sonora-sorda, donde se compararon dos grupos de medidas: las que se extraen de las vocales sostenidas y las basadas en muestras de voz continua. Los experimentos se realizaron en una base de datos de 53 personas sanas y 175 personas patológicas. Se implementaron 9 medidas acústicas, que incluyen el *pitch*, el *shimmer*, las medidas espectrales y de ruido glotal, para evaluar la presencia de trastornos. El rendimiento de estas medidas acústicas se cuantificó mediante LDA, logrando una precisión global alrededor del 94 % para los segmentos sonoros. Cuando se consideraron las medidas individuales de forma aislada, la clasificación fue más precisa para las medidas extraídas de las vocales sostenidas que para las de habla continua; la exactitud fue mejorada en ambos casos cuando se consideraron las combinaciones de parámetros acústicos. Además, en [1] los autores informan tasas de precisión de 96.3 % mediante la red neuronal MLP, considerando catorce MFCC, HNR, NNE, GNE y la primera derivada de cada función, creando un espacio de representación con 36 dimensiones estimadas en un subconjunto de la base de datos MEEI.

En [49], se describe un sistema para la caracterización de patologías laríngeas usando diferentes estadísticas de funciones de dinámica no lineal, en señales de habla continua obtenidas a partir de un subconjunto de la base de datos MEEI. Los resultados reportados por los autores indican tasas de precisión de hasta el 95 %. De igual modo en [50], se presentan diferentes medidas de complejidad sobre el espacio de embebimiento, para el reconocimiento automático de patologías laríngeas. La decisión fue obtenida por una SVM con una precisión de 98,2 %, mejorando el actual resultado en el estado del arte, en cuanto a la clasificación automática de señales de voz patológicas en habla continua. Por otro lado, los autores de [51] realizan un estudio comparativo de la distribución de energía espectral entre voces sanas y disfónicas, tanto para las vocales sostenidas como para el habla continua. Para este propósito, se introduce una nueva medida de estabilidad, tiempo de descorrelación, con el fin de encontrar indicios sólidos de disfonía en el dominio espectral, presentando tasas de acierto entre el 70 % y el 78 %. Algo semejante ocurre en [52], el objetivo de este estudio es comparar el desempeño de la vocal /a/ y el habla continua para el diagnóstico de patologías de la voz, el sistema reconoce entre tres clases, dos conjuntos de personas patológicas y uno de personas de control. Las señales se evalúan utilizando MFCC aplicadas a una SVM y un modelo de mezclas gaussianas (*GMM − Gaussian Mixture Models*). Para el habla continua el sistema GMM llega a una precisión de 74 %, mientras que la SVM obtiene 72 %. Para la vocal sostenida /a/, la exactitud mediante el GMM y la SVM es de 66 % y 69 %, respectivamente, un resultado menor que con el habla continua.

1.1.1. Enfermedad de Parkinson

Luego de haber presentado un estudio de varios de los trabajos más representativos para la detección de patologías de la voz, a continuación, se explorarán los trabajos realizados en cuanto a el reconocimiento automático de la EP.

Considerando los estudios realizados, se ha evidenciado que las personas que padecen

de la enfermedad sufren desórdenes en la comunicación, viéndose afectadas las tres dimensiones del habla, articulación, fonación y prosodia. Para comenzar, se explorará la EP a partir de los trabajos realizados en fonación, debido a que es uno de los aspectos que más afecta a los pacientes con la enfermedad, y la comunidad científica ha orientado gran parte de sus investigaciones al entendimiento de la dinámica de fonación en estos pacientes. En [53], se propone un protocolo para la auto-terapia del habla, diseñado para analizar características particulares de la voz referidas a la comunicación. Reportan resultados exitosos principalmente por su aplicación constante, por lo menos de 6 meses, mostrando efectos positivos en etapas tempranas de la enfermedad. Continuando con el análisis de fonación, en [54] se estudió el contenido de energía sobre combinaciones de consonante-vocal y diptongos, dando como resultado que las palabras son las que mejor reflejan los problemas en el habla, a través de un análisis de varianza (ANOVA). Para el estudio se contó con 19 personas de control y 15 personas con EP. Otro marcador analizado con recurrencia es el *pitch*, en [55] realizan su estimación sobre palabras de tipo vocal-consonante-vocal, concluyendo que los pacientes con EP presentan problemas al pronunciar palabras que exigen una terminación rápida de sonido vocálico, además manifiestan un *pitch* con mayor variación que las personas sanas. Los autores no entregan resultados en cuanto a tasas de acierto, ya que no fue usada una técnica de reconocimiento de patrones. En años recientes, se ha abordado el problema mediante el uso de características clásicamente usadas para detección de patologías de origen orgánico, tales como el análisis de periodicidad y estabilidad, además del estudio del comportamiento no lineal de la voz. Los autores en [56], introducen una nueva medida de disfonía, PPE (*Pitch Period Entropy*), la cual es robusta a entornos acústicos ruidosos. La base de datos utilizada contiene 31 fonaciones sostenidas de personas de control y 23 con EP. El protocolo de evaluación se aplica únicamente a la vocal /ah/ del idioma inglés, y se reportan aciertos de 91.4 % mediante el uso de una SVM. Adicionalmente, se concluye que los métodos no convencionales combinados con el tradicional HNR, tienen una mayor capacidad para diferenciar las personas sanas de aquellas que padecen la EP.

En cuanto al análisis de articulación, en [57] se implementaron varias medidas: el área del espacio vocal (*VSA − Vowel Space Area*), el FCR (*Formant Centralization Ratio*) y su logaritmo natural, y la relación entre el segundo formante de las vocales /i/ y /u/, respectivamente. De acuerdo con los resultados estadísticos, todas las características exceptuando el VSA, presentan una alta capacidad de discriminación entre una voz sana y una con disartria. Así mismo, en [58] se modelan diferentes problemas articulatorios en el habla de las personas con EP, incluyendo la calidad vocal, la coordinación de la laringe, la actividad supralaríngea, la precisión de articulación de consonantes, el movimiento de la lengua, el debilitamiento de oclusión y el momento del habla. A través del estudio de 891 voces sanas y 753 con EP, el mejor resultado reportado fue de 88 % usando una SVM, confirmando las observaciones previas reportadas por otros autores, donde la articulación imprecisa es una de las características más predominante de la EP. En contraste con lo anterior, en [59] se considera el estudio de prosodia en la voz y de articulación mediante análisis de vídeo de problemas motores en personas con EP. Las pruebas fueron realizadas en 16 personas sanas y 16 con EP. La caracterización de las señales de voz fue realizada mediante la tasa de sílabas por segundo, la fracción de longitud en vocales en una frase respecto a la longitud total de la frase (VOT),

entre otros; logrando establecer que en general las medidas de prosodia aportan información para la caracterización de los problemas motores que sufren las personas con EP y que afectan la fluidez en la voz, mediante un análisis de ANOVA.

En los últimos años se ha evidenciado un incremento en el número de publicaciones donde se considera la combinación de, al menos, dos de las tres dimensiones del habla. En [60] se presentó un trabajo donde se evalúa la influencia de la Levodopa en la calidad de voz de pacientes con EP. Para el análisis de fonación se estimaron diferentes características: el promedio del *pitch* y su desviación estándar, NSR (*Net Speech Rate*) y PR (*Pause Ratio*); para el análisis de articulación se implementó VAI (*Vowel Articulation Index*), entre otras características para modelar la prosodia. Las pruebas de voz se realizaron en 23 pacientes con EP y 24 personas de control. Mediante un análisis estadístico, se concluyó que la Levadopa no garantiza una mejora en los parámetros de la voz en paciente con EP. El mismo autor, en [61], presentó un trabajo enfocado al estudio de fonación y prosodia, mediante la evaluación de cuatro frases complejas de 138 pacientes con EP y 50 personas sanas, a través de un análisis estadístico. En cuanto al análisis de fonación, se estimaron medidas basadas en el *pitch*, su varianza y su rango dinámico, encontrando que la variabilidad del *pitch* es menor en personas con EP que en personas sanas. En el caso de la prosodia, los resultados indican que existe una correlación entre algunos de los síntomas de la EP, como la acinesia y las variables de prosodia, relacionadas principalmente con el número de pausas en el habla, la tasa de pausas, entre otras. En [11], los autores consideran un análisis de las tres dimensiones del habla, para la fonación se incluyen medidas de periodicidad, estabilidad y medidas de ruido; para la articulación se realiza un análisis de regularidad y velocidad en la fonación de la palabra /pa-ta-ka/, repetida por los pacientes tan rápido como les sea posible; y, por último, para la prosodia se analiza el número de pausas, el porcentaje de pausas, entre otras medidas. Los datos de la voz se obtuvieron de 46 personas sanas y 24 con EP, el rendimiento global del sistema obtenido mediante una SVM es de 85 %. Como conclusión se encontró que la prosodia es el aspecto más afectado de todas las componentes del habla, sobre todo en etapa temprana de la enfermedad. En [62], los autores realizaron una clasificación automática de la voz de pacientes con EP considerando tres estrategias diferentes para modelar la señal, utilizando un conjunto de 1582 características acústicas extraídas a partir del kit de herramientas conocido como openSMILE [63]. La articulación fue modelada a partir de MFCC, y su primera y segunda derivadas; la prosodia se analiza mediante medidas derivadas de la frecuencia fundamental, la energía, la duración, las pausas, *jitter* y *shimmer*; y la fonación mediante la estimación de parámetros físicos de la glotis. Los experimentos incluyeron 176 hablantes nativos de Alemania, 88 con EP y 88 personas de control, los cuales realizaron un conjunto de tareas: habla espontánea, texto leído, frases, palabras aisladas, vocales sostenidas y la repetición de la sílaba /pa/. La tasa de reconocimiento de pacientes con EP es de 86,5 % evaluando solamente las frases, 94.3 % para fonaciones sostenidas, mientras que es de 81,9 % cuando todas las tareas se combinan con una SVM.

Adicionalmente, en [64] se emplearon características acústicas, de prosodia y de un modelo de las cuerdas vocales, en diferentes tareas del habla: fonaciones sostenidas, repeticiones de sílabas, textos leídos y monólogos. Se realizó una selección con el fin de identificar las características más importantes de cada uno de estos sistemas; para el modelo prosódico

se obtuvo una tasa de reconocimiento del 91 %, con el acústico del 88 % y para el de las cuerdas vocales del 79 %, todos estos resultados fueron obtenidos mediante una SVM. Los autores concluyen que los textos leídos y monólogos son los más significativos cuando se trata de la detección automática de la EP, basado en la articulación y evaluaciones prosódicas. En [65], se realizan experimentos de clasificación automática de la EP, empleando características acústicas, prosódicas y glotales en diferentes tareas del habla: repetición de sílabas, frases, textos y monólogos. Los resultados fueron obtenidos por una SVM para diferenciar entre personas sanas y con EP, alcanzando una tasa de 81.9 %; y de 59.1 % para la escala UPDRS (*Unified Parkinson Rating Scale*), la cual permite medir la gravedad de la enfermedad. Posteriormente, los autores en [22], realizan la detección automática de la EP considerando tres idiomas: alemán, checo y español. El conjunto de grabaciones consideradas incluye: 6 palabras pronunciadas por 176 alemanes (88 con EP y 88 sanos), 13 palabras pronunciadas por un total de 100 colombianos (50 con EP y 50 sanos), y la repetición rápida de /pa/-/ta/-/ka/ que fue pronunciada por 42 checos, así como por los colombianos y alemanes. Se presenta un método basado en la separación de segmentos sonoros y sordos de las señales de voz, considerando los procesos de caracterización y clasificación dependiendo de cada tipo de segmento. Para los sonoros se calcularon 12 MFCC, tres medidas de ruido diferentes y los dos primeros formantes; mientras que para los sonidos sordos se usaron 12 MFCC y la energía de 25 bandas de la escala de Bark. Para el caso de los sonidos sordos, las precisiones máximas reportadas son obtenidas con palabras en español y alemán con un 99 % y 96 %, respectivamente. Para el caso de la repetición /pa/-/ta/-/ka/, se reporta una precisión de 97 % para el alemán y checo, mientras que para el español se alcanza el 99 %. Para finalizar, en [66] es llevada a cabo la evaluación automática de 168 pacientes con EP en diferentes fases, comparando la eficacia de tres estrategias por medio del aprendizaje de regresión regularizada, incluyendo tres tareas diferentes: la fonación sostenida de la vocal /ah/ en inglés, la evaluación diadococinética y el texto leído, todas las grabaciones se realizaron dentro de un tiempo máximo de 4 minutos y fueron adquiridas por un dispositivo portátil. A partir de estas grabaciones, se extrajeron 1582 características para cada sujeto utilizando openSMILE. Los autores redefinieron la extracción de características, para capturar señales relacionados con el tono, incluyendo *jitter* y *shimmer*, con el fin de utilizar con mayor precisión un modelo armónico variable en el tiempo de palabra. Los resultados muestran que la severidad de la enfermedad se puede deducir del habla, con un error medio absoluto de 5.5 explicando el 61 % de la varianza. Según los resultados, las características extraídas del texto leído y tareas diadococinéticas son más eficaces.

La Tabla 1.1 resume los diferentes enfoques descritos hasta el momento, en donde se presentan las características empleadas, la base de datos, la patología que fue estudiada, así como los métodos de clasificación usados para la detección de la enfermedad y sus tasas de acierto.

Tabla 1.1: Resumen del estado del arte de patologías de la voz, incluyendo las características empleadas, la base de datos (normal + patológica), la tarea del habla, el método de clasificación y el mejor acierto.

Característica	Base de datos	Patología	Tarea del habla	Clasificador	(%) Acierto
NNE [23]	MEEI (53 + 173)	Laríngeas	vocal sostenida	LDA	63.2
SFR					98.7
TNI [25]	MEEI (53 + 163)	Laríngeas	vocal sostenida	kNN	96.1
HNR [26]	MEEI (53 + 163)	Laríngeas	vocal sostenida	kNN	94.3
GNE [27]	MEEI (53 + 173)	Laríngeas	vocal sostenida	–	89.9
***Jitter, shimmer*, HNR** [30]	MEEI (573 + 58)	Laríngeas	vocal sostenida	LDA	89.1
MFCC, acústicas [35]	MEEI (53 + 657)	Laríngeas	vocal sostenida	HMM	98.3
MFCC [37]	MEEI (53 + 147)	Laríngeas	vocal sostenida	GMM	94.6
CD, LLE [42]	MEEI (51 + 112)	Laríngeas	vocal sostenida	SVM	98.3
Riqueza espectral, [44]	Privada (108 + 130)	LPH	vocal sostenida	SVM	99.0
Periodicidad,	MEEI (53 + 173)	Laríngeas			99.0
Ruido, DNL	Privada (50 + 50)	Parkinson			91.0
MFCC [52]	MEEI (53 + 724)	Laríngeas	vocal sostenida	GMM, SVM	66 - 69
			habla continua		74 - 72
Jitter, shimmer [46]	Privada (9 + 20)	Laríngeas	habla continua	HMM	85.5
MFCC, ruido [1]	MEEI (23 + 117)	Laríngeas	habla continua	MLP	96.3
Jitter [28]	MEEI (53 + 631)	Laríngeas	habla continua	Umbral	94.8
Energía espectral [51]	MEEI (53 + 173)	Laríngeas	habla continua	Umbral	78.0
fonación [62]	Privada (88 + 88)	Parkinson	vocal sostenida	SVM	94.3
Prosodia			habla continua		86,5
fonación [64]	Privada (23 + 23)	Parkinson	vocal sostenida	SVM	79.0
Acústicas			habla continua		88.0
Prosodia			habla continua		90.5
Acústicas y prosodia [65]	Privada (88 + 88)	Parkinson	habla continua	SVM	81.9
MFCC, Energías de Bark [22]	privadas	Parkinson	habla continua	SVM	84 - 99

1.2. Análisis de las representaciones TF

Después de presentar una revisión del estado del arte sobre detección de patologías de la voz, se busca explorar el uso de técnicas TF para determinar la concentración de energía en el espacio bidimensional. Debido a, que permiten analizar señales cuyas características espectrales varían en el tiempo sin hacer suposiciones de estacionariedad, en muchos casos, a través del análisis de ventanas de la señal mucho mayores a las convencionales (20 y 40 ms), siendo de gran utilidad en escenarios donde se pretende extraer información específica la señal, la cual no puede ser obtenida fácilmente en el dominio del tiempo. Adicionalmente, se presentarán los diferentes tipos de aplicaciones de las técnicas TF en señales biológicas, especialmente en las señales de voz.

En procesamiento de señales, clásicamente ha sido usada la transformada de Fourier, puesto que ha permitido la descomposición de una señal en componentes individuales de frecuencia y ha establecido la intensidad relativa de cada una, consiguiendo una representación de la variación espectral a través del tiempo, es decir, cuáles frecuencias existen mientras dura la señal. Aunque la teoría de Fourier ha sido exitosa en diversas aplicaciones, presenta limitaciones intrínsecas debido a que no puede dar información de la evolución en el tiempo del espectro de la señal [67]. Además, existen señales cuyo contenido espectral cambia tan rápidamente que es muy difícil encontrar una ventana apropiada en el tiempo, debido a que puede existir un intervalo de la señal en que no pueda ser considerada como cuasi-estacionaria, y en casos donde se desea localizar eventos específicos, se debe reducir el

tamaño de la ventana temporal, presentándose una reducción en la resolución de frecuencia, lo que hace evidente el compromiso entre ambas dimensiones [68]. Es por esto, que algunas señales de interés cuyos componentes frecuenciales varían en el tiempo, requieren de un análisis TF que permita simultáneamente analizar estas dos dimensiones, y así aprovechar el poder de la representación en frecuencia sin la necesidad de la caracterización completa en el dominio temporal.

Históricamente, se han propuesto numerosos métodos para el procesamiento y el análisis TF de señales con contenido en frecuencia variante en el tiempo. En 1932, Wigner en el ámbito de la mecánica cuántica, introdujo uno de los primeros conceptos dirigidos al procesamiento de señales en el dominio TF, la distribución Wigner (*WD – Wigner Distribution*), que fue concebida como una distribución estadística bidimensional, que relacionaba los espectros de Fourier de la posición y el momento de una partícula, proporcionando una alta resolución para modulaciones con *chirp* de frecuencia, sinusoides e impulsos [69], aunque en 1948, la distribución fue redefinida por Ville (*WVD – Wigner-Ville Distribution*) [70]. La WVD es la representación cuadrática con interpretación energética que satisface más propiedades matemáticas deseables (real, marginales, conservación de energía, invariante a los desplazamientos en tiempo y frecuencia...). Sin embargo, de acuerdo al principio de incertidumbre, es imposible tener una densidad de energía TF puntual; esta restricción se refleja en el hecho que la WVD puede asumir valores negativos en determinados puntos [71]. Por otro lado, el espectrograma es una de las representaciones más utilizadas para señales de variación lenta o cuasi-estacionarias [19]. Es definido como la magnitud al cuadrado de la transformada de Fourier en tiempo corto *STFT – Short Time Fourier Transform*) y como un esquema que proporciona una interpretación intuitiva del contenido en frecuencia de una señal, además es fácil de calcular [72]. Cabe mencionar, que al ser una función cuadrática sufre de la presencia de términos cruzados, es decir, concentraciones ficticias de energía resultantes de la naturaleza cuadrática de las transformadas, lo cual puede oscurecer las características de interés real de la señal [73]; sin embargo, el espectrograma logra suavizar los términos de interferencia de manera que sólo aparecen en las regiones donde los espectrogramas de cada señal se superponen. Paralelamente, los estudios de Dennis Gabor establecieron las bases de la mayoría de las distribuciones TF (*TFD – Time-Frequency Distribution*) en la forma que hoy se conocen, en cuanto a los conceptos de señal analítica y principio de incertidumbre; diseñó lo que esencialmente es un espectrograma con una ventana gaussiana, reduciendo el efecto del fenómeno de Gibbs [74].

Algunos años después, L. Cohen propuso una clase completa de TFD basadas en la WVD y enfatizó su importancia en el procesamiento de señales. La WVD es la distribución más destacada de esta clase, ya que fue una de las primeras técnicas en obtener información sobre la señal, simultáneamente en el tiempo y la frecuencia, superando las limitaciones de la transformada de Fourier y exhibiendo una mejor resolución, lo que la ha convertido en la distribución más útil y fundamental en el desarrollo de nuevas representaciones o de versiones mejoradas de la misma, a las que se les añade un *kernel* diferente, que define de forma única las propiedades de cada una de las TFD [75]. Por ejemplo, la distribución pseudo-WVD (*PWVD – Pseudo Wigner-Ville Distribution*) y la distribución PWVD suavizada (*SPWVD – Smoothed Pseudo Wigner-Ville Distribution*), utilizan ventanas pasa-bajas con el

fin de suavizar y reducir los términos cruzados que se generan en la WVD. La PWVD realiza el suavizado en frecuencia y atenúa notablemente las interferencias que aparecían con la WVD, así es posible controlar independientemente el suavizado en el eje temporal y en el dominio de la frecuencia, dando lugar a una resolución diferente en cada uno. Tanto para la PWVD como para la SPWVD, si se utilizan ventanas de suavizado pequeñas se reducen notablemente los límites de integración en sus formulaciones y, por lo tanto, también el tiempo de cálculo necesario en relación a la WVD [76, 77]. Otra de las representaciones desarrollada por L. Cohen es la distribución de Born-Jordan (*BJD – Born-Jordan Distribution*) [78], en la que la elección de la función *kernel* es una $sinc$, la cual asegura que se cumplan varias propiedades de la WVD. Si a esta representación se le añade una ventana que realice un suavizado en el dominio de la frecuencia, se obtiene la representación de Zhao Atlas Marks, denominada (*CSKD – Cone-Shaped Kernel Distribution*) [79]. Lo que hace también posible, encontrar la distribución de Page, que se basa en calcular la derivada de la densidad espectral de energía de la señal. En 1968, se dio a conocer la distribución de Rihaczek (*RD – Rihaczek Distribution*), la cual está basada en la construcción de una representación compleja de la densidad de energía en el plano TF [80]. Dado que es una función compleja, la parte real de la distribución también constituye una TFD denominada distribución de Margenau-Hill (*MHD – Margenau-Hill Distribution*), y a diferencia de la WVD, el patrón de términos cruzados generado puede hacer que, en el caso de señales con múltiples componentes, los términos de interferencia se superpongan a los términos propios, este hecho implicaría que las componentes de la señal se acentuaran y, por lo tanto, su detección sea más sencilla [81]. Por otra parte, la distribución de interferencias reducidas (*RID – Reduced Interferences Distribution*) y la distribución de Choi-Williams (*CWD – Choi-Williams Distribution*) son utilizadas comúnmente para alcanzar un compromiso entre los beneficios de la WVD y del espectrograma, por medio de funciones *kernel* unidimensionales, evaluadas con el producto de sus variables de tiempo y frecuencia [82]. La distribución RID emplea una de las ventanas clásicas de duración temporal y normalizada, para satisfacer automáticamente muchas de las propiedades deseadas para una representación TF [83]; mientras que la CWD suaviza los términos cruzados, mediante una función de tipo *kernel* gaussiano que minimiza su contribución y presenta mejores resultados que las demás distribuciones basadas en la WVD. Además, la CWD tiene un factor de escalado que permite seleccionar una buena reducción de los términos cruzados o una buena preservación de los términos propios, pero desafortunadamente no las dos a la vez [82].

En los años 80 se define el escalograma como la magnitud al cuadrado de la transformada *wavelet* (*WT – Wavelet Transform*) [84], el cual permite analizar la información espectral local de una señal en múltiples escalas. Para ello, se utilizan versiones dilatadas y comprimidas de la función *wavelet* madre, dando lugar a una representación denominada multiresolución. A diferencia de la tradicional STFT, que proporciona una resolución fija en el plano TF, la WT ofrece una resolución variable. Para frecuencias elevadas se obtiene una buena resolución temporal, mientras que para las frecuencias bajas se consigue una buena resolución en frecuencia. Uno de los inconvenientes del escalograma es su pobre resolución temporal para las frecuencias bajas, así como la pobre resolución en frecuencia para frecuencias altas.

En contraste con lo anterior, a principios de 1990 Auger y Flandrin acuñaron un nuevo

término denominado reasignación, mostrando que el uso explícito de la fase de la STFT se podía sustituir de manera eficiente mediante una combinación de STFT con ventanas adecuadas [75]. Este fue el punto de partida de su uso en una variedad de nuevos dominios. De forma paralela e independiente, Maes y Daubechies desarrollaron otra técnica basada en la fase que denominaron *synchrosqueezing* [85], su propósito era muy similar al de reasignación, de hecho, es un caso especial, con la ventaja adicional que permite la reconstrucción. A finales de ese mismo año, se vio la introducción de una propuesta radicalmente diferente con la descomposición en modo empírico (*EMD – Empirical Mode Decomposition*) [86], diseñada para extraer componentes AM y FM de los datos de una manera controlada, aunque atractivo por su sencillez y eficacia, carecía de fundamentos matemáticos sólidos. En este contexto, *synchrosqueezing* ha resurgido recientemente como una alternativa más formal, así como una técnica atractiva.

Finalmente, en [87] se propone un nuevo formato de representación, el espectrograma de modulación, que descarta muchos de los detalles espectro-temporal de la señal y en su lugar se centra en la estructura subyacente, en la parte de baja frecuencia de la modulación del espectro distribuido a través de bandas críticas de los bancos de filtros FIR [87].

1.2.1. Aplicaciones en señales biológicas

Las representaciones TF han sido usadas en un sin número de escenarios que incluyen análisis de vibraciones, monitoreo y detección de daños; análisis de vídeos, propagación de la onda electromagnética; ciencias de la tierra como terremotos, mareas, temperatura del agua, aguas subterráneas. Al mismo tiempo, han sido aplicadas en escenarios de las comunicaciones inalámbricas, navegación, radar, sonar, comunicaciones submarinas y sistemas biomédicos [88].

Las primeras aplicaciones en procesamiento de señales usando las representaciones TF, se realizaron con éxito utilizando las ventanas de Gabor, para el procesamiento de audio (eliminación de ruido) e imágenes biomédicas, para problemas tales como la localización de la fuente de la magnetoencefalografía (MEG), debido a su fácil implementación [89, 90]. Por otra parte, EMD es útil para los datos del mundo real que tienen escalas bien separadas, tales como señales fisiológicas y/o ambientales, con un rendimiento de alta resolución en la estimación de la frecuencia instantánea [86], mientras que el método de reasignación ha sido usado en el dominio biológico, en sonidos (o chirridos) de murciélagos, y en señales fisiológicas como son las señales relacionadas con la anestesia, ya que puede proporcionar una representación gráfica que facilita la interpretación de las señales, en particular, en el análisis exploratorio de los datos [91]. La WT ha sido aplicada para detectar el complejo QRS en el electrocardiograma (ECG), debido a la necesidad de encontrar un método eficaz, que sea simple, preciso y que tome poco tiempo de cálculo [92]. En otros trabajos se han comparado diferentes TFD como son el espectrograma, la SPWVD y la WT, para discriminar ECG de pacientes susceptibles de padecer arritmias y controles, con el fin de determinar cuál entrega la mayor cantidad de información [93]. La STFT y la WT han sido evaluadas para analizar su capacidad de caracterizar fonocardiogramas anormales, y examinar la variabilidad intra-paciente e inter-paciente, con el fin de realizar una detección objetiva de los problemas

cardíacos. Los resultados sugieren que las técnicas proporcionan una resolución temporal y frecuencial comparables de los eventos acústicos cardíacos [94]. Con respecto a la CWD, ha sido empleada para estudiar los ritmos anormales en el electrogastrograma teniendo gran éxito, debido a que está libre de suposición de estacionariedad de la señal y, es de hecho, capaz de proporcionar información precisa sobre las variaciones en frecuencia y amplitud [95]. Así mismo, se han aplicado la WT, la WVD y la RID para estudiar la respuesta de las neuronas auditivas tras una estimulación con ruido de banda ancha, sin embargo, no se encontraron diferencias apreciables entre este tipo de representaciones de la señal [96]. En el caso del estudio de los registros de la actividad cerebral, en [97] utilizaron el espectrograma, la WVD y la CWD para caracterizar el comportamiento no estacionario del registro electrocorticográfico (ECoG) de pacientes epilépticos; se muestra que la distribución exponencial representa una mejora considerable sobre el espectrograma en términos de resolución y reduce los términos cruzados presentes en la WV, las representaciones de alta resolución ofrecen unas características temporales-espectrales de las señales que varían rápidamente y son grabadas en la epilepsia del lóbulo temporal.

La WT ha sido también usada para identificar el inicio de ataques epilépticos en el electroencefalograma (EEG) [98], caracterizando los cambios en los potenciales evocados somatosensoriales debido a daños cerebrales, causados por privación de oxígeno [99]. De manera similar, ha sido usada la WT para el estudio de los EEG y los potenciales evocados de los pacientes con Alzheimer [100]. Para el estudio se contó con 10 pacientes con enfermedad de Alzheimer leve y 10 sujetos de control de la misma edad. Los resultados evidencian un rendimiento significativamente mejor de aproximadamente 80 % de sensibilidad contra el 100 % de especificidad, mediante una red neuronal. Además, en [101] algunos métodos fueron aplicados a la superficie EMG de los registros de pacientes con temblor, el análisis TF de cada temblor se calculó a través de la WV filtrada, y se realizó un análisis estadístico para obtener los resultados. En [102], se propuso la utilización de la WT de paquetes (*WPT – Wavelet Packet Transform*) en la codificación de voz que tuvo aplicaciones posteriores en la codificación de música y audio en general [103, 104], y que ha sido utilizado en la mejora de los algoritmos de compresión utilizados en MP3 [105]. Este esquema también fue empleado en [106], para desarrollar un modelo psicoacústico del oído utilizado en aplicaciones para el análisis de la calidad de la voz [107]. Adicionalmente, las técnicas de TF han sido utilizadas para la evaluación de tremor en la enfermedad de Parkinson, en [108] se realizó la evaluación del temblor mediante el análisis de la espiral de Arquímedes, que es un método de análisis del dibujo de espiral para cuantificar la actividad motora y evaluar el posible trastorno. Para la evaluación se hace necesaria la observación clínica para diagnosticar el temblor. Por lo tanto, un método adicional podría ayudar en el proceso de diagnóstico y posiblemente mejorar la detección temprana, así como la medición de la gravedad de la enfermedad. El resultado preliminar de la función seleccionada muestra que el método podría utilizarse para distinguir entre el paciente con el trastorno del movimiento y el paciente de control.

Para finalizar, se muestra un trabajo muy completo con un enfoque metódico para mejorar las TFD cuadráticas mediante el diseño de *kernels* TF adaptados [109], en donde se realizan diagnósticos en tres aplicaciones médicas utilizando el EEG, la variabilidad de la frecuencia cardiaca (HRV) y las señales de voz patológica. La inspección manual y visual de

este tipo de señales multicomponentes no estacionarias, es laborioso, especialmente para grabaciones de larga duración que requiere intérpretes cualificados con posibles fallos y errores subjetivos; para ello es necesario diseñar TFD avanzadas de alta resolución, para la automatización de la clasificación e interpretación. Como los métodos TFD son generales y su cobertura es muy amplia, el artículo se centra en algunas metodologías que utilizan sólo unos pocos problemas. El estudio utilizó 21 cerdos recién nacidos en condiciones de hipoxia, para simular la hipoxia perinatal en los bebés humanos, ya que los cerdos tienen una ontogénesis del sistema nervioso y cardiovascular similar al de los humanos. Tal experimento cuenta con épocas de 5 minutos de EEG, antes y al comienzo de cada estado de hipoxia; además de señales HRV. La clasificación de las señales HRV se realizó a través de una SVM utilizando diferentes tipos de características: el espectrograma, la distribución B-modificada (*MBD – modified-B distributions*) y la MBD extendida, alcanzando tasas de acierto de 91,2 %, 94,9 % y 91,3 %, respectivamente. El método probado con datos de EEG fue la estimación de la frecuencia instantánea usando el algoritmo CPL (*Component Linking*), mostrando un error cuadrático medio menor al 79 %, de las épocas reales. Por último, fueron analizadas las personas con voz esofágica que se caracterizan por no tener laringe, por lo cual producen la señal de voz mediante la expulsión de aire desde el estómago a través del esófago en el tracto vocal. Para los experimentos se tomaron grabaciones de 6 personas y se aplicó el método HNR, en donde se comparó la señal original y la versión mejorada después de aplicar el método, el resultado presentó un promedio de mejora significativa de 2,81 dB.

1.2.2. Aplicaciones en señales de voz

Considerando las ventajas que traen consigo las TFD, se explorarán algunas de las aplicaciones realizadas en las señales de voz, en cuanto a enfermedades que afectan su calidad.

El trabajo presentado en [3], propuso el uso de un enfoque TF para la detección de patologías de la voz. Para los experimentos se contó con 161 registros de control y 51 patológicos de habla continua; en donde las señales de voz se descomponen utilizando una adaptación TF y se extraen varias características como el máximo de octava, media octava, relación de energía, relación de la longitud y proporción de frecuencia. La aproximación logró una precisión global alrededor del 94 %. Además, la WT también se ha utilizado con eficacia en el desarrollo de medidas para la detección de patologías del habla [110]. En [111], se propuso un algoritmo de identificación de trastornos de la voz mediante el uso de la energía de los coeficientes obtenidos de la WT continua (*CWT*). Los autores informaron tasas de acierto alrededor del 85 %, usando un clasificador basado en redes neuronales. Otro enfoque se presenta en [112], donde los autores utilizan un análisis basado en la WT discreta (*DWT – discrete wavelet transform*) y los coeficientes de predicción lineal (*LPC – Linear Prediction Coefficients*), para identificar a los pacientes con nódulos en las cuerdas vocales. En los experimentos fueron utilizados 24 registros patológicos y 24 de control; en donde las señales de voz se descomponen en cuatro niveles, y se caracterizan mediante la raíz cuadrada de los valores de cada nivel. Se presenta un bajo orden en la complejidad computacional en relación con la longitud de la señal de voz, con más del 90 % de precisión en la clasificación. Un trabajo más reciente fue presentado en [113], con un amplio estudio en la identificación de los diferentes trastornos de la voz de

origen orgánico. En primer lugar, se realizó un estudio cualitativo aplicando la STFT y la CWT, con el fin de investigar su capacidad discriminante para identificar voces disfónicas. Se elige como método de parametrización de la señal de voz la WPT, por su capacidad de analizar minuciosamente una señal en varios niveles de resolución. La SVM obtiene resultados de hasta el 100 %.

En [114], se muestra un nuevo enfoque a través del uso de la WT para el análisis de tremor en pacientes con EP. El estudio contó con 52 personas con EP y 30 personas de control; para las pruebas fueron utilizadas diferentes ondas madres de la WT como *Mexican Hat* y *Daubenchies*, con el fin de obtener la serie más adecuada de coeficientes *wavelet*. Además de la supervisión de los coeficientes *wavelet*, que pueden ser considerados como indicadores de la evolución de la enfermedad, el objetivo de la investigación fue correlacionar estos parámetros con la evolución de la enfermedad, aunque no se presentaron tasas de acierto. Por otro lado, en [115] fue usada la WT en un sistema de clasificación de disfonías, aplicando la entropía de Shannon. Se implementaron 5 niveles de la WT para la vocal sostenida /a/, de 13 personas de control y 51 disfónicas; el sistema tuvo una tasa de éxito del 84.3 %. De manera semejante, en [116] se utiliza un método de estimación de frecuencia con base en la CWT, para el estudio del temblor vocal. El método de análisis se compara con un método basado en eventos y la transformada Hilbert, para señales de voz de 10 personas con EP y 8 con voces sanas. Los resultados sugieren que la relación de la energía espectral de la huella de frecuencia vocal en los intervalos $(1-5Hz)$ y $(5-20Hz)$, difieren entre los hablantes.

Aparte de los métodos anteriores, en [117] se presentan dos nuevos métodos basados en EMD para el análisis e identificación de voces patológicas. En primer lugar, se introduce un método que permite la extracción de la frecuencia fundamental de las vocales sostenidas, el cual se basa en el algoritmo conjunto de EMD. Como segunda herramienta basada en EMD, se exploran las propiedades espectrales de las funciones del modo intrínseco, demostrando que simplemente utilizando un algoritmo básico de reconocimiento de patrones como kNN, se pueden alcanzar una buena tasa de acierto. Los resultados indican que es suficiente seleccionar sólo tres modos de las características espectrales para discriminar entre las 53 voces sanas y las 657 patológicas; de esta manera el sistema alcanzó una tasa de precisión de 93.4 %. Aunque una gran parte de las investigaciones se basan en el análisis de medidas acústicas extraídas de vocales sostenidas, los autores en [118] incluyen una metodología basada en EMD para la clasificación de señales de habla continua, de voces sanas y patológicas. EMD se utiliza para descomponer porciones elegidas al azar de señales de voz en funciones del modo intrínseco, que luego se analizan para extraer características temporales y espectrales significativas, incluyendo características instantáneas que pueden capturar información discriminativa en señales ocultas en escalas de tiempo locales. Se extrae un total de seis características, a partir de una base de datos que consta de 51 voces sanas y 161 voces patológicas, para la cual se obtiene una precisión de 95.7 % usando LDA como método de reconocimiento.

Los cambios del espectro acústico son analizados por medio de la técnica conocida como modulación espectral, en donde los espectros de modulación (*MS – Modulation Spectra*) pueden ser vistos como una manera no paramétrica para representar la modulación introducida por la presencia de patología, en donde las voces disfónicas se caracterizan en frecuencia y

se van variando en el tiempo las fluctuaciones de amplitud [119]. Los MS han sido usados en diferentes tipos de aplicaciones, tales como: el reconocimiento de hablante [120] [121] y en detección de patologías [122–124]. En este sentido, en [122] son extraídas 25 características de los MS y utilizadas para la detección automática de voces patológicas. Fueron empleados 173 registros de voces patológicas y 53 de voces sanas de la base de datos MEEI. Los resultados indicaron que la mejor tasa de clasificación fue de 94.1 %. Así mismo, en [123], se realiza un análisis de MS normalizados en la escala de frecuencias Mel, para la detección automática de patologías de la voz. El tamaño del espacio original se reduce utilizando un análisis de descomposición en valores singulares (*SVD − Singular Value Decomposition*). Además, se seleccionan las características más relevantes basadas en la información mutua entre la calidad de la voz subjetiva y las características calculadas, lo que conduce a una adaptación de la representación de clasificación de los MS. Se utilizaron dos bases de datos, un subconjunto de la base de datos MEEI con 173 registros patológicos y 53 sanos; y un subconjunto de la base de datos PdA (*Príncipe de Asturias*), la cual contiene 100 registros de voces sanas y 100 registros de voces disfónicas. Obteniéndose una precisión de hasta 94.1 %. Como trabajo siguiente, en [125] combinan las características MS con MFCC para la detección automática de patologías del habla, reduciendo las dimensiones originales de los MS por medio de la SVD de alto Orden (*HOSVD − Higher Order Singular Value Decomposition*), utilizando las mismas bases de datos del trabajo anterior. La menor tasa de error fue de 3.63 % para la unión de características implementadas, obtenida con una SVM. Luego, en [124] se presenta un estudio de voces patológicas, el cual está basado en la fusión de información obtenida por medio de las características cepstrales y derivadas de los MS; mediante un esquema de dos pasos de clasificación. En primer lugar, se utilizaron las características, MFCC y MS, para alimentar a dos clasificadores diferentes e independientes, y luego se utilizaron las salidas de cada clasificador en una segunda etapa de clasificación. Para el estudio se utilizaron las mismas bases de datos anteriores, en donde las tasas de rendimiento obtenidas con la SVM alcanzan más del 97 %, demostrando que la combinación de los MFCC y características extraídas de los MS, producen una mejora en la precisión. Adicionalmente, en [126] se explora la información proporcionada por una representación conjunta de frecuencia acústica y modulada, para la detección de personas con EP. La base de datos contiene las fonaciones sostenidas de las cinco vocales del idioma español de 50 personas con EP y 50 personas de control. El conjunto de características incluye los centroides y el contenido energético de diferentes bandas de frecuencia de los MS. Además, con el objetivo de eliminar la posible redundancia en la información proporcionada por las características, se aplican dos técnicas de extracción de características, el análisis de componentes principales (*PCA − Principal Component Analysis*) y LDA. Se utiliza un sistema de clasificación basado en GMM, con la mejor precisión alcanzada por la vocal /i/ con 71 %.

Un enfoque más actual se encuentra en [127], el cual explora el análisis de componentes de baja frecuencia de las señales de habla continua de personas con EP, con el fin de detectar cambios en el espectro que podrían estar asociados a la presencia de tremor en la voz. Se utilizan diferentes técnicas TF (MS, WVD y WPT) para la caracterización del contenido de baja frecuencia. El conjunto de variables extraídas de las representaciones TF incluye centroides y la energía a su alrededor, junto con las medidas de entropía y un operador de

energía no lineal, que se utilizan como características para la detección automática de las personas con la EP. La capacidad de discriminación de las características estimadas se evalúa mediante tres estrategias de clasificación: GMM, GMM-UBM y SVM. Además, la información proporcionada por las diferentes técnicas TF se combina utilizando una segunda etapa de clasificación. Los resultados muestran que los cambios en los componentes de baja frecuencia son capaces de discriminar entre las personas con Parkinson y personas sanas con una precisión de 77 %, utilizando una sola frase.

La Tabla 1.2 resume algunos de los trabajo descritos anteriormente, utilizando el análisis TF en detección de patologías de la voz. Prestando especial atención a las características implementadas para la detección de la patología del habla, así como la base de datos usada, los métodos de clasificación y las tasas de acierto obtenidas en la detección de la enfermedad.

Tabla 1.2: Resumen del estado del arte de la aplicación de distribuciones TF en la detección de patologías de la voz, incluyendo las características empleadas, la base de datos (normal + patológica), el método de clasificación y el mejor acierto.

Característica	Base de datos	Patología	Clasificador	(%) Acierto
Adaptación TF [3]	MEEI (161 + 51)	Laríngeas	LDA	94.0
DWT, LPC [112]	Privada (24 + 24)	Laríngeas	SVM	91.6
Energía CWT [111]	MEEI (100 + 95)	Laríngeas	ANN	80 - 85
WPT [113]	MEEI (53 + 67)	Laríngeas	SVM	100
WT, BBA [115]	Privada (13 + 51)	Tremor	ANN	84.3
EMD [117]	Privada (53 + 657)	Laríngeas	kNN	93.4
EMD [118]	Privada (151 + 161)	Laríngeas	LDA	95.7
TFTE [128]	Privada (235 + 249)	Laríngeas	GMM	95.0
MS, MFCC [124]	MEEI (53 + 173)	Laríngeas	SVM	97.0
MS [126]	Privada (50 + 50)	Parkinson	GMM	71.0

A partir de la revisión del estado del arte se puede ver que se ha hecho un trabajo fuerte en el campo de la detección de patologías de la voz, se han analizado señales de voz en vocales sostenidas y voz continua, con un amplio número de variables que dan información de diferentes fenómenos que están involucrados en la producción de la voz, pero en el caso particular de la EP, hay algunas diferencias en términos de lo que le sucede a la señal de voz respecto a otras patologías, como son las de tipo orgánico y las que más se analizan, ya que principalmente en éstas se ven afectados los patrones vibratorios de los pliegues vocales; mientras que, en el caso de la EP, se ven afectadas las diferentes dimensiones de la producción del habla, como son la respiración, fonación, articulación y prosodia, por lo tanto, el análisis de las señales de voz debe incluir el análisis de palabras, frases y/o monólogos, para poder evaluar las habilidades de calidad de voz y comunicación de los pacientes. En general, las técnicas basadas en el análisis de voz continua se basan en algún procedimiento de segmentación para identificar los periodos sonoros y sordos, sin embargo estos procedimientos siguen presentando problemas para la determinación del punto de inicio y final de los segmentos. Es importante resaltar que, el uso de técnicas TF beneficia la posibilidad de caracterizar los cambios espectrales introducidos en la señal de voz a causa de la EP, entre ellos el tremor, ya que al ser una perturbación en baja frecuencia no puede ser detectada mediante segmentos cortos, como los obtenidos mediante la segmentación

(sonoros/sordos), convirtiéndose en un método adecuado para el análisis de señales cuyas características espectrales varían en el tiempo como son las señales de voz, cómo es posible concluir del estado del arte presentado.

Capítulo 2

Análisis y caracterización TF de señales de voz

A través de los años se ha hecho evidente la necesidad del estudio y análisis de las señales del mundo, como son: las ondas sísmicas, las vibraciones de motores, la propagación de la onda electromagnética, los parámetros atmosféricos; al igual que las señales biomédicas: fonocardiogramas, electrocardiogramas y señales de voz. Para el procesamiento de estas señales ha sido ampliamente utilizado el análisis en el dominio temporal, no obstante, también ha sido de gran utilidad disponer de una representación en el dominio de la frecuencia, dado que permite extraer características que no suelen ser evidentes en el dominio temporal y ayudan a comprender su naturaleza; mientras el dominio temporal da información sobre cómo varía la amplitud de la señal a lo largo del tiempo, el dominio de la frecuencia indica con qué frecuencia suceden estas variaciones.

Una herramienta clásica para extraer información de las señales es la transformada de Fourier, la cual determina la composición frecuencial de la señal y, sólo trabaja bien, si la señal a analizar está compuesta de componentes estacionarias o cuasi-estacionarias durante su periodo de análisis. Sin embargo, cuando se trabaja con señales del mundo real cuyas características espectrales varían en el tiempo, conocidas como señales no estacionarias, la información proporcionada por la transformada de Fourier no es suficiente, puesto que indica las componentes frecuenciales de la señal pero no el instante en el que éstas aparecen [68, 129]. Considerando la no estacionariedad de las señales del mundo real, como en las señales de voz, se hace necesario un análisis bidimensional de la señal que permita analizar ambos dominios al mismo tiempo, y es conocido como análisis tiempo-frecuencia (TF).

El análisis TF es un conjunto de técnicas para la caracterización y manipulación de señales cuyas estadísticas varían en el tiempo, en un espacio bidimensional: el plano TF, su objetivo es proveer información directa acerca de las componentes de frecuencia que ocurren para cualquier tiempo dado, combinando la información local del espectro instantáneo con la información global del comportamiento temporal de la señal [130]. Este tipo de análisis es de gran interés cuando el modelo de la señal no está disponible, como en el caso de las señales de voz. Debido a su naturaleza no estacionaria y multicomponente [131], estas señales se pueden representar adecuadamente mediante distribuciones TF (*TFD – Time-*

Frequency Distribution), que pueden mostrar la forma en la cual se distribuye la energía de la señal en el espacio bidimensional, aprovechando las características producidas por la concentración de la energía en estas dos dimensiones en vez de sólo una [132].

Actualmente existe un gran número de TFD, que se pueden clasificar en lineales y cuadráticas. La clase lineal es ampliamente utilizada debido a su sencillez, ya que muestra la señal descompuesta en el plano TF basado en la amplitud de la señal temporal, un gran ejemplo es la transformada de Fourier en tiempo corto (*STFT – Short Time Fourier Transform*). Mientras que la clase cuadrática, también conocida como clase de Cohen, realiza una descomposición de la energía y la distribuye en el plano TF de la señal, uno de sus miembros más representativos es la WVD. Además, es importante resaltar que cada clase se puede definir mediante una expresión común, y cada uno de los miembros se determina con un *kernel* dentro de dicha expresión, así es posible examinar las propiedades de cada TFD estudiando su *kernel* [68].

En el presente capítulo se presentará uno de los aspectos más relevantes para el desarrollo de este trabajo, la etapa de caracterización de un sistema clásico de reconocimiento de patrones. La figura 2.1 muestra un diagrama de bloques con el esquema general del sistema desarrollado para la detección automática de la EP por medio de señales de voz. Las dos etapas dentro del cuadro de líneas discontinuas corresponden a la caracterización basada en el análisis TF.

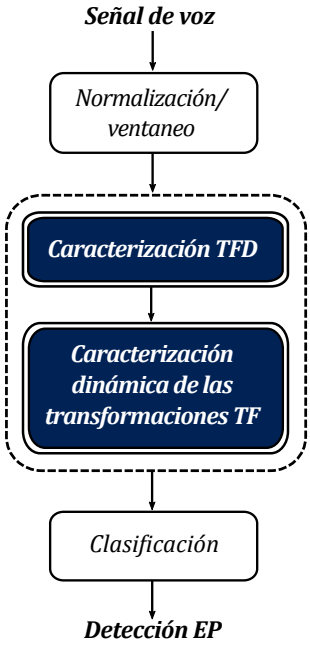

Figura 2.1: Diagrama general de un sistema para la detección automática de las señales de voz con EP.

La intensión de esta metodología, es evaluar la capacidad discriminante de diferentes tipos de características basadas en el análisis TF. En primer lugar, se describen las propiedades generales de las representaciones TF, que permiten el análisis de las ventajas y desventajas de

cada uno de sus miembros. A continuación, se presentan las técnicas de las representaciones TF elegidas en este trabajo, para la detección automática de la EP a través de las señales de voz. Para finalizar, la última sección se centra en las técnicas de caracterización dinámica de los espectros obtenidos por las representaciones TF, como estrategia para disminuir el volumen de datos y extraer la información más relevante de los espectros.

2.1. Definición y propiedades de las TFD

Las TFD permiten representar adecuadamente las señales que no tienen un modelo disponible, debido a su naturaleza no estacionaria y multicomponente, como es el caso de la señal de voz [131]. Las distribuciones tienen como objetivo proveer información directa acerca de las componentes de frecuencia que ocurren para cualquier tiempo dado, combinando la información local del espectro instantáneo, con la información global del comportamiento temporal de la señal; y a su vez muestra la forma en la cual se distribuye la energía de la señal [130].

Existe una gran variedad de TFD con diferentes propiedades, ventajas y desventajas; entre estas se encuentra la clase más significativa de distribuciones que es conocida como la clase cuadrática o de Cohen y se puede expresar mediante [133]:

$$T_x(t,f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{j2\pi\theta(u-t)}\phi(\tau,\theta)x\left(u+\frac{\tau}{2}\right)x^*\left(u-\frac{\tau}{2}\right)e^{-j2\pi f\tau}\,du\,d\tau\,d\theta \tag{2.1}$$

en donde τ y θ, corresponden a los corrimientos en tiempo y en frecuencia, respectivamente, de la señal $x(t)$. Y $\phi(\tau,\theta)$ es una función llamada *kernel*, la cual es análoga a las ventanas utilizadas en el análisis espectral clásico y determina cómo la energía de la señal se distribuye en tiempo y en frecuencia, además define de forma única las propiedades de cada una de las distribuciones pertenecientes a esta clase.

Es importante tener presente, que el dominio (t,f) representa la señal como una función de tiempo y frecuencia reales, mientras que el dominio (τ,θ) la representa como una función de desplazamientos de tiempo y de frecuencia. Dicho lo anterior, la expresión (2.1) se puede escribir de forma aún más sencilla, introduciendo la función de ambigüedad, $A(\tau,\theta)$, que es considerada como una función de autocorrelación conjunta en tiempo y frecuencia, definida como [133]:

$$A(\tau,\theta) = \int_{-\infty}^{\infty} x\left(u+\frac{\tau}{2}\right)x^*\left(u-\frac{\tau}{2}\right)e^{j2\pi\theta u}\,du \tag{2.2}$$

Luego, la expresión (2.1) puede ser reescrita con el *kernel* $\phi(\tau,\theta)$ en el dominio de ambigüedad,

$$T_x(t,f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\tau,\theta)\phi(\tau,\theta)e^{-j2\pi(t\theta+f\tau)}\,d\tau\,d\theta \tag{2.3}$$

Esta expresión es más simple que la (2.1) y permite una mejor interpretación, ya que se puede entender a las TFD de la clase de Cohen como una transformada de Fourier en dos dimensiones del producto entre la función de ambigüedad y el *kernel* de la representación. Mediante esta simplificación se puede facilitar la implementación de las representaciones,

pues hace posible intercambiar las sumatorias anidadas correspondientes a las integrales, por operaciones matriciales e implementaciones de la transformada de Fourier en dos dimensiones.

Se debe agregar, que la cantidad de TFD pertenecientes a la clase de Cohen es infinita, puesto que se puede diseñar un gran número de *kernels* que pueden dar origen a igual cantidad de representaciones y cada una puede satisfacer diferentes propiedades. Sin embargo, en general la clase cuadrática de TFD, sufre de la presencia de términos cruzados, que representan componentes inexistentes en el mapa TF, es decir, concentraciones ficticias de energía resultantes de la naturaleza cuadrática de las transformadas, lo cual puede oscurecer las características de interés real de la señal y que aparecen cuando se analizan señales con múltiples componentes o con características no lineales [73, 134].

Además, la resolución TF es otra característica significativa y varía dependiendo del tipo de aplicación o de la señal que vaya a ser procesada, se pueden tener múltiples componentes con una estrecha separación en tiempo, en frecuencia o en ambas; en estos casos, muchas de las distribuciones pueden fallar a la hora de presentar la verdadera estructura de la señal, dado que se pueden sobreponer varios componentes debido a la baja resolución. A pesar de los esfuerzos por definir distribuciones que reduzcan el efecto de los términos cruzados y al tiempo mejoren la resolución TF, siempre debe existir un compromiso de resolución generado por el principio de incertidumbre, por este motivo la elección adecuada de la distribución debe ser dependiente de la aplicación y de la naturaleza de la señal [129].

2.1.1. Propiedades generales de las TFD

Una forma de seleccionar cuál es la TFD más adecuada, es examinar si cumple con ciertas propiedades necesarias para la aplicación en particular. A continuación, se resumen varias de las propiedades más importantes [135].

- **P1: Real**

  Para que una TFD sea real, ésta debe ser igual a su conjugada compleja, esta propiedad permite que los datos obtenidos por la TFD sean manejables.

$$T_x(t,f) = T_x^*(t,f), \quad \forall x(t) \tag{2.4}$$

- **P2: Positiva**

  Si una TFD se va a interpretar como una distribución bidimensional de energía de la señal, entonces debe ser positiva.

$$T_x(t,f) \geq 0, \quad \forall x(t) \tag{2.5}$$

- **P3: Preservación de los marginales de tiempo y de frecuencia**

  Para interpretar la TFD como una distribución de energía en dos dimensiones en el plano TF, entonces al integrar la variable de frecuencia, debe dar como resultado la energía instantánea de la señal en el dominio del tiempo,

$$\int_{-\infty}^{\infty} T_x(t,f)\,df = |x(t)|^2, \quad \forall x(t) \tag{2.6}$$

  Mientras que al realizar la integral en el eje de tiempo, debe resultar la función de densidad espectral de energía de la señal, $\left|X(f)\right|^2$.

$$\int_{-\infty}^{\infty} T_x(t,f)\,dt = \left|X(f)\right|^2, \quad \forall x(t) \tag{2.7}$$

- **P4: Preservación de la energía**

  Si la TFD es una distribución de la energía de la señal en todo el plano TF, entonces al integrarla se debe obtener la energía total de la señal, Ex.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} T_x(t,f)\,dt\,df = \int_{-\infty}^{\infty} |x(t)|^2\,dt = \int_{-\infty}^{\infty} \left|X(f)\right|^2 df = E_x \tag{2.8}$$

  La propiedad se cumple automáticamente si los marginales lo hacen, aunque lo contrario no se cumple.

- **P5: Preservación de los momentos en tiempo y en frecuencia**

  Esta propiedad establece que el valor del n-ésimo momento de tiempo de la energía instantánea de la señal, $|x(t)|^2$, y el n-ésimo momento de tiempo de la TFD, deben ser idénticos.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} t^n T_x(t,f)\,dt\,df = \int_{-\infty}^{\infty} t^n |x(t)|^2\,dt \tag{2.9}$$

  De la misma forma pasa con el n-ésimo momento de la densidad espectral de potencia, $\left|X(f)\right|^2$, y el n-ésimo momento de frecuencia de la TFD.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f^n T_x(t,f)\,dt\,df = \int_{-\infty}^{\infty} f^n \left|X(f)\right|^2 df \tag{2.10}$$

- **P6: Covarianza al desplazamiento en el tiempo y en la frecuencia**

  Una TFD ideal debe ser invariante a cualquier desplazamiento constante en el retraso de grupo de la señal.

$$y(t) = x(t-t_0) \;\Rightarrow\; T_y(t,f) = T_x(t-t_0,f) \tag{2.11}$$

Equivalentemente, si la señal se modula o se desplaza en frecuencia en una cantidad f_0, entonces la TFD de esta señal se debe correr f_0 en frecuencia. Siendo una propiedad muy importante para analizar una gran variedad de señales como voz, música o sonar.

$$y(t) = x(t)e^{j2\pi f_0 t} \;\Rightarrow\; Y(f) = X(f - f_0) \;\Rightarrow\; T_y(t,f) = T_x(t, f - f_0) \tag{2.12}$$

- **P7: Covarianza a la escala**

Si el eje de tiempo de la señal se comprime por un factor de escala a, entonces el eje de tiempo de la TFD se debe comprimir por el mismo factor a, y el eje de frecuencia se expande por el factor $1/a$.

$$y(t) = \sqrt{|a|}x(at) \;\Rightarrow\; T_y(t,f) = T_x\left(at, \frac{f}{a}\right) \tag{2.13}$$

- **P8: Soporte finito de la señal**

Si una señal comienza en el tiempo t_1 y termina en t_2, entonces una TFD ideal debe comenzar y terminar al mismo tiempo. Esta propiedad es muy intuitiva, ya que la TFD no toma valores mientras no lo hace la señal, sin embargo, no garantiza que la distribución será igual a cero cada vez que la señal o su espectro sean iguales a cero.

$$x(t) = 0 \;\text{ para }\; t \neq (t_1, t_2) \Rightarrow T_x(t,f) = 0 \;\text{ para }\; t \neq (t_1, t_2),\;\; t_1 < t_2 \tag{2.14}$$

Al mismo tiempo, se establece que si la transformada de Fourier de la señal es de banda limitada, entonces su TFD debe tener el mismo soporte diferente de cero en el dominio de la frecuencia.

$$\left|X(f)\right| = 0 \;\text{ para }\; f \neq (f_1, f_2) \Rightarrow T_x(t,f) = 0 \;\text{ para }\; f \neq (f_1, f_2),\;\; f_1 < f_2 \tag{2.15}$$

- **P9: Frecuencia instantánea**

Una de las formas de describir la evolución frecuencial de una señal a lo largo del tiempo, es a través de su frecuencia instantánea. Esta propiedad establece que el primer momento normalizado en frecuencia de la TFD, debe ser igual a la frecuencia instantánea de la señal. Así, el valor medio de la distribución o el centro de gravedad en la dirección de la frecuencia, debe corresponder a la frecuencia instantánea.

$$\frac{\int_{-\infty}^{\infty} f\, T_x(t,f)\, df}{\int_{-\infty}^{\infty} T_x(t,f)\, df} = f_x(t) = \frac{1}{2\pi}\frac{d}{dt}\arg\{x(t)\} \tag{2.16}$$

- **P10: Retraso de grupo**

  Esta propiedad es dual a P9, y establece que el valor medio normalizado de la TFD o centro de gravedad en la dirección del tiempo, debe ser igual al retraso de grupo de la señal.

$$\frac{\int_{-\infty}^{\infty} t T_x(t,f)\,dt}{\int_{-\infty}^{\infty} T_x(t,f)\,dt} = -\frac{1}{2\pi}\frac{d}{df}\arg\{X(f)\} \tag{2.17}$$

- **P11: Localización en tiempo y en frecuencia**

  Si la señal es un impulso perfectamente localizado en el tiempo t_0, entonces su TFD también debe estar concentrada en el tiempo t_0.

$$x(t) = \delta(t-t_0) \;\Rightarrow\; T_x(t,f) = \delta(t-t_0) \tag{2.18}$$

  donde δ es la función impulso unitario. Similarmente, si la señal es una sinusoide compleja cuya transformada de Fourier está perfectamente concentrada alrededor de cierta frecuencia f_0, entonces su TFD también debe estar perfectamente concentrada alrededor de esta misma frecuencia.

$$X(f) = \delta(f-f_0) \;\Rightarrow\; T_x(t,f) = \delta(f-f_0) \tag{2.19}$$

2.2. Técnicas y análisis TF

2.2.1. Distribución Wigner-Ville (WVD)

La WVD es la distribución más destacada de la clase de Cohen, ya que fue una de las primeras técnicas en obtener información sobre la señal, simultáneamente en el tiempo y la frecuencia, superando las limitaciones de la transformada de Fourier y exhibiendo una mejor resolución, lo que la ha convertido en la distribución más útil y fundamental en el desarrollo de nuevas representaciones o de versiones mejoradas de la misma.

La WVD se caracteriza por tener un *kernel* $\phi(\tau,\theta) = 1$, y al reemplazarlo en la expresión general (2.1) de las TFD, se obtiene su expresión:

$$T_{WVD}(t,f) = \int_{-\infty}^{\infty} x\left(t+\frac{\tau}{2}\right) x^*\left(t-\frac{\tau}{2}\right) e^{-j2\pi f\tau}\,d\tau \tag{2.20}$$

A partir de la expresión (2.3) se puede observar que la WVD en el plano de ambigüedad, corresponde a un par de transformadas de Fourier tal como se puede observar,

$$T_{WVD}(t,f) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(\theta,\tau)\,e^{-j2\pi(t\theta+f\tau)}\,d\tau\,d\theta \tag{2.21}$$

Además, esta distribución ha sido de gran interés, debido a que es la interpretación energética que satisface más propiedades matemáticas deseables (P1, P3-P11, entre otras) y algunas de las más importantes son:

– Real (P1).
– Satisface los marginales (P3).
– Conserva la energía (P4).
– Invariante a los desplazamientos en tiempo y en frecuencia (P6).
– Soporte finito en tiempo y en frecuencia (P8).

Por el contrario, no satisface la condición de positividad, tomando valores negativos en algunos puntos; que se debe al hecho de que la WVD es una distribución bidimensional de la energía de la señal en el plano TF y, de acuerdo al principio de incertidumbre, es imposible tener una densidad puntual de energía TF [133, 135].

Hay que mencionar, además, que para obtener la WVD en un tiempo particular, es necesario sumar segmentos formados por el producto de la señal en el pasado, multiplicados por la señal en un tiempo futuro, siendo igual el intervalo de señal que se toma en el tiempo pasado y en el futuro. Esta operación hace que en general la WVD no sea necesariamente cero en los instantes de tiempo en los cuales la señal es cero y tampoco puede que sea cero para frecuencias que no existen en el espectro de la señal. Los componentes causados por este fenómeno son llamados términos cruzados o de interferencia, la causa de este comportamiento se atribuye al hecho que la WVD es bilineal con respecto a la señal $x(t)$ (producto de la señal consigo misma) [133]. Para ilustrar este efecto, se puede expresar la $x(t)$ como la suma de dos señales $x_1(t)$ y $x_2(t)$,

$$x(t) = x_1(t) + x_2(t) \tag{2.22}$$

aplicando la definición (2.20) a la señal $x(t)$, se tiene

$$T_{WVD}(t,f) = T_{WVD}(t,f)^{11} + T_{WVD}(t,f)^{22} + T_{WVD}(t,f)^{12} + T_{WVD}(t,f)^{21} \tag{2.23}$$

donde,

$$T_{WVD}(t,f)^{12} = \int_{-\infty}^{\infty} x_1\left(t+\frac{\tau}{2}\right) x_2^*\left(t-\frac{\tau}{2}\right) e^{-j2\pi f\tau}\, d\tau \tag{2.24}$$

La expresión es llamada WVD cruzada y además es compleja, sin embargo, ${T_{WVD}}^{12} = {T^*_{WVD}}^{21}$ y ${T_{WVD}}^{12} + {T_{WVD}}^{21}$ son reales. De esta manera, la expresión (2.23) se puede simplificar como,

$$T_{WVD}(t,f) = T_{WVD}(t,f)^{11} + T_{WVD}(t,f)^{22} + 2\Re\left\{T_{WVD}(t,f)^{12}\right\} \tag{2.25}$$

De modo que la WVD de la suma de dos señales no es igual a la suma de las WVD de cada señal, sino que los dos primeros términos son las auto-componentes y el tercero es un término adicional $2\Re\left\{T_{WVD}(t,f)^{12}\right\}$, que corresponde al término cruzado o de interferencia, y trae como consecuencia artefactos en la representación TF [133].

La localización de los términos cruzados se ilustra en la figura 2.2 para una señal de dos componentes. En el dominio TF, las dos componentes de la señal (auto-términos) se encuentran concentradas alrededor de los puntos (t_1, f_1) y (t_2, f_2), mientras que el término cruzado (t_x, f_x) es el punto medio entre los auto-términos, $t_x = (t_1 + t_2)/2$ y $f_x = (f_1 + f_2)/2$, como se puede observar en la figura 2.2a. Además, a medida que aumenta la distancia entre los auto-términos en el plano TF aumenta la frecuencia de oscilación, la cual se presenta en

el término cruzado con un periodo de $1/\theta_x$ respecto al tiempo y $1/\tau_x$ respecto a la frecuencia. En el dominio de ambigüedad (figura 2.2b), se cambia la ubicación de los términos, ahora los términos cruzados corresponden a $\tau_x = t_1 + t_2$ y $\theta_x = f_1 + f_2$, mientras que el auto-término se ubica en el origen [134].

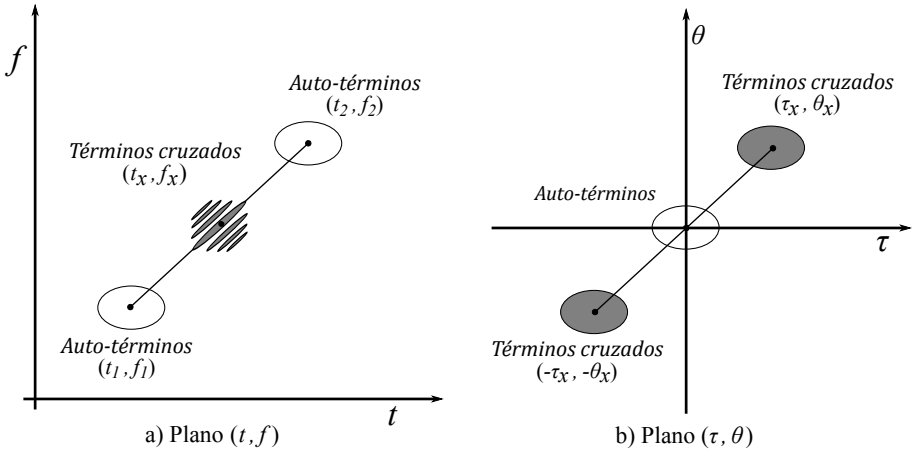

Figura 2.2: Localización de los términos cruzados

Debido a que los términos cruzados pueden enmascarar los términos propios, se ha hecho necesario buscar representaciones TF mejoradas que permitan mantener los auto-términos y suavizar los términos cruzados al mismo tiempo, aprovechando el hecho de que estos términos son oscilatorios, en general pueden ser eliminados realizando un suavizado o filtrado pasa-bajas sobre la superficie TF, principalmente en el dominio de ambigüedad, ya que al concentrarse los auto-términos en el origen se hace más fácil eliminar las interferencias presentes en la distribución de energía (figura 2.2b).

Considerando que la WVD se relaciona con la función de ambigüedad por medio de una transformada de Fourier en dos dimensiones, la forma más simple de reducir los términos cruzados es realizando un filtrado en el dominio de ambigüedad, antes de aplicar la transformada de Fourier que hace regresar al dominio TF. Por esta razón, se utiliza la función *kernel*, $\phi(\tau,\theta)$, que actúa como un filtro bidimensional de tal forma que deja pasar la región del plano de ambigüedad cercana al origen (donde se encuentran los auto-términos), y al mismo tiempo atenúa el resto del plano. Además, se observa un compromiso entre la reducción de términos cruzados y la resolución, pues cualquier truncado de los auto-términos causado por el filtro $\phi(\tau,\theta)$, resulta en dispersión de las componentes en el dominio TF, y por tanto, en pérdida de la resolución de la representación. De esta forma, cuando se define el *kernel* en el dominio de ambigüedad y el ajuste de sus parámetros, se debe tener en cuenta el mejor compromiso entre resolución TF y supresión de términos cruzados [136]. Se hace importante resaltar, que cualquier TFD de la clase de Cohen puede considerarse como una variante de la WVD a la que se añade un *kernel* diferente, que generalmente corresponde a un filtrado de la WVD con el fin de suavizar los términos cruzados sin modificar las propiedades más importantes de la distribución.

Por otro lado, en este trabajo durante la implementación de la WVD, se tuvo en cuenta un tamaño de ventana lo suficientemente grande con el fin de poder caracterizar los cambios espectrales introducidos en la señal de voz a causa de la EP, entre ellos el tremor, ya que al ser una perturbación en baja frecuencia no puede ser detectada mediante segmentos cortos, sino que implica analizar segmentos que oscilan entre los 100 y los 500 ms. La figura 2.3 presenta una espectro obtenido al analizar una frase pronunciada por una persona con voz sana y una con EP, el cual fue calculado sobre ventanas de 262 ms como se sugiere en [137], con 256 puntos en frecuencia.

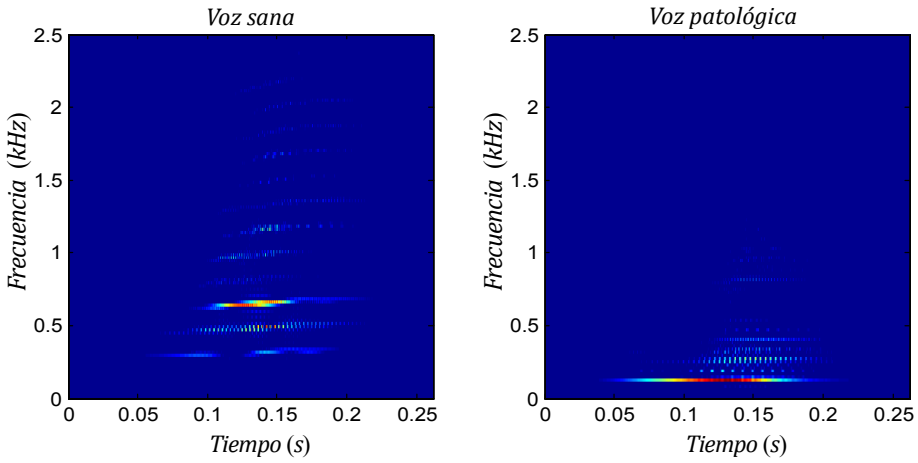

Figura 2.3: WVD para una señal de voz sana y una voz con EP.

En esta representación se visualiza fácilmente la evolución en el tiempo de los contenidos de frecuencia de la señal, aunque en esta imagen las interferencias de los términos cruzados son muy pronunciada, lo que indica que a pesar de las buenas propiedades matemáticas de la WVD no necesariamente corresponden a un gráfico limpio, ya que las componentes de la señal son bastante afectadas por los términos de interferencia. Sin embargo, se evidencia como las concentraciones de energía presentes en el espectro de las voces con EP se localizan en frecuencias más bajas que en las sanas, lo que podría ser un bio-marcador relevante para el diagnóstico y evaluación del tratamiento de los pacientes con EP.

2.2.2. Distribución pseudo Wigner-Ville (PWVD)

Las distribuciones suavizadas de WVD se idearon principalmente como un método para atenuar las interferencias, a través del ajuste de los parámetros apropiados de la distribución. Entre ellas se encuentra la pseudo WVD, que es uno de los métodos más simples para reducir las interferencias y permite representar los cambios dependientes del tiempo, además es ideal para la descripción de los fenómenos no estacionarios [132].

La PWVD consiste en una versión de la WVD en tiempo corto, mediante el uso de una ventana de análisis deslizante, la cual sirve para atenuar los términos cruzados presentes en la WVD. De esta manera, la ecuación (2.20) se multiplica por una función de ventana $g(t)$,

con el fin de concentrar el análisis de las propiedades de la señal en el tiempo. La descripción matemática de la PWVD está definida como:

$$T_{PWVD}(t,f) = \int_{-\infty}^{\infty} g(\tau) x\left(t+\frac{\tau}{2}\right) x^*\left(t-\frac{\tau}{2}\right) e^{-j2\pi f\tau} d\tau \tag{2.26}$$

La función de esta ventana superpuesta aporta a la supresión de términos cruzados de la representación WVD, realizando un gran suavizado respecto a la frecuencia, pero que al mismo tiempo produce una ligera distorsión en los términos propios. Dado que en la práctica no se tiene la capacidad para integrar desde menos infinito a más infinito, entonces se estudia la señal en un rango de tiempo limitado. Desde luego, esta mejora se produce a costa de sacrificar algunas de las propiedades que antes se cumplían y ahora ya no, como son las propiedades de los marginales (P3) y la frecuencia instantánea (P9) [133].

Adicionalmente, se presenta en la figura 2.4 una representación de la PWVD, la cual fue calculada sobre la mismas señales de voz que en la WVD y en la misma ventana de 262 ms.

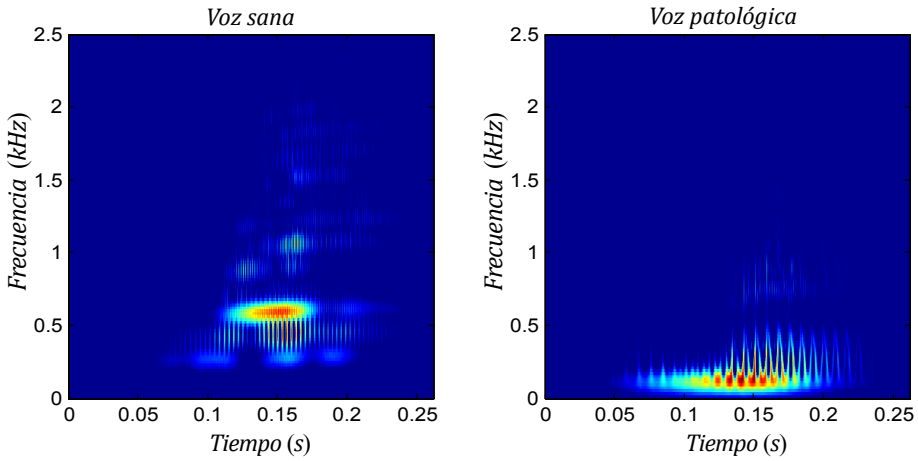

Figura 2.4: PWVD para una señal de voz sana y una voz con EP

Para su implementación se tiene en cuenta una ventana tipo Hamming de 128 puntos en el dominio del tiempo y de 256 puntos en la frecuencia.

En comparación con la WVD, se evidencia que han disminuido los artefactos e interferencias que la WVD introducía al tener por ejemplo términos negativos, mediante el suavizado respecto a la frecuencia. Lo que permite visualizar unos espectros más claros y con una mayor concentración de energía de las componentes reales de la señal de voz.

2.2.3. Distribución pseudo Wigner-Ville suavizada (SPWVD)

Otra de las representaciones suavizadas de la WVD y una de las más utilizadas en el análisis TF, es la SPWVD, la cual permite una mejora considerable en cuanto a la resolución de la PWVD, ya que la PWVD solamente realiza un suavizado en dirección de la frecuencia conservando las concentraciones presentes en el tiempo de la WVD, por lo que no se atenúan

los términos que oscilan en la dirección del tiempo. Esta restricción es superada con la SPWVD [138], la cual es una PWVD con un suavizado adicional en la dirección del tiempo.

La SPWVD es una de las representaciones más interesantes, ya que proporciona un control independiente sobre la resolución en el tiempo y la frecuencia, además es el único miembro de la clase del Cohen que hace uso de un *kernel* separable y da la posibilidad de controlar de forma independiente la cantidad de filtrado en tiempo y frecuencia. Esta representación es conceptualmente simple para reducir los efectos de las interferencias, permitiendo una elección fácil y flexible de las características de suavizado, también es considerada una implementación más eficiente. Sin embargo, no satisface la mayoría de las propiedades matemáticas que satisface la WVD a causa del suavizado que se aplica y solamente satisface algunas propiedades: es real (P1), conserva la energía (P4), y es invariante a los desplazamientos en tiempo y en frecuencia (P6) [139].

La SPWVD se define como una WVD en tiempo corto y suavizada, empleando una ventana de análisis $g(t)$ y una ventana de suavizado en el tiempo $h(t)$. El *kernel* de esta distribución es:

$$\phi(\tau,\theta) = g_1(\tau)S(\theta) \tag{2.27}$$

reemplazándolo en la ecuación (2.1), se obtiene la expresión general de la SPWVD que es definida como:

$$T_{SPWVD}(t,f) = \int_{-\infty}^{\infty} g_1(\tau)h(t)x\left(t+\frac{\tau}{2}\right)x^*\left(t-\frac{\tau}{2}\right)e^{-j2\pi f\tau}d\tau \tag{2.28}$$

donde $g_1(\tau) = g\left(\frac{1}{2}\tau\right)g^*\left(-\frac{1}{2}\tau\right)$ y $S(\theta)$ es la transformada de Fourier de $h(t)$. Ya que el *kernel* es separable, sus factores $S(\theta)$ y $g_1(\tau)$, dependen de la ventana de análisis $h(t)$ y de la ventana de suavizado $g(t)$, respectivamente. De esta forma, la cantidad de suavizado en tiempo y en frecuencia se pueden controlar de forma independiente, eligiendo la longitud de las ventanas $h(t)$ y $g(t)$, respectivamente. De esta manera, entre más larga sea $h(t)$ se produce más suavizado en el tiempo, mientras que una $g(t)$ más larga producirá un menor suavizado en frecuencia. Al utilizar ventanas convencionales como Hamming y Hanning, para $h(t)$ y $g(t)$, el suavizado de la SPWVD corresponderá a un filtro pasa-bajas bidimensional, y el *kernel* $\phi(\tau,\theta)$ será típicamente similar a una función gaussiana en dos dimensiones. Debido a que el *kernel* tiene una forma muy simple, la atenuación de términos cruzados y la concentración de energía de la distribución no son muy dependientes de los detalles de la estructura TF de la señal, lo que permite en general su aplicación a cualquier tipo de señal, independiente de la forma en la cual se encuentren distribuidos sus componentes en el plano TF [139].

Con el fin de mostrar la supresión de los términos cruzados al utilizar esta representación, en la figura 2.5 se presenta el espectro obtenido con la SPWVD, la cual es un poco más complicada en su ajuste, pues permite ajustar independientemente el ancho del *kernel* tanto en el dominio de retrasos de tiempo como en el dominio de la frecuencia. Para la representación se utilizaron las mismas señales de voz que en las distribuciones ya presentadas, además como parámetros de ajuste se utilizaron ventanas de tipo Hamming con 128 puntos en el dominio del tiempo y 256 puntos en el dominio de la frecuencia.

La representación SPWVD ofrece una mejor resolución TF, además el filtrado realizado para la supresión de los términos cruzados es muy conveniente, ya que permite obtener una

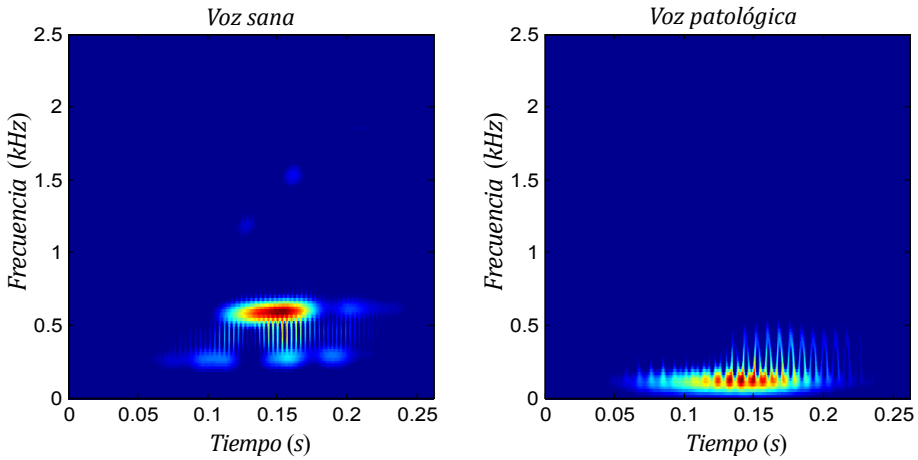

Figura 2.5: SPWVD para una señal de voz sana y una voz con EP

gráfica más clara que las obtenidas por las representaciones anteriores.

2.2.4. Distribución Choi-William (CWD)

La CWD o también llamada distribución exponencial, suaviza los términos cruzados mediante una función de tipo *kernel* gaussiano que minimiza su contribución, con el objetivo de alcanzar un compromiso entre los beneficios de la WD y del espectrograma. Esta distribución es una versión WVD suavizada que mantiene un gran número de propiedades matemáticas deseables en las distribuciones TF, aún mas que la PWVD y la SPWVD, y aún así efectúa una reducción adecuada de los términos cruzados [139]. Algunas de las propiedades que satisface la CWD son: P1, P3-P7, P9-P11.

El *kernel* de la CWD está dado por [82],

$$\phi(\tau,\theta) = e^{\frac{-\theta^2\tau^2}{\sigma}} \tag{2.29}$$

y la distribución se encuentra definida por medio de la expresión:

$$T_{CWD}(t,f;\sigma) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\sqrt{4\pi\tau^2/\sigma}} e^{-\frac{\sigma(t-u)^2}{4\tau^2}} x\left(u+\frac{\tau}{2}\right) x^*\left(u-\frac{\tau}{2}\right) e^{-i2\pi f\tau} du d\tau \tag{2.30}$$

donde σ es un parámetro positivo que controla la concentración del *kernel* alrededor del origen del plano de ambigüedad, y por tanto, la cantidad total de suavizado de la representación. Entre mayor sea σ, será más ancho el *kernel* y se efectuará una menor cantidad de suavizado. No obstante, la desventaja de la representación es que no se puede elegir de forma independiente la cantidad de suavizado en tiempo y en frecuencia, pero es mejor en cuanto a la implementación. El *kernel* de la CWD depende solamente del producto $\tau\theta$, que tiene una forma característica de cruz, y específicamente $\phi(\tau,0) = \phi(0,\theta) = 1$, lo cual indica que el *kernel* no decae en el eje θ o τ. Este comportamiento es necesario para satisfacer las propiedades de los marginales, provocando una limitación en la atenuación de términos

cruzados, especialmente en aquellas componentes que ocurren al mismo tiempo o en la misma frecuencia.

La concentración TF de la CWD depende fuertemente de la señal específica con la cual se esté trabajando, y para que sea buena, los términos de la señal deben estar en la región de paso del *kernel* en el plano de ambigüedad. Consecuentemente, la representación es muy buena tanto para señales impulsivas en el tiempo, como para señales de frecuencia constante, puesto que estas señales se encuentran concentradas en el eje τ y θ respectivamente, en los cuales $\phi(\tau,\theta) = 1$. Para el resto de señales, el *kernel* producirá un truncado de los auto-términos de la señal, lo cual corresponde a una dispersión en la CWD [139].

Al igual que en las distribuciones anteriores, se presenta una representación de la CWD de las mismas señales de voz. La figura 2.6 presenta los espectros obtenidos con la CWD y para su implementación se usó $\sigma = 0.5$ y 256 puntos en el dominio de la frecuencia.

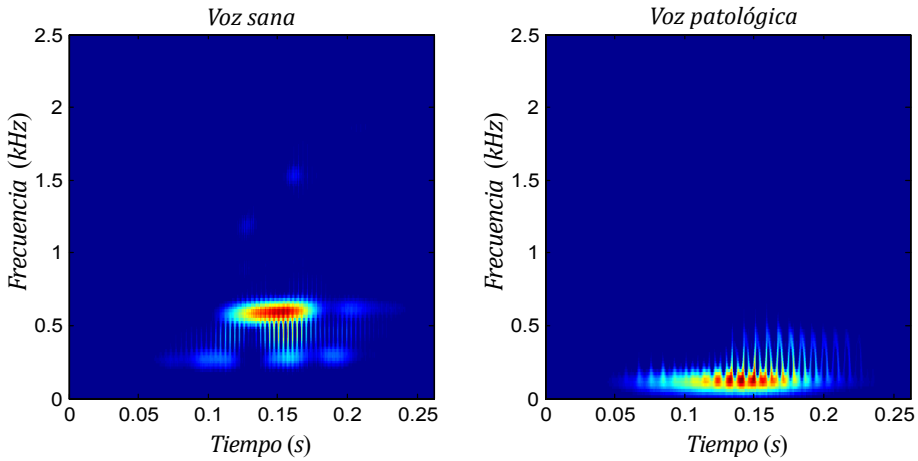

Figura 2.6: CWD para una señal de voz sana y una voz con EP.

La representación es clara y fácilmente interpretable, y presenta una buena disminución de los términos de interferencia, sobretodo en el espectro de la señal de voz con EP, se evidencia una disminución de los términos de interferencia con respecto a la SPWVD. Por lo que la CWD presenta una mejor representación en cuanto a resolución TF sobre las distribuciones de energía presentadas anteriormente; aunque debido al suavizado que realiza, la energía se dispersa en bandas en sentido del eje frecuencial, resultando estructuras en forma de hilos que en ciertas ocasiones dificultan la interpretación gráfica.

2.2.5. Transformada *wavelet* de paquetes (WPT)

La transformada *wavelet* (WT), a diferencia de las presentadas anteriormente, es una de las representaciones más utilizadas de la clase afín, las cuales se caracterizan por ser covariantes a los cambios en la escala y a traslaciones en el tiempo, permitiendo realizar un análisis multiresolución. Esta clase de distribuciones hacen parte importante de la clase lineal. La función *wavelet* es una pequeña onda cuya energía se encuentra concentrada en

el tiempo y sirve como herramienta para el análisis no estacionario, convirtiéndose en una herramienta poderosa para diversas aplicaciones en el procesamiento de señales.

El análisis wavelet se basa, al igual que la teoría de Fourier, en el concepto de aproximación de señales usando la teoría de superposición. La WT tiene ventajas sobre la transformada de Fourier tradicional, ya que las funciones *wavelet* varían tanto en frecuencia como en escala, lo que permite una buena representación para las señales no estacionarias con discontinuidades y picos intensos [140], [141].

La WT genera bloques de información en escala y tiempo de una señal, los cuales se crean desde una única función fija llamada *wavelet* madre $\psi(t)$, que se define como:

$$\psi_{a,b}(t) = \psi\left(\frac{t-b}{a}\right) \tag{2.31}$$

Se debe agregar, que la función de análisis $\psi(t)$ se escala y no se modula como la función de ventana $h(t)$ de la STFT, por lo que el análisis *wavelet* es denominado tiempo-escala y no TF.

La descripción matemática de la WT se presenta en (2.32), que se calcula como el producto interno entre la señal $x(t)$ y versiones trasladadas y escaladas de la función $\psi(t)$:

$$W_\psi x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi * \left(\frac{t-b}{a}\right) dt \tag{2.32}$$

donde a es el parámetro de escala que controla las dilataciones y contracciones de la señal, mientras b se relaciona con una traslación en el tiempo. Una variación en el retraso de tiempo b y/o en el parámetro de escala a, no tiene efecto en la forma del *kernel* de transformación de la WT, sin embargo, la resolución en tiempo y en frecuencia depende de a. Para frecuencias altas (a pequeño) se tiene una buena localización en el tiempo, pero baja resolución en frecuencia. Por otro lado, en frecuencias bajas, se tiene buena resolución en frecuencia, pero baja resolución en el tiempo.

Se debe tener en cuenta que en términos de cálculo computacional es imprescindible discretizar la transformada, y la suposición más lógica es que tanto los valores de escala como traslación sean discretos. La forma más común de discretizar los valores de a y b, es utilizar una red diádica [142], es decir, $a = 2^{-j}$ y $b = k2^{-j}$, de tal manera el conjunto de funciones con parámetros discretizados se convierte en:

$$\psi_{j,k}(t) = a_0^{-j/2}\psi(a^{-j}t - kb_0) \tag{2.33}$$

con j y k enteros que escalan y dilatan la función madre $\psi(t)$, para generar la familia de *wavelets* discretas, en donde j indica la anchura de la *wavelet* y k determina la posición. De esta manera se introduce la transformada *wavelet* discreta (DWT), que permite la descomposición de una señal en otras dos por medio de serie de filtros pasa-alto y pasa-bajo. El contenido de alta frecuencia a través de un filtro de pasa-alto se conserva como "detalles" (D), de la misma manera, el contenido de baja frecuencia se conserva como "aproximaciones" (A), y sólo las aproximaciones se pueden descomponer de forma iterativa, como se puede observar en la figura 2.7. En particular para las señales de voz, el contenido en baja frecuencia es la parte más importante, ya que proporciona la identidad de la señal, mientras que el

contenido de alta frecuencia imparte sabor o matiz [143]. Es importante tener en cuenta que si se retiran los componentes de alta frecuencia de la señal de voz, la voz sonará diferente, pero el discurso todavía puede ser entendido.

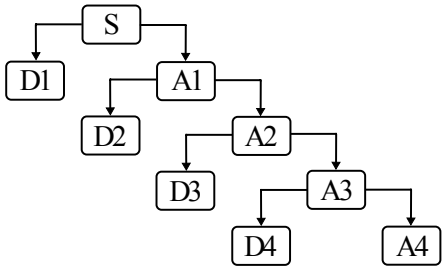

Figura 2.7: Árbol de descomposición *wavelet* con tres niveles.

Por otra parte, una nueva propuesta de la WT es la WPT, que es considerada como una versión extendida de la DWT. La WPT permite ampliar el estudio que se realiza mediante *wavelets* y posee una mejor localización en frecuencia, ya que mientras en un análisis DWT se descompone únicamente las aproximaciones en D y A, con el uso de la WPT se realiza una descomposición recursiva tanto D y A, en lugar de solamente realizar el proceso de descomposición en A. La figura 2.8 muestra la descomposición de 7 niveles, utilizando la WPT.

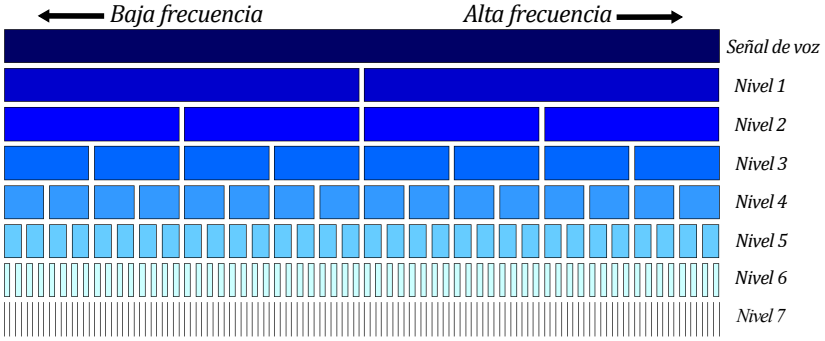

Figura 2.8: WPT convencional con 7 niveles de descomposición.

En este trabajo, las señales de voz se descomponen mediante el uso de la WPT con una *wavelet Daubechies*, la cual fue creada especialmente para el análisis de señales de voz. Durante se implementación se tuvo en cuenta varios niveles, considerando todos los coeficientes del segundo al quinto nivel de descomposición para la caracterización. Obteniendo conjuntos de características desde 4 hasta 32 sub-bandas, cabe resaltar que el número de conjuntos de características se duplica con la descomposición adicional que presenta la WPT.

2.2.6. Modulación espectral

En años recientes, una de las técnicas TF más representativa y novedosa es conocida como modulación espectral que, aunque no está relacionada directamente con las técnicas que veníamos trabajando, ha sido muy usada para la detección de voces patológicas, alcanzando altas tasas de reconocimiento, ya que permite el análisis de los cambios del espectro acústico. Los espectros de modulación (*MS – Modulation Spectra*) obtenidos mediante esta técnica, pueden ser vistos como una forma no paramétrica para representar las bandas de frecuencias dependientes de las modulaciones presentes en el discurso, introducidas por la presencia de patologías [144]. Es considerada una representación espectral de la trayectoria temporal de las características y proporciona información de los cambios lentos de la señal de voz, ofreciendo una forma compacta para fusionar los fenómenos que se presentan durante la producción del habla, que proporciona información dinámica importante y complementaria para la detección de voces patológicas [145].

Para el análisis de modulación espectral, se parte de un banco de filtros, seguido por la detección de envolvente por sub-banda y su análisis respectivo. En la figura 2.9 se muestra el proceso para la estimación los espectros de modulación.

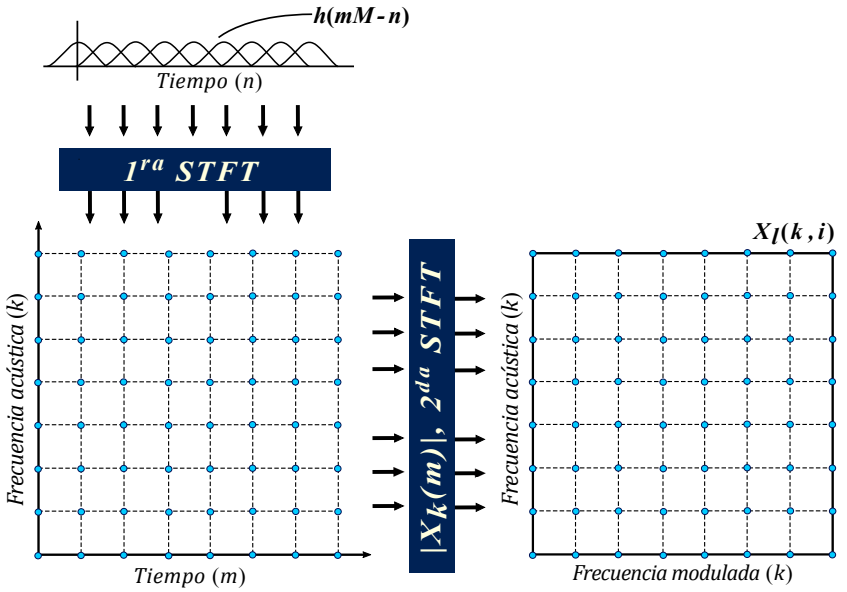

Figura 2.9: Diagrama para el cálculo de los MS.

Inicialmente, la estimación del banco de filtros se realiza mediante el uso de la STFT, y la detección de envolvente es definida como la magnitud o la magnitud al cuadrado de cada sub-banda [124], tal como se muestra en la ecuación 2.34, para una señal de tiempo discreto, $x(n)$, se emplea la STFT para una ventana de tiempo, m, la cual está dada por $X_\kappa(m)$:

$$X_\kappa(m) = \sum_{n=-\infty}^{\infty} h(mM-n)x(n)e^{-j\frac{2\pi}{K}\kappa n} \tag{2.34}$$

$$\kappa = 0, ..., K-1$$

donde $h(\cdot)$ es la ventana de análisis de las frecuencias acústicas con un tamaño de salto de M muestras y una longitud de K muestras. En esta etapa, la resolución en el dominio de la frecuencia es también de K muestras. La distribución de la envolvente de amplitud de la voz, posee una fuerte componente exponencial, por lo tanto se utiliza una normalización logarítmica de los valores de amplitud $|X_\kappa(m)|$ y se sustrae la media logarítmica de su amplitud, como se expresa en:

$$\widehat{X}_\kappa(m) = log|X_\kappa(m)| - \overline{log|X_\kappa(m)|} \tag{2.35}$$

donde $\{\bar{\cdot}\}$ denota el operador promedio aplicado sobre cada m. A continuación, se aplica una segunda $STFT$, para detectar el contenido de frecuencias de $|X_k(m)|$,

$$X_l(\kappa, s) = \sum_{m=-\infty}^{\infty} d(lL-M)|\widehat{X}_\kappa(m)|e^{-j\frac{2\pi}{S}sm} \tag{2.36}$$

$$s = 0, ..., S-1$$

y $d(\cdot)$ es denominado como la ventana de análisis de las frecuencias de modulación, y L es el tamaño de salto, en muestras; κ y s, se refieren a las frecuencias "acústicas" y de "modulación", respectivamente. Las ventanas, $h(\cdot)$ y $d(\cdot)$, se utilizan para reducir los lóbulos laterales de las dos estimaciones de frecuencia.

La representación de la modulación espectral, muestra la energía de $|X_l(\kappa, s)|$ (magnitud de la envolvente espectral de cada sub-banda) sobre el plano de frecuencias acústicas en función de las frecuencias de modulación. Posteriormente se aplica una normalización de características por sub-banda, donde cada sub-banda de frecuencias acústicas fue normalizada con la frecuencia de modulación marginal [124], como finalmente se indica en la expresión:

$$X_{l,sub}(k, i) = \frac{X_l(\kappa, i)}{\sum_i X_l(\kappa, i)} \tag{2.37}$$

De acuerdo a los estudios realizados, se ha sugerido que las frecuencias de modulación en el habla sean de $2-8Hz$, para reflejar la estructura temporal silábica y fonética. Por lo mismo, para estimar la energía en la frecuencia de modulación más sensible de la audición humana, alrededor de los $4Hz$ [145], es necesario usar una ventana para el análisis al menos de 250 ms. Para este trabajo las señales de voz se calcularon sobre ventanas de 262ms desplazadas por 64ms como se propone en [137]. Se aplicó un filtrado lineal con diferentes números de bandas, mientras que el tamaño de la transformada de Fourier para la transformación de dominio de tiempo se establece en 257. Por lo tanto, cada espectro de modulación consiste en 257 frecuencias acústicas y 257 frecuencias de modulación, lo que resulta en una imagen de 257×257, por cada ventana.

En la figura 2.10 se presentan dos MS, obtenidos para una señal de voz de una persona sana y una con EP. En ella se puede observar como la energía en las modulaciones correspondientes a la frecuencia fundamental y sus armónicos, se localiza en las frecuencias acústicas más bajas para señales de voz patológicas, mientras que para las voces sanas no se observa este comportamiento. A pesar de que esta representación no se parece a las que antes fueron

presentadas anteriormente, se sigue presentando el mismo comportamiento, en donde el espectro de las personas con EP tienen las mayores componentes de energía, en una frecuencia más baja que en las personas sanas, además no se evidencian problemas de términos cruzados.

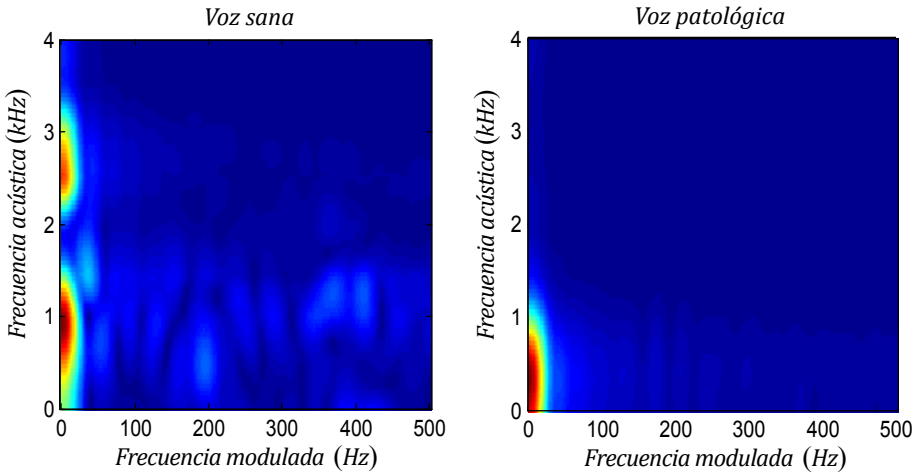

Figura 2.10: Espectros de Modulación, de una voz sana y una voz con EP

2.3. Técnicas de caracterización de las transformaciones TF

Las transformaciones TF tienen como objetivo principal obtener la mejor representación posible para la subsiguiente etapa de detección automática o clasificación de señales. Sin embargo, muchas de estas técnicas al tener una alta resolución en TF, producen un alto grado de detalle en todas las regiones del plano y consecuentemente un volumen de datos excesivo, lo cual es a veces innecesario, pues la información de la señal no se encuentra distribuida en todo el plano, sino que se concentra en determinadas regiones en donde se debe enfocar el análisis, convirtiéndose en un problema para cualquier tipo de clasificador, si no se realiza algún tipo de procesamiento de reducción de dimensión, ya que contienen mucha información que puede no ser importante para identificar diferencias entre los tipos de señales que se quieren discriminar [146].

En este sentido, una vez se tiene una representación TF que represente adecuadamente el comportamiento espectral de la señal a lo largo del tiempo, se debe reducir la dimensión de los datos. Una vía para mejorar este problema es remuestrear la representación en tiempo y en frecuencia, sin embargo, esta operación puede remover el detalle que se obtiene al utilizar una representación más compleja. Como alternativa se puede realizar una exploración para encontrar la información con mayor concentración de la señal y así enfocar el análisis en regiones específicas de la representación. De esta manera se pueden obtener características puntuales a partir de las representaciones en dos dimensiones, por este motivo, es recomendable extraer variables dinámicas con información dinámica de las distribuciones TF, es

decir, características que varían a lo largo del tiempo y que contienen menos datos que la distribución TF completa, pero con suficiente información para el proceso de interés [146].

A continuación, se presentan los métodos de estimación de características dinámicas sobre los espectros obtenidos de las distribuciones TF y la caracterización de los niveles *wavelet*, usados en este trabajo para llevar a cabo la detección de la enfermedad de Parkinson mediante el análisis de la señal de voz.

2.3.1. Energía instantánea

Teniendo en cuenta que uno de los propósitos fundamentales de las TFD, es la caracterización de la energía de la señal en sus diferentes bandas y a lo largo del tiempo, la primera opción para reducir el volumen de datos, ha sido utilizar medidas de energía sobre el espectro TF, al igual que en los niveles de la WPT.

Para entender este concepto se parte matemáticamente de un detector lineal de energía $E_L(n)$, que está constituido por un filtro lineal pasabanda seguido por el cuadrado de la magnitud, para una señal discreta $x(n)$, se puede expresar como [147]:

$$E_L(n) = \left| \sum_k x(n-k) h_L(k) \right|^2 \tag{2.38}$$

donde $h_L(k)$ es la respuesta al impulso de un filtro pasa-banda. Para una señal real, la anterior expresión se puede expandir como

$$E_G(n) = \sum_k \sum_l x(n-k-l) x(n-k+l) h_L(k+l) h_L(k-l) \tag{2.39}$$

La expresión (2.39) es una forma generalizada de expresar la energía instantánea de una señal, y a partir de ésta se pueden derivar múltiples estimadores de la energía instantánea, entre los que cuales encontramos: la energía normalizada, la energía de Shannon, el operador de Teager y la energía por filtrado homomórfico.

- **Energía normalizada:**

  La energía normalizada de la señal $x(n)$ se define mediante la expresión,

$$E_n(n) = \frac{1}{M} \sum_{m=0}^{M-1} |x(n-m)|^2 \tag{2.40}$$

- **Energía de Shannon:**

  La cual se define como la entropía de Shannon del cuadrado de la señal, y se calcula mediante la expresión [148],

$$E_S(n) = -\frac{1}{M} \sum_{m=0}^{M-1} x(n-m)^2 log\left(x(n-m)^2\right) \tag{2.41}$$

en la cual se enfatiza la intensidad media de la señal sobre amplitudes muy bajas o muy altas, de esta manera permite reducir la diferencia en la intensidad de la envolvente de la señal entre los segmentos de baja intensidad y los de alta intensidad. Al mismo tiempo, el promedio dado por la sumatoria en la ecuación (2.41), permite suavizar cambios abruptos y picos que puede presentar la señal durante el análisis [149].

- **Energía por filtrado homomórfico:**

El filtrado homomórfico es utilizado para extraer una envolvente de energía suavizada, su principal ventaja es que produce una suavidad escalable, con la cual se evitan problemas de picos separados o cortados [150]. Esta técnica involucra una transformación logarítmica, la cual convierte una combinación no lineal de señales (multiplicadas en el dominio del tiempo) en una combinación lineal. De esta forma, el espectro resultante se puede tomar como una combinación de dos componentes, una que varía lentamente y la otra que varía rápidamente, en donde se remueve esta última componente a través de un filtro pasa-bajas [151].

Si $x(n)$ representa la señal, ésta se puede expresar como un producto de una señal de baja frecuencia $\beta(n)$ y una de alta frecuencia $\alpha(n)$,

$$x(n) = \beta(n)\alpha(n) \tag{2.42}$$

Luego, la operación de multiplicación se convierte en una suma a través de una transformación logarítmica,

$$z(n) = log\, x(n) = log\, \beta(n) + log\, \alpha(n) \tag{2.43}$$

La componente de alta frecuencia se caracteriza por tener variaciones rápidas en el tiempo, así que se aplica un filtro lineal pasa-bajas $F_L\{\cdot\}$ con el fin de filtrar las componentes de $\alpha(n)$,

$$z(n) = F_L\{z(n)\} \tag{2.44}$$

Asumiendo que la transformación logarítmica no afecta la separabilidad de las componentes de Fourier de $\beta(n)$ y $\alpha(n)$ y con $F_L\{\cdot\}$ lineal se tiene,

$$z(n) = F_L\{log\, \beta(n)\} + F_L\{log\, \alpha(n)\} \approx log\, \beta(n) \tag{2.45}$$

Finalmente, aplicado la operación exponencial,

$$E_H(n) = e^{z_l(n)} \approx \beta(n) \tag{2.46}$$

La aproximación de $\beta(n)$ dada por (2.46), corresponde a la estimación de la envolvente de energía de la señal calculada a través de filtrado homomórfico [150].

- **Operador de energía de Teager (TEO):**

TEO es un operador no lineal que permite estimar la energía que consume un oscilador para generar una señal sinusoidal, teniendo en cuenta tanto su amplitud como su frecuencia, ya que se necesita más energía para generar señales de mayor frecuencia [152]. Para encontrar este operador se debe tener en cuenta, que la energía de un movimiento oscilatorio simple, es proporcional al cuadrado de la amplitud y al cuadrado de la frecuencia de oscilación.

$$E \propto A^2\omega^2 \tag{2.47}$$

Sea $x(n)$ la señal que representa el movimiento oscilatorio de un cuerpo,

$$x(n) = Acos(\Omega n + \varphi) \tag{2.48}$$

en donde $\Omega = 2\pi f / F_s$ es la frecuencia digital, siendo f la frecuencia analógica, F_s la frecuencia de muestreo y φ el ángulo de fase inicial.

Los parámetros A, Ω y φ, pueden ser estimados bajo ciertas restricciones, a partir de tres muestras de la señal $x(n)$. Por conveniencia se toman tres puntos adyacentes:

$$\begin{aligned} x(n) &= Acos(\Omega n + \varphi) \\ x(n+1) &= Acos\big((n+1)\Omega + \varphi\big) \\ x(n-1) &= Acos\big((n-1)\Omega + \varphi\big) \end{aligned} \tag{2.49}$$

Usando identidades trigonométricas se obtiene,

$$x(n+1)x(n-1) = A^2cos^2(\Omega n + \varphi) - A^2sen^2(\Omega) \tag{2.50}$$

Se observa que el primer término de la derecha corresponde al cuadrado de la señal $x(n)$, la cual se sustituye y se obtiene,

$$A^2sen^2(\Omega) = x(n)^2 - x(n+1)x(n-1) \tag{2.51}$$

Con valores pequeños de Ω, se cumple que $sen(\Omega) \approx \Omega$. Ahora, si se limita el valor de Ω a $\Omega < \pi/4$, se obtiene,

$$A^2\Omega^2 \approx x(n)^2 - x(n+1)x(n-1) \tag{2.52}$$

De esta forma la siguiente expresión constituye la energía Teager en cualquier señal de una componente, de acuerdo con lo definido por Kaiser en 1990 [152]:

$$E_T = x(n)^2 - x(n+1)x(n-1) = A^2sen^2(\Omega) \approx A^2\Omega^2 \tag{2.53}$$

2.3.2. Frecuencia instantánea

La frecuencia instantánea (FI) es una característica muy importante para señales cuyas componentes espectrales varían a lo largo del tiempo, puesto que define la ubicación del pico espectral de la señal a medida que cambia con el tiempo y, además, podría dar información puntual de la variación de cada componente.

La definición de FI fue planteada por Van der Pol, la cual puede interpretarse como la frecuencia de una onda sinusoidal que se ajusta localmente a la señal bajo análisis [153],

$$x(t) = Acos\left[\int_0^t 2\pi F_i(t)dt + \varphi_0\right] \tag{2.54}$$

en donde φ_0 es un ángulo de fase arbitraria, y la frecuencia instantánea es,

$$F_i(t) = \frac{1}{2\pi}\frac{d\varphi(t)}{dt} \tag{2.55}$$

siendo $\varphi(t)$ la fase, la cual se puede relacionar con la señal compleja o analítica, que se obtiene a partir de la señal real $x(t)$ a través de la transformada Hilbert. Mediante este procedimiento se calcula la trasformada de Fourier de la señal y se eliminan las frecuencias negativas, luego se multiplican por dos las frecuencias positivas. Lo anterior es equivalente a,

$$z(t) = x(t) + jH\{x(t)\} = A(t)e^{j\varphi(t)} \tag{2.56}$$

donde $H\{\cdot\}$ denota la transformada Hilbert de la señal.

Finalmente Ville definió la FI de la señal $x(t)$, y demostró que la frecuencia media en el espectro de la señal es igual a la media en el tiempo de la FI, como,

$$F_i(t) = \frac{1}{2\pi}\frac{d}{dt}\left[\arg z(t)\right] \tag{2.57}$$

Usando estos resultados, Ville formuló la WVD, en la cual el primer momento con respecto a la frecuencia produce la FI [68, 153]:

$$F_i(t) = \frac{\int_{-\infty}^{\infty} f T_{WVD}(t,f)df}{\int_{-\infty}^{\infty} T_{WVD}(t,f)df} \tag{2.58}$$

La importancia de la FI radica en que permite tener una medida de la ubicación en frecuencia de la concentración de energía de la señal como una función del tiempo. Esta propiedad explica la importancia de la FI en reconocimiento de señales, estimación y modelado. Sin embargo, este método solamente tiene significado físico en señales monocomponente, en donde sólo hay una frecuencia o un rango estrecho de frecuencias, las cuales varían como función del tiempo. Para señales multicomponente, la noción de una FI para cada instante de tiempo pierde significado, y se hace necesario realizar un análisis de FI de cada componente por separado [154].

Estimación de FI en TFD

Existen diferentes técnicas y metodologías para la estimación de FI en señales no estacionarias. En el caso de las señales multicomponentes existe una aproximación que se basa en el análisis de los mapas TF [153, 155], en los cuales se aprovecha la información de la transformada bidimensional, y se realiza la estimación a partir de la representación. Los métodos que utilizan esta aproximación son menos sensibles al ruido, la estimación es más robusta y además permiten ser aplicados a señales de varias componentes, lo cual puede ser una gran ventaja para gran parte de las aplicaciones. La estimación de FI a partir de las TFD se puede hacer a partir de dos aproximaciones: basada en los momentos o en los picos de las TFD. La expresión en tiempo discreto para el cálculo de la FI a partir del primer momento de la TFD es la siguiente:

$$F_i(n) = \frac{\sum_k f\, T_x(n,k)}{\sum_k T_x(n,k)} \tag{2.59}$$

donde $T_x(n,k)$ es la versión discreta de la representación TF $T(t,f)$ y n, k corresponden a los índices de tiempo y frecuencia discretos respectivamente.

Por otro lado, la estimación de FI se puede realizar a través de la detección de los picos de las TFD, ya que la estimación de máxima verosimilitud para la frecuencia de una señal sinusoidal embebida en ruido blanco gaussiano estacionario, está dada por el pico del espectro de la señal. Generalizando este resultado a señales no estacionarias, entonces la estimación de la FI se puede realizar detectando el pico o los picos de la TFD en cada instante de tiempo [155]. Las TFD en forma discreta se pueden representar como matrices bidimensionales, las cuales se pueden interpretar y analizar mediante técnicas de procesamiento de imágenes. Por tal motivo, en [156] se propone un método para la estimación de FI a partir de la representación TF de la señal, tratando la TFD como una imagen discreta. El procedimiento a seguir involucra al gradiente en el sentido del eje de la frecuencia para obtener los máximos de la representación. Posteriormente, se aplica procesamiento de imágenes con el fin de enlazar componentes y eliminar máximos espurios causados por ruido y perturbaciones.

El primer paso para la estimación de la FI es convertir la TFD en una imagen binaria; esto se hace asignando a las ubicaciones de picos locales el valor de 1 y asignando 0 a las demás. Los picos se hallan aplicando el gradiente en el sentido de la frecuencia así,

$$B(n,k) = \begin{cases} 1 \text{ si } \left[\frac{\partial T_x(n,k)}{\partial k} = 0\right] \oplus \left[\frac{\partial^2 T_x(n,k)}{\partial k^2} < 0\right] \\ 0 \text{ otro} \end{cases} \tag{2.60}$$

Si la TFD tiene varios máximos en cada instante de tiempo, solamente se elige aquel con amplitud máxima. Luego, en la imagen binaria $B(n,k)$, las componentes se enlazan usando un conjunto de 10-vecinos como se presenta en [156]. Se tiene también un umbral de longitud mínima de píxeles conectados P_L, el cual define cuál componente es verdadero, y se ajusta con el fin de remover componentes detectados de forma falsa. P_L se elige como la duración mínima para la cual un componente de la señal puede existir. Finalmente, la expresión para calcular la frecuencia instantánea es,

$$F_i(n) = \arg\max\{B(n,k)\} \tag{2.61}$$

2.3.3. Ancho de banda

La energía de las señales se puede concentrar en diversas partes del espectro, puede estar localizada en las frecuencias bajas, altas o medias. En el caso particular de las señales de tipo biológico, como la señal de voz, se clasifican como señales pasa-banda, por lo que se hace necesario expresar de forma cuantitativa el rango de frecuencias sobre el cual se concentra la densidad espectral de potencia o energía, que es conocido como ancho de banda de una señal.

Existen diversos métodos para la estimación del ancho de banda de una señal, uno de ellos es cuando la potencia del espectro decae a la mitad del máximo y también a través de la desviación estándar del espectro. Sin embargo, los métodos convencionales no pueden ser aplicados correctamente a espectros multicomponente, por lo que existen diversos métodos para su estimación como son los anchos de banda equivalentes (*EBW – Equivalent Bandwidth*) para una señal aleatoria, y los que se han usado más frecuentemente son: el EBW rectangular, el ancho de banda de Blakman-Tukey, el EBW basado en la información de Fisher de la media y la varianza, el ancho de banda de la entropía espectral de Campbell y el EBW igual al tamaño del soporte del espectro [157]. Los EBW de procesos aleatorios fueron unificados y se pueden representar a través de la entropía de Rènyi de orden α, definido con base en la teoría de información y más precisamente a través de una medida de entropía. Puesto que el EBW indica el número efectivo de variables no correlacionadas por unidad de tiempo o la tasa de coeficientes de una señal aleatoria [157].

Para un proceso aleatorio estacionario $x_s(t)$ con espectro de potencia $X_s(f)$, el EBW de orden α se define como [158],

$$BW_s^{(\alpha)} = \frac{1}{2}\left[\int_{-\infty}^{\infty} X_s^{\alpha}(f)df\right]^{\frac{1}{1-\alpha}} \tag{2.62}$$

en donde el espectro de potencia $X_s(f)$ se encuentra normalizado para que el área bajo la curva sea igual a uno, es decir, $\int_{-\infty}^{\infty} X_s(f)df = 1$.

En el caso de una TFD de un proceso aleatorio no estacionario, por analogía con (2.62), el ancho de banda equivalente variante en el tiempo de orden α asociado al proceso $x(t)$ está dado por [158],

$$BW^{(\alpha)}(t) = \frac{1}{2}\left[\int_{-\infty}^{\infty} T_x^{\alpha}(f|t)df\right]^{\frac{1}{1-\alpha}} \tag{2.63}$$

en donde la densidad de probabilidad condicional $T_x(f|t)$ se define como,

$$T_x(f|t) = \frac{T_x(t,f)}{\int_{-\infty}^{\infty} T_x(t,f)df} \tag{2.64}$$

Consecuentemente, para una TFD no negativa que satisfaga los marginales de la señal, es decir, si la distribución se puede considerar como una función de probabilidad, es posible estimar el EBW de un proceso aleatorio no estacionario. Si $\alpha = 2$ la ecuación (2.63) corresponde al segundo ancho de banda equivalente de la densidad espectral blanca, de banda

limitada, la cual satisface

$$T_x(f|t) = \begin{cases} \xi, & |f| \leq BW^2(t) \\ 0, & |f| > BW^2(t) \end{cases} \tag{2.65}$$

donde ξ es una constante. La implementación de (2.63) se puede realizar a través de sumatorias en los dominios discretos de tiempo y de frecuencia,

$$BW^{(\alpha)}(t) = \frac{1}{2}\left(\sum_k T_x^{\alpha}(k|n)\right)^{\frac{1}{1-\alpha}} \tag{2.66}$$

donde,

$$T_x(k|n) = \frac{T_x(n,k)}{\sum_k T_x(n,k)} \tag{2.67}$$

2.3.4. Centroides de subbanda espectral

Otro tipo de técnicas de caracterización de los espectros, son los conocidos centroides de subbanda espectral [159], que combinan eficientemente la frecuencia y la magnitud de la información del espectro de potencia, para representar correctamente señales no estacionarias.

Con el fin de definir los centroides de sub-banda espectral y la energía alrededor de las diferentes bandas de frecuencias de los espectros, se divide la banda de frecuencia en un número fijo de sub-bandas, desde 0 a $F_s/2$, donde F_s es la frecuencia de muestreo y se calcula el centroide de cada sub-banda usando el espectro de potencia de la señal. Esta definición involucra la elección de diversos parámetros, entre los cuales se encuentran: la forma en la cual se divide el espectro en sub-bandas, es decir, las frecuencias centrales y de corte de cada filtro y el solapamiento, si es que es necesario, la forma del filtro, entre otros.

Asumiendo que el espectro de frecuencia se divide en V sub-bandas, utilizando los filtros $H_v(f)$, se define el $v-$ésimo centroide de subbanda espectral C_v [160], como se presenta en la siguiente ecuación:

$$C_v = \frac{\int_{-\infty}^{\infty} f H_v(f) X^{\gamma}(f) df}{\int_{-\infty}^{\infty} H_v(f) X^{\gamma}(f) df} \tag{2.68}$$

en donde $X(f)$ es el espectro de potencia y γ es una constante que controla su rango dinámico. Ajustando $\gamma < 1$, el rango dinámico del espectro de potencia se puede reducir, y si $\gamma \to 0$, el centroide se ubicará en el centro de la subbanda y no contendrá información alguna. Por otro lado si $\gamma \to \infty$, el centroide corresponderá a la ubicación del pico mayor del espectro de potencia en la subbanda, y puede producir estimados ruidosos [161].

Los filtros $H_v(f)$ se pueden distribuir linealmente a lo largo del dominio de la frecuencia, o se pueden ubicar de acuerdo a escalas de frecuencia perceptuales, tales como la escala Bark o la escala Mel. En cualquier caso, es importante que la distribución de los puntos en el dominio de la frecuencia esté de acuerdo con la distribución de los filtros, con el fin de obtener histogramas no sesgados [161].

La ecuación 2.68 se puede generalizar para el caso no estacionario, e implementar en tiempo discreto mediante la siguiente expresión,

$$C(n) = \frac{\sum_w w H_v(w) T_x^{\gamma}(n,k)}{\sum_w H_v(w)] T_x^{\gamma}(n,w)} \tag{2.69}$$

donde $H_v[w]$ es la versión discreta del banco de filtro $H_v(f)$. Aunque la ubicación de los centroides puede ser una característica relevante para diferenciar señales pertenecientes a diferentes clases, también puede ser importante tener en cuenta la energía que se encuentra concentrada alrededor de dichos centroides en un ancho de banda fijo, y menor al ancho de banda del filtro $H_v(f)$ correspondiente al centroide bajo análisis. Así, la energía alrededor de los centroides se puede calcular como,

$$\xi(n) = \sum_{W=C_v(n)-\Delta w}^{C_v(n)+\Delta w} T_x(n,w), \qquad 1 \leq v \leq V \tag{2.70}$$

donde Δw corresponde al ancho de banda alrededor del centroide sobre el cual se calcula la energía.

2.3.5. Coeficientes cepstrales

Los coeficientes cepstrales han sido comúnmente utilizados en algunas tareas de reconocimiento de señales [162], donde el cepstrum es una transformación homomórfica, la cual permite separar la respuesta de un filtro de la señal de entrada, teniendo solamente a disposición la señal de salida. Una transformación homomórfica $\widehat{x}(n) = D\{x(n)\}$ es una transformación que convierte una convolución $x(n) = e(n) * h(n)$ en una suma $\widehat{x}(n) = \widehat{e}(n) + \widehat{h}(n)$. El cepstrum real de una señal digital $x(n)$ se define como,

$$\zeta(n) = \int_{-\infty}^{\infty} \ln\left|X(f)\right| e^{j2\pi f n} df \tag{2.71}$$

Si la señal $x(n)$ es real, entonces su cepstrum también será una señal real [162].

Es posible extender la definición de coeficientes cepstrales en (2.71) para que pueda ser aplicada a señales no estacionarias a través de representaciones TF. En primer lugar, se divide el espectro de la señal en V sub-bandas a través de los filtros pasa-banda $H_v[w]$, los cuales pueden estar distribuidos de forma lineal y son conocidos como coeficientes cepstrales de frecuencia lineal (*LFCC*), o según alguna escala perceptiva,

$$\zeta_v(n) = \log\left(\sum_w T_x(n,w) H_V(w)\right), \quad 1 \leq v \leq V \tag{2.72}$$

La transformada inversa de Fourier en (2.71) se convierte en transformada discreta del coseno en (2.73), ya que $\zeta_v(n)$ es una función par en v, pues proviene de la representación espectral de una señal real. Así, se obtienen P coeficientes cepstrales en cada instante de tiempo n

mediante la expresión,

$$\zeta_p(n) = \sum_{v=1}^{V} \zeta_v(n) \cos\left[p\left(v - \frac{\pi}{2P}\right)\right], \quad 1 \leq p \leq P \tag{2.73}$$

Es importante notar que la representación (2.73) no es una transformación homomórfica, la cual sería homomórfica si se invirtiera el orden de la sumatoria y del operador de logaritmo en (2.72)

$$\zeta_v(n) = \sum_{w} \log\left(T_x[n,w] H_v[w]\right), \quad 1 \leq v \leq V \tag{2.74}$$

Coeficientes cepstrales en la escala de Mel (MFCC)

Los MFCC son una representación definida como el cepstrum real de una señal por ventanas de corta duración derivadas del espectro de la FFT. La ventaja principal de los parámetros MFCC es que no requieren una estimación del tono. Este es un problema común con la mayoría de los parámetros acústicos que se encuentran en el estado del arte [34].

Para el cálculo de los MFCC se aplican las expresiones (2.72) y (2.73), pero los filtros $H_v(w)$ se distribuyen de acuerdo a una escala no lineal, la cual se aproxima al comportamiento del sistema auditivo, y permite su aplicación a señales cuyas componentes de frecuencia se encuentran dentro del rango auditivo humano [38].

La expresión general de la distribución uniforme en la escala Mel de los filtros, está dada por:

$$f_{Mel}(n) = \frac{K}{Fs} B^{-1}\left[B(f_l + n\frac{B(f_h - B(f_l))}{N+1}\right] \tag{2.75}$$

donde la escala Mel es definida como,

$$Mel = 2595 log_{10}\left(700 + \frac{f(Hz)}{700}\right) \tag{2.76}$$

Capítulo 3

Detección de Parkinson en voz

Este capítulo se refiere a la etapa de clasificación para la detección de la enfermedad de Parkinson en señales de voz, como se presenta en la figura 3.1. La detección de patologías de la voz se ha asumido típicamente como una tarea similar a la verificación de hablante, por tanto, las técnicas que se han utilizado son muy similares. Aquí, se presentan las técnicas utilizadas y las metodologías de clasificación desarrolladas, así como las estrategias empleadas para combinar la información obtenida de los espectros con cada una de las representaciones TF descritas en el capítulo anterior.

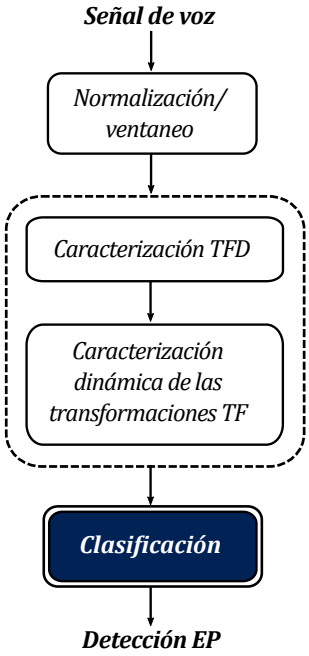

Figura 3.1: Diagrama general de un sistema para la detección automática de las señales de voz con EP.

3.1. Estrategias básicas de clasificación

En el capítulo anterior, se presentó el primer enfoque para tomar una decisión acerca de la presencia o ausencia de la patología, el cual está relacionado con el uso de un conjunto de características, que conforman un espacio de representación como entrada de un sistema de clasificación, y que proporcionará una función que representa un límite de decisión entre las diferentes clases.

En este trabajo se utilizó para la etapa de clasificación, técnicas que han sido usadas clásicamente en detección de patologías, como son las técnicas basadas en GMM y la SVM, ya que presentan muy buenas capacidades de modelado. Los clasificadores basados en GMM se ajustan a la distribución de los datos observados por medio de un conjunto de funciones gaussianas ponderadas, las ventajas de utilizarlos es que son de bajo costo computacional y tienen una capacidad de modelar las distribuciones arbitrariamente complejas de múltiples maneras [5], siendo ampliamente utilizados para la detección de voces patológicas [37]. Por otro lado, la SVM al realizar un mapeo no lineal, maximiza las capacidades de generalización del clasificador, y es una de las técnicas que mejores tasas de reconocimiento han alcanzado en el estado del arte [163].

3.1.1. Modelos de mezclas gaussianas

Un modelo GMM es una función de densidad de probabilidad paramétrica representada como una suma ponderada (mezcla) de las densidades de las componentes gaussianas. Los GMM se utilizan comúnmente como un modelo paramétrico de la distribución de probabilidad de las mediciones continuas o características en un sistema biométrico, como el tracto vocal, relacionadas con características espectrales en un sistema de reconocimiento de hablante, el cual se considera un problema muy similar al tratado en este trabajo [5].

Un GMM es una suma ponderada de M componentes de las densidades gaussianas, dada por la ecuación (3.1):

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i N(\mathbf{x}|\mu_i, \Sigma_i) \tag{3.1}$$

donde $\mathbf{x}$ es un vector de características de dimensión D, wi, $i = 1,\ldots,M$ son los pesos de las mezclas gaussianas que representan la probabilidad de cada distribución gaussiana, y $N(\mathbf{x}|\mu_i, \Sigma_i)$ son las densidades de componentes gaussianas de la forma:

$$N(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)\right\} \tag{3.2}$$

con un vector de medias μ_i y matriz de covarianza Σ_i. Los pesos de las mezclas deben satisfacer las limitaciones $w_i \geq 0$ y $\sum_{i=1}^{M} w_i = 1$. El GMM completo está parametrizado por los vectores de medias, matrices de covarianza y pesos de las mezclas de todas las componentes de las densidades gaussianas. Estos parámetros están representados colectivamente por la notación,

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, \quad i = 1,\ldots,M \tag{3.3}$$

Hay diversas variantes de los GMM, en relación a la ecuación anterior. Las matrices de covarianza pueden ser de rango completo o diagonal. Utilizando la matriz diagonal, se trabaja sólo con un vector que contiene la varianza, esto se puede hacer porque la combinación lineal de la diagonal de la matriz de covarianza es capaz de modelar la correlación entre los elementos de los vectores de características. Además, el efecto de utilizar un conjunto de M matrices de covarianza completas, puede obtenerse igualmente al usar un conjunto mayor de gaussianas de covarianza diagonal [164]. Adicionalmente, los parámetros pueden ser compartidos entre las componentes gaussianas, utilizando una matriz de covarianza común para todas las componentes. La elección de la configuración del modelo (número de componentes, matriz de covarianza y el parámetro vinculante), a menudo está determinada por la cantidad de datos disponibles para la estimación de los parámetros de los GMM y cómo los GMM se utilizan en una aplicación en particular [5].

- **Estimación de máxima verosimilitud**

Hay varias técnicas disponibles para la estimación de los parámetros de un GMM. Sin embargo, tradicionalmente el método más empleado y bien establecido es el de la estimación de la máxima verosimilitud (*ML − Maximum Likelihood*). El objetivo de la estimación ML es encontrar los parámetros del modelo que maximizan la probabilidad conjunta del GMM, dados los datos de entrenamiento, que se suponen son independientes e idénticamente distribuidos (*iid − Independent and Identically Distributed*).

Dado un conjunto $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, con una secuencia de N vectores de entrenamiento, la probabilidad de los GMM puede ser descrita como:

$$p(X|\lambda) = \prod_{t=1}^{N} p(\mathbf{x}_t|\lambda) \tag{3.4}$$

Aunque esta expresión no es una función lineal de los parámetros λ y la maximización directa no es posible. La probabilidad conjunta puede ser maximizada con un procedimiento de actualización simple y eficiente llamado algoritmo de Máxima-Esperanza (*EM − Expectation Maximization*) [163]. La idea básica del algoritmo EM es comenzar con un modelo inicial λ, para calcular un nuevo modelo $\overline{\lambda}$, de tal manera que se presente la relación $p(X|\overline{\lambda}) \geq p(X|\lambda)$. El nuevo modelo se convierte entonces en el modelo inicial para la siguiente iteración, y el proceso se repite hasta que se alcanza el umbral de convergencia.

El algoritmo EM es un método eficaz para la búsqueda de soluciones ML, en aquellos modelos con variables latentes. Si se conocen las variables latentes, que en el caso de los GMM son los pesos w_i, la estimación de los parámetros restantes sería sencilla. El logaritmo de la función de verosimilitud utilizando por el algoritmo EM para un GMM está dado por:

$$\mathscr{L}(\lambda) = \sum_{t=1}^{N} \log\left(\sum_{i=1}^{M} w_i N(\mathbf{x}_t|\mu_i, \Sigma_i)\right) \tag{3.5}$$

el $\log(\mathbf{x})$ es una función estrictamente creciente y el valor de λ que maximiza $p(X|\lambda)$ también maximiza $\mathscr{L}(\lambda)$. Con el objetivo de maximizar la probabilidad, luego de aplicar el logaritmo,

se deriva la función de la expresión (3.5) con respecto a la media μ_i y se iguala el otro lado de la función a cero [165]:

$$0 = -\sum_{t=1}^{N} \underbrace{\frac{w_i N(\mathbf{x}_t|\mu_i,\Sigma_i)}{\sum_k w_k N(\mathbf{x}_t|\mu_k,\Sigma_k)}}_{\gamma(z_{ti})} \Sigma_i(\mathbf{x}_t - \mu_i) \tag{3.6}$$

siendo $\gamma(z_{ti})$ la probabilidad condicional de la mezcla i dada la muestra $\mathbf{x}_t$. La estimación γ corresponde a la etapa de expectativa o esperanza del algoritmo EM. Multiplicando por Σ_i (la cual se supone que no es singular), y reordenando la ecuación (3.6) es posible obtener:

$$\mu_i = \frac{1}{N_i}\sum_{t=1}^{N} \gamma(z_{ti})\mathbf{x}_t \tag{3.7}$$

donde,

$$N_i = \sum_{t=1}^{N} \gamma(z_{ti}) \tag{3.8}$$

N_i puede ser interpretado como el número efectivo de puntos asignados a la agrupación i [165]. Ajustando la derivada de la ecuación (3.5) con respecto a Σ_i en cero y siguiendo un procedimiento similar al que se realizó antes, es posible obtener:

$$\Sigma_i = \frac{1}{N_i}\sum_{t=1}^{N} \gamma(z_{ti})(\mathbf{x}_t - \mu_i)(\mathbf{x}_t - \mu_i)^T \tag{3.9}$$

Finalmente, es necesario maximizar la ecuación (3.5) con respecto a los pesos w_i, pero teniendo en cuenta la restricción $\sum_{i=1}^{M} w_i = 1$. Esto se puede lograr usando un multiplicador de Lagrange y maximizando la siguiente cantidad:

$$\mathscr{L}(\lambda) + \kappa\left(\sum_{i=1}^{M} w_i - 1\right) \tag{3.10}$$

lo cual da,

$$0 = \sum_{t=1}^{N} \frac{N(\mathbf{x}_t|\mu_i,\Sigma_i)}{\sum_k w_k N(\mathbf{x}_t|\mu_k,\Sigma_k)} + \kappa \tag{3.11}$$

Multiplicando ambos lados por w_i y sumando sobre i, haciendo uso de la restricción anterior, es posible encontrar $\kappa = -N$. Usando esto para eliminar κ y reordenando, los pesos w_i pueden ser actualizados empleando [165]:

$$w_i = \frac{N_i}{N} \tag{3.12}$$

de esta manera se termina el paso de maximización del algoritmo de EM.

Una vez los parámetros GMM han sido calculados, el sistema de detección es un clasificador sencillo de máxima verosimilitud. Para que cada clase pueda ser reconocida (patológica

y/o normal), se estiman diferentes parámetros del GMM: λ_1 que corresponde a la clase patológica y λ_2 a la clase normal. Por lo tanto, para la evaluación se lleva acabo el cociente de probabilidad, en el que para cada GMM se estima la probabilidad a posteriori de una secuencia características particular $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Aplicando la regla de Bayes y descartando constantes de probabilidad a priori, la razón o cociente de verosimilitud en el dominio logarítmico, se convierte en [164]:

$$\Lambda(X) = \log p(X|\lambda_1) - \log p(X|\lambda_2) \tag{3.13}$$

El cociente de probabilidad se compara con un umbral T_d con el fin de tomar una decisión acerca de la presencia o ausencia de la patología, el cual mide esencialmente la cantidad de modelos pertenecientes a la clase patológica. Vale la pena destacar, que los términos de la razón del logaritmo de la verosimilitud deben ser calculados como:

$$\log p(X|\lambda) = \frac{1}{N}\sum_{t=1}^{N} \log p(\mathbf{x}_t|\lambda) \tag{3.14}$$

Donde se utiliza $\frac{1}{N}$ como factor de escala para normalizar la probabilidad con respecto a la duración de la señal de voz, evitando un posible sesgo debido a las diferentes longitudes de las grabaciones de la clase patológica y la clase normal.

3.1.2. Modelos de mezclas gaussianas adaptados con modelos universales

Los GMM adaptados con modelos universales, se conocen como *GMM-UBM* (*UBM - Universal Background Model)*, los cuales fueron introducidos por Reynolds para ser usados en problemas de reconocimiento de hablante y desde entonces se han venido utilizando para diferentes problemas tanto en verificación de hablante como en detección de patologías [5].

Un modelo universal (UBM) es un modelo GMM entrenado a partir de conjuntos de muestras de diferentes clases (patológica y/o normal), con el fin de representar el comportamiento general de "toda" la población. Para entender el desarrollo y el uso de un UBM, primero se debe describir la razón de verosimilitud que se debe llevar a cabo. Teniendo en cuenta un segmento de voz O y una clase hipotética S, la tarea de detección consiste en determinar si O pertenece a la clase S. Esta tarea puede ser reformulada como una decisión entre dos hipótesis:

- H_0 : O pertenece a la clase S
- H_1 : O no pertenece a la clase S (hipótesis alternativa)

Para decidir entre estas dos hipótesis, se utiliza una razón de probabilidad dada por:

$$\frac{p(O|H_0)}{p(O|H_1)} \begin{cases} \geq \Theta \text{ Acepta } H_0 \\ < \Theta \text{ Rechaza } H_0 \end{cases} \tag{3.15}$$

donde $p(O|H_0)$ es la función de densidad probabilidad (verosimilitud) para evaluar la hipótesis H_0 en la observación de la voz O y para la hipótesis H_1 es $p(O|H_1)$. El umbral de decisión

para aceptar o rechazar H_0 es Θ.

El primer paso en un sistema de reconocimiento es extraer las características de los segmentos de voz, que transmiten la información dependiente de la persona y la clase a la cual pertenece (patológica y normal), como las medidas espectrales relacionadas con el tracto vocal. La salida de esta etapa es típicamente una secuencia de vectores de características que representan los segmentos de voz, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Los vectores de características se utilizan entonces para calcular las probabilidades de H_0 y H_1, en donde, H_0 está representada por un modelo denotado como λ_p, que caracteriza la distribución de características derivadas de O en el espacio de características de X. Se puede suponer que un GMM representa lo mejor posible la distribución de vectores de características para H_0, de modo que λ_p contiene los parámetros del modelo (vector de pesos, vector de medias y matriz de covarianza), y el modelo $\lambda_{\overline{p}}$ representa la hipótesis alternativa H_1. Dicho lo anterior, la razón de verosimilitud está dada por:

$$LR(X) = \frac{p(X|\lambda_p)}{p(X|\lambda_{\overline{p}})} \tag{3.16}$$

En el dominio logarítmico, se convierte en:

$$\Lambda(X) = \log p(X|\lambda_p) - \log p(X|\lambda_{\overline{p}}) \tag{3.17}$$

Mientras que el modelo para H_0 está bien definido y se puede estimar a partir de muestras de entrenamiento de S, el modelo de S no está tan bien definido, ya que potencialmente debe representar todo el espacio de posibles alternativas a la clase S.

$$p(X|\lambda_{\overline{p}}) = F\left(p(X|\lambda_1), \ldots, p(X|\lambda_N)\right) \tag{3.18}$$

donde $F(\cdot)$, es el promedio de los valores de probabilidad del conjunto de "toda" la población del UBM.

Teniendo en cuenta los datos para entrenar un UBM, hay muchos enfoques que se pueden utilizar para obtener el modelo final. El primero, consiste en entrenar de manera individual un UBM sobre cada una de las clases, como se muestra en la figura 3.2a). Este método tiene la ventaja que permite utilizar datos no balanceados y se puede controlar el proceso final de composición del UBM. Otro método para generar el UBM y el utilizado en este trabajo, consiste en unir todos los datos de la población (patológica y normal), para entrenarlos posteriormente mediante el algoritmo EM, como se presenta en la figura 3.2b). Es muy importante que, a la hora de agrupar los datos de los subconjuntos de diferentes clases, éstos estén balanceados [5].

- **Adaptación MAP**

La idea básica de GMM-UBM, consiste en obtener el modelo del conjunto de datos (patológicos y normales) mediante la adaptación de los parámetros del UBM. A este entrenamiento se le llama *adaptación Bayesiana* o estimación MAP (*Maximum A Posteriori*). El cual es diferente al acercamiento estándar de maximizar la verosimilitud de un modelo para que la clase sea independiente del UBM, este enfoque llamado adaptación MAP, consiste en obtener

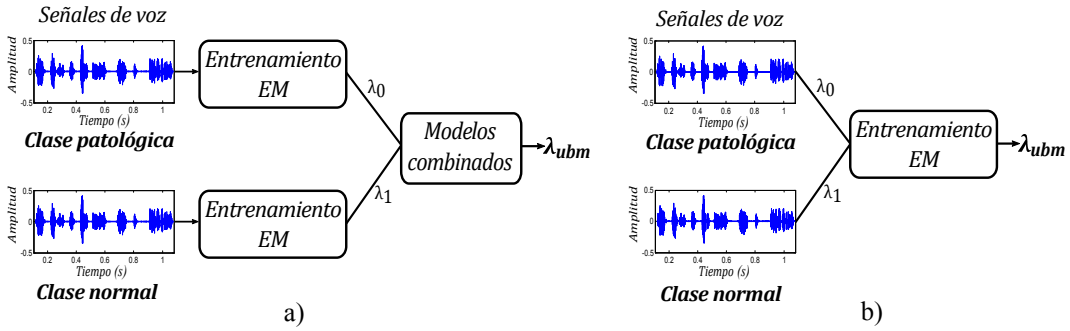

Figura 3.2: Enfoques del modelo UBM. a) Los modelos de cada clase son entrenados de manera individual y luego agrupados para crear un UBM final. b) Los datos de las clases se reúnen antes del entrenamiento del UBM mediante el algoritmo EM [5].

el modelo del conjunto de datos mediante la actualización de los parámetros entrenados en el UBM, lo que proporciona una conexión más estrecha entre el modelo de la clase y el UBM, que no sólo genera un mejor funcionamiento que en los modelos independientes, sino que también tiene en cuenta una técnica rápida para calcular la verosimilitud [5].

Como el algoritmo de EM, la adaptación MAP es un proceso de estimación de dos etapas. El primer paso es idéntico al primero del algoritmo EM, donde se calculan las estadísticas de los datos de entrenamiento de la clase para cada mezcla en el UBM. La diferencia se encuentra en el segundo paso, ya que la adaptación de las "nuevas" estimaciones estadísticas de los parámetros del modelo de la clase se combinan con las estimaciones de los parámetros del UBM (anteriores). La combinación se realiza mediante un coeficiente, el cual tiene que garantizar que las mezclas con gran cantidad de datos de la clase se basen en las "nuevas" estadísticas, y que las mezclas con poca cantidad de datos de la clase, se base en las estadísticas "anteriores", para la estimación de los parámetros finales.

Los detalles de la adaptación son los siguientes: dado un modelo anterior y los vectores de entrenamiento, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, en primer lugar se determina la alineación probabilística de los vectores de entrenamiento en los componentes de la mezcla anterior (prior) como se presenta en la figura 3.3a.

Es decir, para la mezcla i en el UBM, se calcula:

$$Pr(i|\mathbf{x}_t, \lambda_{\text{prior}}) = \frac{w_i N(\mathbf{x}_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k N(\mathbf{x}_t|\mu_k, \Sigma_k)} \tag{3.19}$$

A continuación, se utiliza $Pr(i|\mathbf{x}_t, \lambda_{\text{prior}})$ y $\mathbf{x}_t$ para calcular las estadísticas para el peso, la media y la varianza de los parámetros:

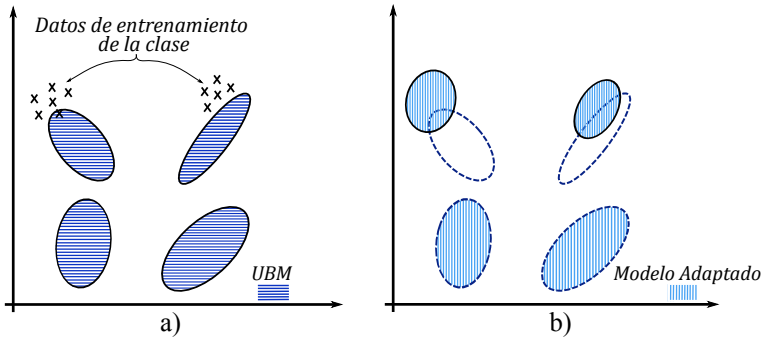

Figura 3.3: Ilustración de los dos pasos en la adaptación MAP. a) Los vectores de entrenamiento (x) son probabilísticamente proyectados en las mezclas del modelo UBM. b) Los parámetros de la mezcla adaptados se derivan utilizando las estadísticas de los nuevos datos y los parámetros UBM [5].

$$
\begin{aligned}
&\rightarrow \text{Pesos de las mezclas:} && n_i = \sum_{t=1}^{T} Pr(i|\mathbf{x}_t, \lambda_{\text{prior}}) \\
&\rightarrow \text{Medias:} && E_i(\mathbf{x}) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|\mathbf{x}_t, \lambda_{\text{prior}}) \mathbf{x}_t \\
&\rightarrow \text{Varianzas:} && E_i(\mathbf{x}^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|x_t, \lambda_{\text{prior}}) \mathbf{x}_t^2
\end{aligned}
\tag{3.20}
$$

Hasta aquí se ha realizado el primer paso, que es el mismo que el de "esperanza" en el algoritmo EM. Finalmente, los nuevas estadísticas a partir de los datos de entrenamiento se utilizan para actualizar los parámetros "anteriores" del UBM de la $i-$ésima mezcla, para crear los parámetros adaptados de la mezcla i (figura 3.3b), de acuerdo a las siguientes expresiones:

$$
\begin{aligned}
&\rightarrow \text{Pesos de las mezclas adaptadas:} && \hat{w}_i = \left[\frac{\alpha_i^w n_i}{T} + (1 - \alpha_i^w) w_i \right] \gamma \\
&\rightarrow \text{Medias adaptadas:} && \hat{\mu}_i = \alpha_i^m E_i(\mathbf{x}) + (1 - \alpha_i^m) \mu_i \\
&\rightarrow \text{Varianzas adaptadas:} && \hat{\sigma}_i^2 = \alpha_i^v E_i(\mathbf{x}^2 + (1 - \alpha_i^v)(\sigma_i^2 - \mu_i^2) - \hat{\mu}_i^2
\end{aligned}
\tag{3.21}
$$

Los coeficientes de adaptación que controlan el equilibrio entre las estimaciones anteriores y nuevas son, $\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$ para los pesos, las medias y las varianzas, respectivamente. El factor de escala γ es utilizado en los pesos de las mezclas adaptadas para asegurar que su suma sea igual a uno. Para cada mezcla y cada parámetro, un coeficiente de adaptación dependiente de los datos, α_i^ρ, $\rho \in \{w, m, v\}$, se utiliza en las ecuaciones de anteriores y se define como:

$$
\alpha_i^\rho = \frac{n_i}{n_i + r^\rho} \tag{3.22}
$$

donde r^ρ, es un factor de relevancia. Si un componente de las mezclas tiene una probabilidad n_i baja en los nuevos datos, entonces $\alpha_i^\rho \rightarrow 0$, causando poco énfasis de los nuevos parámetros (potencialmente no entrenados) y enfatizando en los anteriores (parámetros mejor entrenados), mientras que cuanto las probabilidades son altas $\alpha_i^\rho \rightarrow 1$, se causa un

mayor énfasis en los nuevos parámetros y poco en los anteriores.

Es común en aplicaciones de reconocimiento de hablante, utilizar un coeficiente de adaptación para todos los parámetros. Reynolds utiliza, en un sistema GMM-UBM un sólo coeficiente de adaptación para todos los parámetros [5],

$$\alpha_i^w = \alpha_i^m = \alpha_i^v = \frac{n_i}{n_i + r^\rho} \tag{3.23}$$

con un factor de relevancia $r = 16$. Cabe tener en cuenta que la razón de verosimilitud para los vectores de características X se calcula como:

$$\Lambda(X) = \log p(X|\lambda_p) - \log p(X|\lambda_{ubm}) \tag{3.24}$$

Puesto que la adaptación es dependiente de los datos y no todas las gaussianas en el UBM son adaptadas durante el entrenamiento del conjunto de datos de las diferentes clases. El hecho de que el modelo de la clase patológica y/o normal es adaptado del UBM, permite que sea un método más rápido para calcular la verosimilitud que la evaluación estándar de los GMM. Debido a que, cuando un GMM grande se evalúa para una matriz de características, sólo algunas de las mezclas contribuyen significativamente al valor de la verosimilitud, ya que el GMM representa una distribución sobre un espacio grande, pero es un sólo vector el que estará cerca de algunos componentes del GMM. Así, los valores de la verosimilitud se pueden aproximar muy bien usando solamente las mejores componentes de las mezclas M [5].

3.1.3. Máquina de soporte vectorial

La SVM es un potente clasificador discriminativo utilizado en reconocimiento de patrones. Formalmente, la SVM es una máquina de decisión que proporciona un hiperplano de separación óptimo para alcanzar un margen máximo de separación entre las clases, de esta manera cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas en una u otra clase.

El concepto de SVM, parte del problema de clasificación de dos clases usando modelos lineales de la forma [165]:

$$f(x) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{3.25}$$

donde $\phi(\mathbf{x})$ denota una transformación fija en el espacio de características X, y b el parámetro de sesgo. Dado un conjunto de datos de entrenamiento $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, $y_i \in \{-1, 1\}$, el objetivo es encontrar una función $f(\mathbf{x})$ que separe las muestras positivas de la negativas (hiperplano de separación). Los puntos $\mathbf{x}$ que se encuentran en el hiperplano satisfacen $\mathbf{w}^T \phi(\mathbf{x}) + b = 0$, donde $\mathbf{w}$ es normal al hiperplano, $b/\|\mathbf{w}\|$ es la distancia mínima desde el hiperplano al origen y $\|\mathbf{w}\|$ es la norma euclidiana de $\mathbf{w}$.

Suponiendo que el conjunto de datos de entrenamiento es linealmente separable en el espacio de características, por definición, existe al menos una elección de los parámetros $\mathbf{w}$ y b, tales que una función de la forma (3.25) satisface $f(\mathbf{x}_i) > 0$ para los puntos que tienen

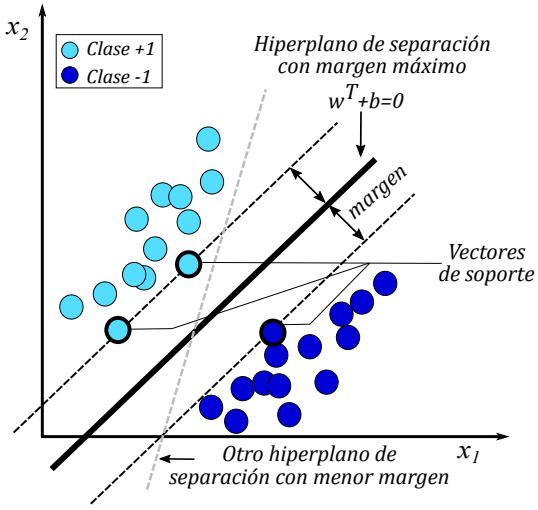

Figura 3.4: Principio de la SVM

$y_i = 1$ y $f(\mathbf{x}_i) < 0$ para los puntos que tienen $y_i = -1$, de modo que $y_i f(\mathbf{x}_i) > 0$ para todos los puntos de datos de entrenamiento.

La distancia mínima de un punto $\mathbf{x}$ a el hiperplano definido por $f(\mathbf{x}) = 0$, está dado por $\left|f(\mathbf{x})\right| / \|\mathbf{w}\|$. Por lo tanto, la distancia de un punto $\mathbf{x}_i$ a la superficie de decisión es dada por [165]:

$$\frac{y_i f(\mathbf{x}_i)}{\|\mathbf{w}\|} = \frac{y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b)}{\|\mathbf{w}\|} \tag{3.26}$$

El margen está dado por la distancia perpendicular al punto más cercano $\mathbf{x}_i$ del conjunto de datos, y como se desea optimizar los parámetros $\mathbf{w}$ y b con el fin de maximizar esta distancia, el margen máximo se encuentra resolviendo:

$$\arg\max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} \min_n \left[y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \right] \right\} \tag{3.27}$$

donde hemos tenido el factor $1/\|\mathbf{w}\|$ fuera de la optimización de más de i porque $\mathbf{w}$ no depende de i.

La razón que hace a las SVM más robustas que otros clasificadores es su criterio de entrenamiento, que consiste en un compromiso entre la minimización del riesgo empírico y del riesgo estructural. Éste último evita un posible sobreajuste de la máquina al conjunto de entrenamiento. Así que este problema puede ser escrito como un problema de optimización convexa [166]:

$$\begin{array}{ll} \min & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{sujeto a} & y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq \varepsilon \end{array} \tag{3.28}$$

donde ε es la distancia mínima permitida entre el hiperplano de separación y cualquier punto en el conjunto de datos. En la mayoría de los problemas reales las diferentes clases en los conjuntos de datos no son separables, por lo tanto, deben introducirse las variables de

holgura ξ_i (slack variables) para hacer frente a las limitaciones del problema de optimización y la solución viene dada por la siguiente formulación:

$$\begin{aligned} \min \quad & \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{sujeto a} \quad & \begin{cases} y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b) \geq \varepsilon - \xi_i \\ \varepsilon_i \geq 0 \end{cases} \end{aligned} \tag{3.29}$$

siendo C el factor de ponderación entre el riesgo empírico y el riesgo estructural, cuanto mayor sea el valor de C, mayor es el riesgo. Con el fin de resolver este problema, se introducen los multiplicadores de Lagrange y la expresión (3.29) se puede reescribir como:

$$\mathscr{L}(\mathbf{w},b,\mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}a_i\left(y_i f(\mathbf{x}_i) - \varepsilon + \xi_i\right) - \sum_{i=1}^{N}c_i\xi_i \tag{3.30}$$

donde $a_i \geq 0$ y $c_i \geq 0$ son los multiplicadores de Lagrange. Mediante la optimización de $\mathscr{L}$ con respecto a las variables primarias $(\mathbf{w}, b, \xi_i)$, es posible obtener [165]:

$$\begin{aligned} \frac{\partial\mathscr{L}}{\partial\mathbf{w}} = 0 &\Rightarrow \mathbf{w} = \sum_{i=1}^{N}a_i y_i \mathbf{x}_i \\ \frac{\partial\mathscr{L}}{\partial b} = 0 &\Rightarrow \sum_{i=1}^{N}a_i y_i = 0 \\ \frac{\partial\mathscr{L}}{\partial\xi_i} = 0 &\Rightarrow a_i = C - c_i \end{aligned} \tag{3.31}$$

Reemplazando estos resultados en la ecuación (3.30), se obtiene la representación dual del problema de máximo margen:

$$\mathscr{L}(\mathbf{a}) = \sum_{i=1}^{N}a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_i a_j y_i y_j \left\langle \mathbf{x}_i, \mathbf{x}_j \right\rangle \tag{3.32}$$

esta expresión es denominada como la expansión de vectores de soporte, es decir, $\mathbf{w}$ puede ser completamente descrito como una combinación lineal de las muestras de entrenamiento $\mathbf{x}_i$. El subconjunto de datos para los que $a_i \geq 0$ se llaman los vectores de soporte y, por lo tanto, son el conjunto de puntos que se encuentran en el margen. Teniendo en cuenta que $a_i, c_i \geq 0$ son necesarios porque estos son los multiplicadores de Lagrange, implica que $a_i \leq C$. Por lo tanto, la función (3.32) tienen que ser minimizada con respecto a a_i sujeta a

$$\begin{aligned} & 0 \leq a_i \leq C \\ & \sum_{i=1}^{N}a_i y_i = 0 \end{aligned} \tag{3.33}$$

Después de haber resuelto el problema de programación cuadrática, para determinar el parámetro b, es necesario tener en cuenta que los vectores de soporte para los que $0 \leq a_i \leq C$

tienen $\xi_i = 0$ de manera que $y_i f(\mathbf{x}_i) = \varepsilon$ y por lo tanto va a satisfacer

$$y_i \left(\sum_{j \in S} a_j y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + b \right) = \varepsilon \tag{3.34}$$

donde S denota el conjunto de índices de los vectores de soporte. Aunque se puede resolver esta ecuación para b, eligiendo arbitrariamente un vector de soporte $\mathbf{x}_i$, se hace uso de un valor predeterminado para $\varepsilon = 1$, lo que hace posible encontrar una solución numéricamente más estable, dada por [165]:

$$b = \frac{1}{N_{\mathscr{M}}} \sum_{n \in \mathscr{M}} \left(y_i - \sum_{j \in S} a_j y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right) \tag{3.35}$$

donde $\mathscr{M}$ es el conjunto de los índices de los datos con $0 \leq a_i \leq C$.

Este procedimiento se puede extender para el caso de funciones no lineales, sustituyendo el producto de punto $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ en la ecuación (3.32) para una función de *kernel* $K(\mathbf{x}_i, \mathbf{x}_j)$, que puede ser pensado como un producto escalar en un espacio de características F, el cual podría alcanzarse simplemente procesando las muestras $\mathbf{x}_i$ por un mapa no lineal $\Psi : X \rightarrow F$. La ventaja de este enfoque es que no es necesario conocer Ψ sino sólo la función del *kernel*. Para garantizar que $K(\mathbf{x}_i, \mathbf{x}_j)$ corresponde a un producto escalar de alguna característica del espacio F, debe satisfacer las condiciones de Mercer [166]. Con el fin de clasificar los nuevas muestras utilizando el modelo de entrenamiento, la función (3.25) se convierte para el caso no lineal:

$$f(\mathbf{x}) = \sum_{j=1}^{N} a_j y_j K(\mathbf{x}_i, \mathbf{x}_j) + b \tag{3.36}$$

El valor de salida que la SVM entrega no es una probabilidad a posteriori como en el caso de los clasificadores basados en GMM, sino que se puede interpretar como la probabilidad que el vector pertenece a una clase específica. En este trabajo, se tiene en cuenta que la voz está descrita por varias ventanas y para cada una de ellas se obtiene un valor de salida de la SVM, estos valores son normalizados con respecto a la duración de la señal de voz, y este valor se llamará *score*.

3.2. Estrategias para la fusión de información

Hasta ahora, se ha abordado la cuestión de cómo tomar una decisión sobre la presencia o ausencia de la EP, mediante un conjunto de características en un sistema de clasificación. En reconocimiento de patrones un sistema desarrollado para detectar automáticamente las voces patológicas, tiene como objetivo principal lograr la máxima precisión posible y esto se puede lograr mediante la combinación de información entregada por cada conjunto de características, la cual puede ser complementaria, puesto que considera varios fenómenos que participan en el proceso de producción de la voz y pueden ser de gran utilidad para la detección automática de la EP.

En los sistemas de reconocimiento de patrones, hay principalmente dos enfoques diferentes para la fusión de información: la fusión pre-clasificación y la fusión posterior a la clasificación. La fusión pre-clasificación combina la información antes de utilizar cualquier clasificador, tal combinación puede llevarse a cabo a nivel de características [167]. Mientras que, en la fusión posterior a la clasificación, la información se combina después de haber obtenido las primeras decisiones de los clasificadores, y puede hacerse de dos maneras, a nivel de *scores* o a nivel de decisión [168]. A continuación, se presentarán las estrategias para la fusión de información en este trabajo.

3.2.1. Fusión a nivel de características

La fusión a nivel de características se refiere a la combinación de diferentes vectores de características que se obtienen usando múltiples algoritmos en las mismas señales de voz (fusión pre-clasificación). Sin embargo, sólo cuando los vectores de características son homogéneos o compatibles, es posible concatenar los vectores para formar un único vector de características. En la práctica, la integración a nivel de característica es difícil de lograr, debido a que los espacios de características de los diferentes sistemas pueden no tener las mismas condiciones iniciales, ya que, al ser obtenidos mediante diferentes técnicas, se pueden utilizar diferentes longitudes de la señal, procedimientos de pre-procesamiento, o incluso diferentes tareas del habla de los pacientes. Además, es importante resaltar que la concatenación de vectores de características puede resultar en un vector de características con gran dimensionalidad, que es conocido como un problema general en la mayoría de las aplicaciones de reconocimiento de patrones [167].

En la figura 3.5, se presenta el esquema de fusión de información a nivel de características utilizada en este trabajo, la cual es posible, ya que la mayoría de las técnicas de caracterización basadas en el análisis TF fueron implementadas bajo las mismas condiciones, exceptuando la WPT.

Para la fusión, se realizó la concatenación de las características dinámicas obtenidas de los espectros de las 5 TFD, las cuales tienen diferentes números de variables, dependiendo del método que se utilizó. A continuación, se muestra una lista de las variables dinámicas estimadas y del número de características:

– Energía instantánea - 3.
– Frecuencia instantánea - 1.
– Ancho de banda - 1.
– Centroides espectrales - el número de características es variable, ya que depende del número de bandas en el cual se divide el espectro (2 a 20), éste se optimiza de acuerdo al rendimiento máximo dado por un clasificador.
– Coeficientes cepstrales - número de características variable, dependiente del número de coeficientes que se quieran tomar (2-32), éste se optimiza de acuerdo al rendimiento máximo dado por un clasificador.

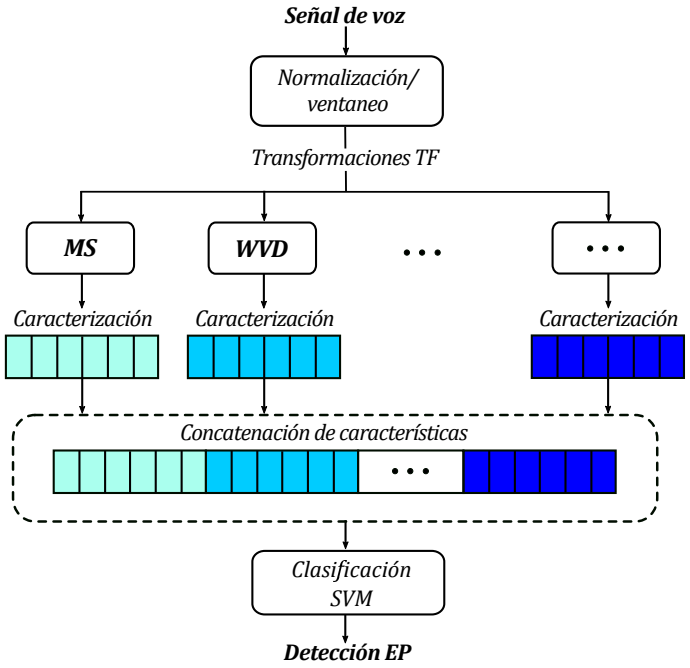

Figura 3.5: Diagrama de fusión a nivel de características

3.2.2. Fusión a nivel de *scores*

En el caso de la fusión posterior a la clasificación, la información se combina después de haber obtenido las primeras decisiones de los clasificadores [168]. En este trabajo se utiliza la fusión a nivel de *scores*, la cual es considerada como un esquema simple que se puede utilizar para combinar las primeras salidas de los clasificadores como características para alimentar una segunda etapa de clasificación. Una de las aplicaciones más robusta y sencilla de las estrategias de combinación de *scores* es utilizar un clasificador adicional para combinar las salidas de un conjunto anterior de clasificadores para tomar la decisión final, es decir, las salidas de los clasificadores individuales se utilizan para construir un nuevo espacio de características en la que otro clasificador está entrenado. Este enfoque es capaz de combinar las capacidades de generalización de varias técnicas de clasificación con el fin de encontrar una decisión final óptima; el clasificador más usado en este contexto es la SVM.

Una de las ventajas en la combinación de salidas de diferentes clasificadores en lugar de características que se fusionan, es que la estructura del espacio de características utilizado para alimentar a cada clasificador es mucho más simple. Aunque algunos de los clasificadores podrían presentar un mejor rendimiento que otros, el conjunto de registros de las señales de voz que son incorrectamente clasificados no necesariamente se superponen; por lo tanto, la combinación de sus salidas podría mejorar el rendimiento global del sistema. Además, los diferentes clasificadores entrenados utilizando los mismos datos pueden no sólo diferir en sus desempeños globales, sino que también pueden tener su propia región en el espacio de características donde cada uno realiza el mejor rendimiento [167].

La figura 3.6 muestra un esquema general del sistema de detección automática de voces patológicas, mediante dos etapas de clasificación, para el cual se utiliza las características dinámicas obtenidas de las 6 TFD.

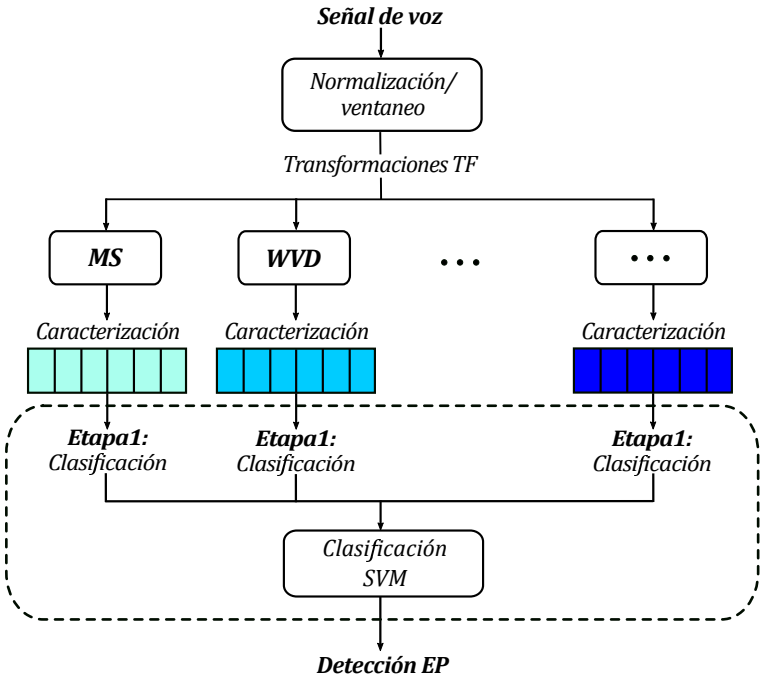

Figura 3.6: Diagrama de fusión a nivel de scores

3.2.3. Fusión mediante el aprendizaje con múltiples *kernels*

El aprendizaje con múltiples *kernels* (*MKL - Multiple Kernel Learning*) fue desarrollado en la última década para resolver problemas de clasificación binaria. Esta técnica permite la combinación convexa de *kernels*, de manera que preserva la estructura de los datos infiriendo un único modelo. Esta fusión de información, tiene un razonamiento similar a la combinación de diferentes clasificadores, es decir, en lugar de elegir una sola función de *kernel* para un conjunto de datos, se utiliza un algoritmo para hacer su combinación. Los diferentes *kernels* corresponden a diferentes nociones de similitud y en lugar de tratar de encontrar el que funciona mejor para el conjunto, se puede utilizar un método de aprendizaje que puede utilizar una combinación de *kernels* para una mejor solución, que puede evitar una fuente de sesgo. Los diferentes *kernels* pueden estar utilizando entradas provenientes de diferentes representaciones o diferentes subconjuntos de características, por lo que la combinación de *kernels* es una manera posible de combinar múltiples fuentes de información y es denominada combinación intermedia, como se presenta en la figura 3.7 [169].

La SVM es uno de los métodos basados en *kernels* que ha demostrado ser de gran alcance para una amplia gama de diferentes problemas de análisis de datos, un tema importante durante el entrenamiento es la selección de la función del *kernel* $K(\mathbf{x}_i, \mathbf{x}_j)$ y sus parámetros.

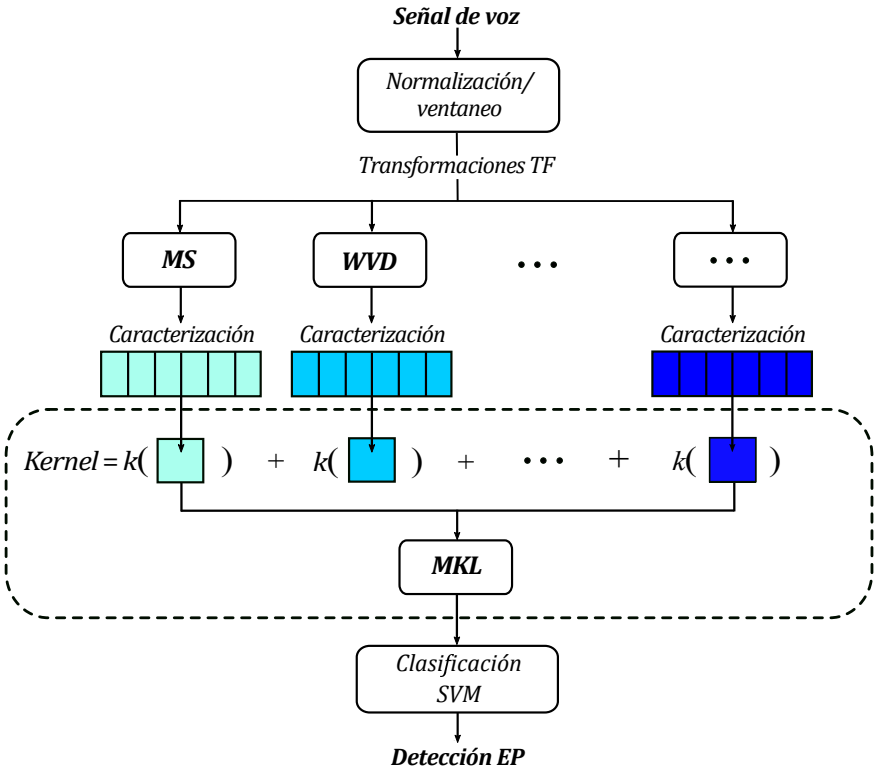

Figura 3.7: Diagrama de fusión mediante MKL

Por lo general, se utiliza un procedimiento de validación cruzada para elegir el mejor rendimiento de la función del *kernel*, entre un conjunto de funciones de *kernels* de un conjunto de validación independiente del conjunto de entrenamiento [169]. En los últimos años, se han propuesto diferentes métodos de MKL, donde se utilizan múltiples *kernel* [169]

$$K_\eta(\mathbf{x}_i,\mathbf{x}_j) = f_\eta\left(\left\{K_j(\mathbf{x}_i^m,\mathbf{x}_j^m)\right\}_{m=1}^{P}\right) \tag{3.37}$$

la función de combinación, $f_\eta : \mathbb{R}^P \rightarrow \mathbb{R}$, puede ser un una función lineal o no lineal. Las funciones del *kernel*, $\left\{K_m : \mathbb{R}^{D_m} \times \mathbb{R}^{D_m} \rightarrow \mathbb{R}\right\}_{m=1}^{P}$, pueden tomar P representaciones de características (no necesariamente distintas) $\mathbf{x}_i = \left\{\mathbf{x}_i^m\right\}_{m=1}^{P}$ donde $\mathbf{x}_i^m \in \mathbb{R}^{D_m}$ y D_m es la dimensionalidad de la representación correspondiente. η parametriza la función de combinación y la aplicación más común es,

$$K_\eta(\mathbf{x}_i,\mathbf{x}_j) = f_\eta\left(\left\{K_m(\mathbf{x}_i^m,\mathbf{x}_j^m)\right\}_{m=1}^{P}|\eta\right) \tag{3.38}$$

donde los parámetros son usados para combinar un conjunto de *kernels* predefinidos durante el entrenamiento. Los métodos de combinación lineal son los más populares y tienen dos categorías básicas: suma no ponderada (es decir, el uso de suma o promedio de los *kernels*

como *kernel* combinado) y suma ponderada. En el caso de suma ponderada, se parametriza linealmente la función de combinación:

$$K_\eta(\mathbf{x}_i,\mathbf{x}_j) = f_\eta\left(\left\{K_m(\mathbf{x}_i^m,\mathbf{x}_j^m)\right\}_{m=1}^P|\eta\right) = \sum_{m=1}^P \eta_m K_m(\mathbf{x}_i^m,\mathbf{x}_j^m) \tag{3.39}$$

donde η denota los pesos del *kernel*. Diferentes versiones de este enfoque se diferencian en la forma en que ponen restricciones en η: la suma lineal (es decir, $\eta \in \mathbb{R}^P$), la suma cónica (es decir, $\eta \in \mathbb{R}_+^P$), o la suma convexa (es decir, $\eta \in \mathbb{R}_+^P$ y $\sum_{m=1}^P \eta_m = 1$). Como puede verse, la suma cónica es un caso especial de la suma lineal y la suma convexa es un caso especial de la suma cónica. Las sumas cónicas y convexas tienen dos ventajas sobre la suma lineal en términos de interpretabilidad. En primer lugar, cuando se tienen pesos positivos del *kernel*, se puede extraer la importancia relativa de los *kernel* combinados con mirarlos. En segundo lugar, cuando se restringen los pesos del *kernel* para ser no negativos, esto corresponde al ajuste de los espacios de características y el uso de la concatenación de ellos como la representación de característica combinadas. Es decir, los coeficientes de combinación optimizada pueden ser utilizados para entender cuáles características son de importancia para la discriminación. Los parámetros de combinación también se pueden restringir usando limitaciones adicionales, tales como la norma-ℓp relativa a los pesos del *kernel*. Por ejemplo, la norma-$\ell 1$ promueve la dispersión en el nivel del *kernel*, que puede ser interpretado, de manera similar a cómo se hace en la técnica de selección de características LASSO [170].

En el problema de aprendizaje de múltiples *kernels* para clasificación binaria dados N muestras $(\mathbf{x}_i, y_i)$ $(y_i \in \{+1,-1\})$, $\mathbf{x}_i$ se traduce a través de P mapeos $\Phi_m(\mathbf{x}) \to \mathbb{R}^{D_m}$, $m = 1,\ldots,P$ desde la entrada en P espacios de características $\Phi_1(\mathbf{x}_i),\ldots,\Phi_P(\mathbf{x}_i)$ donde D_m es la dimensionalidad de la representación correspondiente. La formulación inicial de MKL propone [171]:

$$\begin{aligned} \text{minimizando} \quad & \frac{1}{2}\left(\sum_{m=1}^P d_m \|\mathbf{w}_m\|_2\right)^2 + C\sum_{i=1}^N \xi_i \\ \text{con respecto a} \quad & \mathbf{w}_m \in \mathbb{R}^{S_m},\ \xi \in \mathbb{R}_+^N,\ b \in \mathbb{R} \\ \text{sujeto a} \quad & y_i\left(\sum_{m=1}^P \left\langle \mathbf{w}_m, \Phi_m(\mathbf{x}_i^m)\right\rangle + b\right) \geq 1-\xi_i, \quad \forall i \end{aligned} \tag{3.40}$$

donde el espacio de características construido $\Phi_m(\cdot)$ tiene la dimensionalidad S_m y el peso d_m. La solución del problema puede ser escrito como $\mathbf{w}_m = \eta_m \mathbf{w}'_m$ con $\eta_m \geq 0$, $\forall \sum_{m=1}^P \eta_m = 1$ [171]. Por lo tanto, la norma-$\ell 1$ de η está limitada a uno, mientras que uno está penalizando la norma-$\ell 2$ de $\mathbf{w}_m$ en cada espacio de característica m de manera separada. La idea es que la norma-$\ell 1$ restrinja las variables que tienden a tener soluciones óptimas dispersas, mientras que las variables penalizadas por la norma-$\ell 2$ no lo hacen. Así, el problema de optimización anterior ofrece la posibilidad de encontrar soluciones dispersas en el espacio de características con soluciones no dispersas. Si se tiene en cuenta este problema de optimización como un problema de programación cónica de segundo orden, se obtiene la siguiente formulación

dual:

$$\begin{aligned} \text{minimizado} \quad & \frac{1}{2}\gamma^2 - \sum_{i=1}^{N} \alpha_i \\ \text{con respecto a} \quad & \gamma \in \mathbb{R}, \alpha \in \mathbb{R}_+^N \\ \text{sujeto a} \quad & \gamma^2 d_m^2 \geq \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K_m(\mathbf{x}_i^m, \mathbf{x}_j^m) \quad \forall m \\ & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \quad \forall i \end{aligned} \tag{3.41}$$

donde de nuevo se obtienen los pesos óptimos del *kernel* a partir de las variables duales, estos pesos satisfacen $\sum_{m=1}^{P} d_m^2 \eta_m = 1$. El problema dual es exactamente equivalente a la formulación de programación cuadrática con restricciones cuadráticas (QCQP) [172].

$$\begin{aligned} \text{minimizado} \quad & \gamma - \sum_{i=1}^{N} \alpha_i \\ \text{con respecto a} \quad & \gamma \in \mathbb{R}, \alpha \in \mathbb{R}^N \\ \text{sujeto a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \\ & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_m y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \leq \gamma, \quad \forall m \end{aligned} \tag{3.42}$$

donde $K_m(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi_m(\mathbf{x}_i), \Phi_m(\mathbf{x}_j) \rangle$. Se debe tener en cuenta, que se tiene una restricción cuadrática por *kernel* ($S_m(\alpha) \leq \gamma$). En el caso de $P = 1$, el problema anterior se reduce a la original SVM dual.0

En el presente trabajo, la implementación de MKL es realizada mediante el *toolbox* Shogun [173], teniendo en cuenta que los *kernels* tienen que ser elegidos a priori y que C es un parámetro de regularización pre-especificado.

Capítulo 4

Experimentos y resultados

En el presente capítulo se presentan los experimentos y resultados obtenidos a partir de la utilización de las técnicas TF, las diferentes estrategias de clasificación y de fusión de información, con el fin de evaluar su impacto en el sistema para la detección automática de voces con EP.

En primer lugar, se describe la base de datos utilizada en los experimentos y la metodología de validación, que es una parte importante del trabajo para permitir comparaciones con trabajos anteriores y futuros, además de las respectivas medidas de rendimiento.

Luego se presentan los resultados obtenidos para la detección automática de la EP. Los experimentos iniciales fueron realizados sobre vocales sostenidas, seguidos por los realizados sobre habla continua mediante el análisis de frases de la base de datos de Parkinson.

4.1. Base de datos

La base de datos utilizada en este trabajo se describe completamente en [174]. Incluye varios ejercicios del habla como vocales sostenidas, palabras aisladas, frases y un monólogo. Las muestras de voz fueron pronunciadas por 50 pacientes con EP y 50 personas sanas, 25 mujeres y 25 hombres en cada grupo. Todos los participantes son hablantes de español nativo de Colombia. Las edades de los hombres con EP oscilan entre los 33 y 77 años (promedio 62.2 ± 11.2), y las edades de las mujeres con EP oscilan entre los 44 y 75 años (promedio 60.1 ± 7.8). Para el caso de las personas sanas, la edad de los hombres está en el intervalo de 31 a 86 años (promedio 61.2 ± 11.3) y la edad de las mujeres se encuentra entre los 43 a los 76 años (promedio 60.7 ± 7.7). Se puede notar que la base de datos se encuentra balanceada en términos de edad y género. Las grabaciones fueron capturadas en condiciones de ruido controlado, en la cabina sonoamortiguada de la Clínica Noel, en Medellín, Colombia; los registros fueron muestreados a 44100 Hz con una resolución de 16 bits. Además, es importante resaltar que todos los pacientes fueron diagnosticados por expertos neurólogos y ninguna de las personas en el grupo de personas sanas tiene antecedentes de síntomas relacionados con la EP o cualquier otro tipo de síndrome de trastorno del movimiento. Los valores medios de la evaluación neurológica según la escala UPDRS-III y Hoehn & Yahr [175] son $36,7 \pm 18,7$ y $2,29 \pm 0,8$, respectivamente.

En este trabajo fueron utilizadas las vocales sostenidas y el conjunto de las seis frases, las cuales se proporcionan en la tabla 4.1.

Tabla 4.1: Frases de la base de datos, catalogadas como simples y complejas desde el punto de vista de la sintaxis.

	Frases	Textos
	1	Los libros nuevos no caben en la mesa de la oficina (simple)
	2	Laura sube al tren que pasa (compleja)
-	**3**	Luisa Rey compra el colchón duro que tanto le gusta (compleja)
	4	Mi casa tiene tres cuartos (simple)
	5	Omar, que vive cerca, trajo miel (compleja)
	6	Rosita Niño, que pinta bien, donó sus cuadros ayer (compleja)

4.2. Metodología de validación

Para la evaluación del sistema se empleará la metodología propuesta en [17], con el fin de comparar los resultados de los experimentos de forma que se puedan cuantificar sus diferencias y decidir objetivamente cuáles son mejores, para la detección de la EP. De acuerdo con esta metodología, las capacidades de generalización del sistema tienen que ser probados siguiendo un esquema de validación cruzada con diferentes conjuntos de entrenamiento y validación *(k-folds)*.

4.2.1. Validación cruzada

En este trabajo se aplica una metodología de validación cruzada, conocida como *k-fold*, en donde el conjunto de datos se divide de forma aleatoria en k subconjuntos independientes de aproximadamente igual tamaño. Se efectúa el entrenamiento y prueba del modelo k veces, dejando fuera del entrenamiento un subconjunto diferente cada vez. Con este subconjunto se valida el funcionamiento del modelo entrenado con los $k-1$ subconjuntos restantes. Los valores típicos de k suelen ser del orden de 5 a 20. Al término del proceso, se han aprovechado todos los datos disponibles para entrenar y validar el modelo. La estimación de la generalización del modelo es el promedio de las tasas de clasificación obtenidas con cada uno de los subconjuntos de prueba, lo que permite calcular intervalos de confianza para la presentación de resultados, dando mayor robustez al sistema. El estimador obtenido no está sesgado puesto que en cada resultado parcial no se usan los mismos datos para entrenar que para probar. Sin embargo, puede tener más varianza que otros métodos [17].

Cuando se entrenan varios modelos usando un conjunto de entrenamiento y se usa un segundo conjunto de datos (conjunto de validación) para decidir qué modelo es el mejor, se debe usar un tercer conjunto (conjunto de prueba o test) para obtener una estimación no sesgada del error de generalización del modelo elegido. También se puede, una vez elegido el modelo con los valores de los parámetros que funcionan mejor, entrenar de nuevo ese modelo con los conjuntos de entrenamiento y de validación juntos, y medir su capacidad de generalización con el conjunto de test.

En este trabajo k es igual a 10 particiones y los conjuntos de datos o *folds* usados son los mismos para cada uno de los experimentos realizados con las diferentes técnicas TF, lo

que permite una comparación directa. Los resultados van a estar en términos de eficiencia, sensibilidad, especificidad, curva ROC y el área bajo la curva (AUC).

4.2.2. Medidas de rendimiento

Para presentar los resultados de detección de la EP y, en general, para cualquier sistema de reconocimiento de patrones es común el uso de la matriz de contingencia o de confusión, que recoge el número de aciertos y fallos del sistema para cada una de las clases en que se divide el problema, dado un valor fijo del umbral de decisión [17]. La tabla 4.2 muestra la matriz de confusión en la que la suma de los elementos de cada fila debe ser igual a 100 %.

Tabla 4.2: Matriz de confusión

		Clase real	
		Patológica	Normal
Clase estimada	Patológica	TP	FP
	Normal	FN	TN

De acuerdo con esta matriz y tomando como referencia la clase patológica, se definen los siguientes conceptos:

- *Detección correcta o aceptación verdadera (TP, True Positive)*: el número o porcentaje de patrones de la clase 0 que el sistema clasifica correctamente como pertenecientes a la clase 0.
- *Falso rechazo (FN, False Negative)*: el número o porcentaje de patrones de la clase 0 que el sistema clasifica incorrectamente como pertenecientes a la clase 1.
- *Falsa aceptación (FP, False Positive)*: el número o porcentaje de patrones de la clase 1 que el sistema clasifica incorrectamente como pertenecientes a la clase 0.
- *Rechazo verdadero (TN, True Negative)*: el número o porcentajes de patrones de la clase 1 que un sistema clasifica correctamente como pertenecientes a la clase 1.

Nótese que cuando los valores se representan en porcentaje, $TP+FN=100$ y $FP+TN=100$. A partir de dichos valores, se pueden estimar:

- *Tasa de acierto o eficiencia (CCR, Correct Classification Rate)*: es la proporción de patrones correctamente clasificados por el sistema.

$$CCR = \frac{TP+TN}{TP+FN+FP+TN} \tag{4.1}$$

- *Tasa de error (ER, Error Rate)*: es el complemento a la tasa de acierto, es decir, la proporción de patrones erradamente clasificados.

$$ER = 1-CCR = \frac{FN+FP}{TP+FN+FP+TN} \tag{4.2}$$

Idealmente, la tasa de acierto debería ser del 100 % y la tasa de error del 0 %. Aunque ambas medidas pueden usarse indistintamente, en tareas donde la tasa de aciertos es muy alta se utiliza a menudo la tasa de error como indicador único. Si el número de patrones de las clases no están balanceados (no tienen el mismo número de patrones entre clases), el acierto y el error no reflejan realmente el funcionamiento del sistema, para corregir este problema se emplean estos otros parámetros:

- *Sensibilidad (S)*: da una indicación de la capacidad del sistema para detectar los patrones de la clase de referencia. Cuando los valores se representan en porcentaje, la sensibilidad coincide con TP.

$$S = \frac{TP}{TP + FN} \tag{4.3}$$

- *Especificidad (E)*: da una idea de la capacidad del sistema para rechazar los patrones que no pertenecen a la clase de referencia. Cuando los valores se representan en porcentaje, la especificidad coincide con TN.

$$E = \frac{TN}{TN + FP} \tag{4.4}$$

En el caso ideal, S y E deben ser 1 (o el 100 % si se miden en porcentaje).

Curva ROC

La curva ROC (Característica de Operación del Receptor) es otra de las herramientas más utilizadas en tareas de detección de patologías, la cual expresa el rendimiento en términos de la sensibilidad y 1-especificidad. La forma y posición de la ROC depende de la forma y del solapamiento de las distribuciones subyacentes de las voces patológicas y normales, por lo que el punto de trabajo del sistema vendrá determinado por el valor de umbral elegido. También se han propuesto varias medidas teóricas para reducir la curva ROC a un único indicador de la precisión del diagnóstico, la más utilizada es el área bajo la curva (AUC). El área bajo la ROC es equivalente al test del signo-rango de Wilcoxon y puede usarse como una estimación de la probabilidad de que la patología detectada permita una identificación correcta. Este índice varía entre 0,5 (no hay precisión aparente) y 1 (precisión perfecta) a medida que la curva ROC se mueve hacia los márgenes izquierdo y superior. Cuanto mayor sea el área bajo la curva, mejor será el rendimiento del sistema [17].

La figura 4.1 muestra un esquema del funcionamiento de la curva ROC y sus posibles puntos de operación.

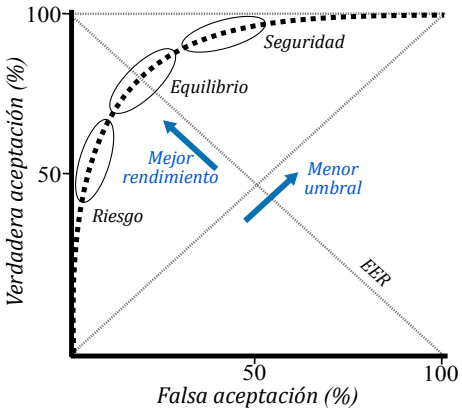

Figura 4.1: Resumen del comportamiento de la curva

4.3. Resultados

A continuación, se presentarán los experimentos y los mejores resultados obtenidos con la metodología y técnicas propuestas a lo largo de este trabajo, inicialmente sobre vocales sostenidas y después en habla continua. Es importante resaltar que los resultados de clasificación son presentados en cuanto a personas, es decir, que por cada una de las ventanas se tiene un valor de salida entregado por el clasificador y estos valores son normalizados con respecto al número de ventanas pertenecientes a la señal de voz de la persona. Los resultados van a estar en términos de eficiencia, sensibilidad, especificidad, curva ROC y AUC, pero debido a la gran cantidad de experimentos realizados se presentarán sólo los mejores resultados.

4.3.1. Resultados en vocales

Los primeros experimentos se realizaron sobre vocales sostenidas, con el fin de tener una base en el análisis de señales de voz, específicamente en la tarea de detección automática de la EP, para lo cual se construyó dos espacios de representación: el primero está constituido por medidas clásicas en las tareas de procesamiento de señales de voz como son NNE, HNR, GNE, *jitter*, *shimmer* y MFCC; las cuales fueron calculadas sobre tamaños de ventanas convencionales de 40 ms, con un solape del 50 % [1]. El segundo espacio, está constituido por la información proporcionada por la representación conjunta en frecuencia acústica y de modulación, los MS, con una ventana de análisis de 262 ms con un incremento de 64 ms, como se sugiere en [137].

Adicionalmente, con el objetivo de eliminar posibles redundancias en la información proporcionada por las características, se aplican dos técnicas clásicas de extracción de características, PCA y LDA. En cuanto a la tarea de clasificación, ésta es llevada a cabo por medio de los tres clasificadores presentados en el capítulo anterior. En el caso de GMM y GMM-UBM, se realizó un ajuste del número de Gaussianas M, entre 2 y 6; mientras que para la SVM, se realizó una ajuste de los parámetros mediante la variación de $C = [0.1,\ 1,\ 10,\ 100,\ 1000]$ y

γ = [0.0001, 0.001, 0.01, 0.1, 1, 10, 100], con el fin de encontrar los que mejor se ajusten a los datos. Durante los experimentos se consideró la caracterización dinámica, en la que cada registro de voz fue segmentado en ventanas, y cada una fue caracterizada y utilizada durante la etapa de clasificación.

En el primer experimento, se utilizan las características clásicas (NNE, HNR, GNE, *jitter*, *shimmer* y 12 MFCC) con cada una de las cinco vocales y las tres estrategias de clasificación, los resultados son obtenidos utilizando todo el espacio de características y luego de aplicar los dos métodos de extracción. La tabla 4.3 presenta un resumen de los resultados, mediante los porcentajes de eficiencias.

Tabla 4.3: % Eficiencia en medidas clásicas de procesamiento de voz en vocales.

Clasificador	ME	/a/	/e/	/i/	/o/	/u/
GMM	SE	**69.00 ± 10.75**	64.67 ± 13.35	63.67 ± 14.56	68.00 ± 12.49	62.33 ± 16.74
	LDA	**65.33 ± 10.77**	65.67 ± 16.27	64.00 ± 15.11	64.67 ± 12.67	57.33 ± 12.45
	PCA	**71.00 ± 10.01**	67.00 ±11.59	61.00 ± 16.67	69.33 ± 10.31	62.67 ± 12.18
GMM-UBM	SE	65.67 ± 10.75	63.33 ± 12.56	60.67 ± 10.62	**69.00 ± 11.36**	58.67 ± 13.60
	LDA	66.67 ± 14.45	66.67 ± 13.98	62.00 ± 15.65	**66.67 ± 7.60**	59.00 ± 15.06
	PCA	66.67 ± 10.22	64.33 ± 13.67	62.67 ± 15.97	**68.67 ± 13.92**	61.00 ± 9.78
SVM	**SE**	**72.33 ± 7.16**	67.33 ± 10.93	66.33 ± 11.30	68.00 ± 10.46	67.00 ± 11.59
	LDA	**67.00 ± 8.36**	61.33 ± 8.19	63.00 ± 14.72	65.00 ± 14.78	62.67 ± 10.46
	PCA	70.00 ± 8.43	67.67 ± 5.17	65.33 ± 13.60	**70.33 ± 11.20**	65.67 ± 10.01

ME: Método de extracción, SE: Sin extracción

Para el clasificador GMM, la vocal /a/ alcanzó la mejor tasa de acierto del 71 % utilizando PCA; en el caso de GMM-UBM fue de 69 % para la vocal /o/ con todo el espacio de representación; y la SVM alcanzó un 72.33 %, también con todo el espacio de representación. Se puede observar que la vocal /a/ fue la que mejor desempeño global presentó. Luego de revisar los resultados, se puede inferir que en los casos en que fueron utilizadas las técnicas de extracción de características, no se alcanzaron tasas lo suficientemente altas al reducir la dimensionalidad de los datos originales, respecto a los resultados obtenidos con todo el espacio de representación. También se comparó el desempeño de los tres clasificadores, y en promedio la SVM fue la que mejor desempeño y estabilidad presentó en la mayoría de los resultados, mas aún evidencia que cuando se utilizó todo el espacio de representación se alcanzaron las eficiencias más altas como se destaca en la tabla.

Teniendo en cuenta que la vocal /a/ fue la que tuvo mejores resultados, en la tabla 4.4 se muestran las medidas de rendimiento en términos de eficiencia, sensibilidad y especificidad, además de los parámetros de los clasificadores.
La mejor tasa de acierto para la vocal /a/ fue de 72.33 %, mediante la SVM sin usar técnicas de extracción de características. Además, se refleja una disminución en los porcentajes de varianza de alrededor 2 puntos porcentuales, con respecto al uso de los clasificadores basados en GMM. En cuanto a las medidas de rendimiento, en la mayoría de los casos, la sensibilidad mantiene un porcentaje mayor que la especificidad, lo cual indica que hay un número mayor de pacientes detectados correctamente con EP que aquellos que son detectados como sanos.

El segundo experimento en vocales sostenidas, se realizó mediante la implementación de los MS, cuyo conjunto de características incluye los centroides y el contenido energético

Tabla 4.4: Resultados de las medidas clásicas para la vocal /a/.

Clasificador	ME	% Eficiencia	% Sensibilidad	% Especificidad	% Reducción	Parámetros
GMM	SE	69.00 ± 10.75	78.67 ± 12.49	73.77 ± 13.49	–	M=6
	LDA	65.33 ± 10.77	76.00 ± 18.91	54.67 ± 11.24	37.5	M=6
	PCA	71.00 ± 10.01	82.00 ± 12.59	60.00 ± 11.76	43.75	M=5
GMM-UBM	SE	65.67 ± 10.75	81.33 ± 17.76	76.70 ± 20.17	–	M=5
	LDA	66.67 ± 14.45	70.67 ± 21.59	62.67 ± 15.14	56.25	M=2
	PCA	66.67 ± 10.22	65.33 ± 14.67	68.00 ± 17.44	81.25	M=2
SVM	**SE**	**72.33 ± 7.16**	81.33 ± 13.98	63.33 ± 9.56	–	C=1000, γ=0.01
	LDA	67.00 ± 8.36	85.33 ± 11.67	48.67 ± 12.59	37.5	C=0.1, γ=1
	PCA	70.00 ± 8.43	87.33 ± 12.35	52.67 ± 10.16	43.75	C=0.1, γ=1

ME: Método de extracción, SE: Sin extracción, –: No aplica

de diferentes bandas de frecuencia en los espectros obtenidos, por lo que se hizo necesario, primero realizar un ajuste de los parámetros de los centroides: el número de subbandas, gamma y el porcentaje de energía. Para el ajuste fino de estos parámetros se realiza un barrido desde 2 hasta 20 subbandas con un paso de 2; mientras el valor de gamma y el porcentaje de energía fue variado en el rango de 0 a 1. Con la finalidad de encontrar los mejores valores, se utilizó el clasificador GMM y las dos técnicas clásicas de extracción de características. La tabla 4.5 presenta los mejores parámetros de los centroides obtenidos mediante el barrido, el objetivo de este paso es conseguir los parámetros que contribuyen significativamente al proceso de clasificación entre las personas con EP y las sanas.

Tabla 4.5: Mejores parámetros para los centroides de los MS en vocales.

		Subbandas	(%) Gamma	(%) Energía
/a/	LDA	8	0.8	0.2
	PCA	12	0.1	0.9
/e/	LDA	12	0.8	0.3
	PCA	12	0.8	0.5
/i/	LDA	20	0.7	0.3
	PCA	16	0.8	0.2
/o/	LDA	20	0.6	0.3
	PCA	20	0.6	0.3
/u/	LDA	18	0.4	0.2
	PCA	12	0.5	0.8

Luego de haber ajustado los parámetros, se llevaron a cabo los experimentos con los MS, utilizando los tres tipos de clasificadores. En la tabla 4.6, se presenta un resumen con los porcentajes de eficiencia de las cinco vocales, utilizando todo el espacio de representación y aplicando las técnicas de extracción de características.

El clasificador GMM presenta la mejor tasa de acierto con la vocal /e/ del 69.67 %, luego de aplicar PCA; en el caso de GMM-UBM fue de 70 % para la vocal /i/, también utilizando PCA; y la SVM alcanzó un 67.67 % con todo el espacio de representación. Al analizar los resultados conseguidos por los clasificadores GMM y GMM-UBM, se evidencia que después de utilizar las técnicas de extracción de características, el desempeño es mayor en hasta 10 puntos porcentuales que cuando se utiliza todo el espacio de representación. Además, después de comparar cada uno de los resultados, se puede determinar que la mayoría de

Tabla 4.6: % Eficiencia de los MS en vocales.

Clasificador	ME	/a/	/e/	/i/	/o/	/u/
GMM	SE	56.33 ± 6.90	**64.67 ± 10.56**	61.33 ± 6.86	52.67 ± 7.12	61.67 ± 10.57
	LDA	65.00 ± 10.57	**69.67 ± 14.41**	68.00 ± 10.56	62.00 ± 16.28	67.33 ± 11.33
	PCA	61.67 ± 8.85	**69.33 ± 9.40**	67.33 ± 13.23	60.67 ± 6.96	63.67 ± 12.78
GMM-UBM	SE	54.67 ± 8.97	**67.33 ± 10.83**	61.33 ± 10.97	58.67 ± 10.77	61.00 ± 9.89
	LDA	64.00 ± 12.89	66.67 ± 14.30	**70.00 ± 12.29**	64.33 ± 16.06	63.33 ± 12.02
	PCA	64.67 ± 9.33	67.67 ± 12.02	**68.00 ± 10.67**	58.67 ± 12.22	59.67 ± 7.22
SVM	**SE**	67.33 ± 9.29	**67.67 ± 12.02**	66.00 ± 10.52	61.67 ± 16.88	59.00 ± 11.16
	LDA	**66.33 ± 9.71**	66.33 ± 11.40	66.00 ± 13.56	65.00 ± 17.72	62.67 ± 10.83
	PCA	67.00 ± 10.90	**67.00 ± 6.90**	67.33 ± 8.67	57.67 ± 12.91	63.33 ± 9.45

ME: Método de extracción, SE: Sin extracción

los experimentos presentan una mejor tasa de acierto para la vocal /e/ y que la estrategia de clasificación con mejor rendimiento es la SVM. El desempeño de la SVM presenta tasas de acierto muy similares cuando se utiliza todo el espacio de características, que cuando se usan las técnicas de extracción; por lo que, se puede observar que la SVM es bastante estable frente al conjunto de características de los MS, y que además, se puede prescindir del uso de técnicas de extracción, las cuales acarrean un costo computacional bastante elevado no por la técnica misma, sino debido a que se implementó un método tipo *wrapper*, el cual generó una dificultad a la hora de seleccionar el número óptimo de componentes durante el proceso de validación, ya que se varió sobre diferentes porcentajes de varianza, entre el 65 % y el 95 %, lo cual significa que las simulaciones realizadas se incrementan en el número de porcentajes de varianza utilizados por cada una de las técnicas de clasificación usadas, las cuales requieren a su vez el ajuste de los parámetros propios de cada técnica. Los costos computacionales de los algoritmos de extracción están asociados con la inversión de la matriz de covarianza que usualmente tiene el número de dimensiones igual al número de variables, y que sólo en casos extremos puede representar un tiempo elevado. Es importante aclarar que, en este caso, cuando nos referimos al aumento en el costo computacional, estamos limitándonos al costo durante el proceso de validación debido al aumento en el número de simulaciones que es necesario realizar. Para dar una idea, en una máquina con determinadas especificaciones, el tiempo por cada una de las evaluaciones fue entre 20 y 40 horas teniendo en cuenta que el proceso se realizó para las 5 vocales y diferentes tipos de características.

La tabla 4.6 muestra las medidas de rendimiento y los parámetros de los clasificadores sobre la vocal /e/, que fue la que mejores resultados tuvo. La mejor eficiencia fue de 69.67 %, obtenida con el clasificador GMM y la técnica de extracción LDA; aunque se debe resaltar que, la SVM presenta un desempeño muy similar, destacando que los porcentajes de sensibilidad son mayores que los de especificidad, lo que no se evidencia en los resultados obtenidos con los clasificadores basados en GMM.

En la tabla 4.8 se presenta un resumen con los resultados obtenidos mediante los dos conjuntos de características en las cinco vocales, con el clasificador SVM que fue el de mejor rendimiento, lo cual permite realizar una comparación directa entre las características clásicas y la representación TF (MS).

Las tasas de acierto más altas fueron de 72.33 % y 67.67 %, para las características clásicas y los MS, respectivamente. En ambos conjuntos, la sensibilidad presenta un porcentaje mayor

Tabla 4.7: Resultados de los MS para la vocal /e/.

Clasificador	ME	% Eficiencia	% Sensibilidad	% Especificidad	% Reducción	Parámetros
GMM	SE	64,67 ± 10,56	57,33 ± 13,77	62,28 ± 10,88	-	M=5
	LDA	**69,67 ± 14,41**	62,67 ± 25,18	76,67 ± 11,44	62,5	M=3
	PCA	69,33 ± 9,40	64,00 ± 20,66	74,67 ± 12,88	62,5	M=4
GMM-UBM	SE	67,33 ± 10,83	61,33 ± 21,95	67,17 ± 10,67	-	M=3
	LDA	66,67 ± 14,30	61,33 ± 26,44	72,00 ± 12,09	83,3	M=5
	PCA	67,67 ± 12,02	58,00 ± 26,30	77,33 ± 14,47	75	M=2
SVM	SE	67,67 ± 12,02	77,33 ± 12,25	58,00 ± 22,67	-	C=100, γ=10
	LDA	66,33 ± 11,40	74,00 ± 17,62	58,67 ± 19,32	75	C=100, γ=10
	PCA	67,00 ± 6,90	77,33 ± 14,81	56,57 ± 14,28	83,33	C=10, γ=0,1

ME: Método de extracción, SE: Sin extracción

Tabla 4.8: Resultados de las características clásicas y MS, con la SVM.

	Características	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
/a/	MS	67.33 ± 9.29	68.00 ± 23.48	66.67 ± 16.33	C=1000, γ=0.01
	Clásicas	**72.33 ± 7.16**	81.33 ± 13.98	63.33 ± 9.56	C=1000, γ=0.01
/e/	**MS**	**67.67 ± 12.02**	77.33 ± 12.25	58.00 ± 22.67	C=0.1, γ=0.01
	Clásicas	67.33 ± 10.93	74.67 ± 11.24	60.00 ± 14.05	C=10, γ=0.01
/i/	MS	66.00 ± 10.52	76.00 ± 18.65	56.00 ± 18.91	C=10, γ=0.001
	Clásicas	**66.33 ± 11.30**	70.67 ± 13.41	62.00 ± 15.09	C=1000, γ=0.1
/o/	MS	61.67 ± 16.88	70.00 ± 20.43	53.33 ± 24.34	C=10, γ=0.001
	Clásicas	**68.00 ± 10.46**	74.00 ± 11.09	62.00 ± 17.79	C=10, γ=0.01
/u/	MS	59.00 ± 11.16	80.67 ± 14.56	37.33 ± 16.98	C=1000, γ=0.001
	Clásicas	**67.00 ± 11.59**	79.33 ± 14.21	54.67 ± 19.83	C=1, γ=0.1

que la especificidad. Los resultados obtenidos muestran, en general, que las características clásicas presentan mejores resultados en las vocales en comparación con los MS, aunque los intervalos de confianza permiten observar que las diferencias entre ambos no son estadísticamente significativas. Adicionalmente, un punto importante a tener en cuenta es que el objetivo fundamental del trabajo es proporcionar medidas y estrategias de caracterización que proporcionen información adicional y complementaria a las medidas convencionales de análisis.

Por otra parte, se presenta en la figura 4.2 las curvas ROC de los resultados obtenidos con las características clásicas y los MS. Además, se incluyen las correspondientes AUC dentro del recuadro.

Las curvas ROC presentan el mismo comportamiento que se evidenciaba en las tablas presentadas anteriormente, en donde las características clásicas presentan el mejor comportamiento con la vocal /a/, la cual tiene una AUC de 0.79 y como se observa en la gráfica tiende a tener una mejor detección que las demás vocales. En el caso de los MS, se observa un deterioro en el rendimiento y en este caso, a diferencia de las medidas anteriores, la vocal de mejor rendimiento fue la vocal /i/ con una AUC de 0.72.

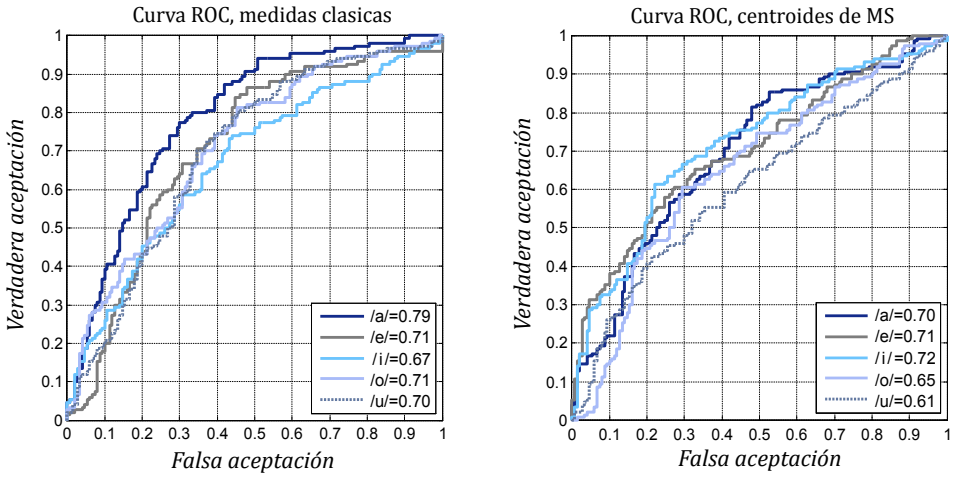

Figura 4.2: Curvas ROC y AUC, de las medidas clásicas y los MS para las vocales.

4.3.2. Resultados en frases

Después de analizar los resultados en vocales sostenidas, se prosiguió con habla continua mediante el uso de frases de la base de datos de Parkinson. La metodología propuesta (ver figura 3.1) no utiliza técnicas de extracción de características, ya que como se explicó anteriormente implican un elevado costo computacional, y al examinar el costo-beneficio no presenta muchas ventajas para este trabajo.

En el primer experimento, se utilizan las características clásicas (NNE, HNR, GNE, *jitter*, *shimmer* y 12 MFCC) con cada una de las 6 frases, con el fin de realizar un contraste con las técnicas TF propuestas en este trabajo.

Tabla 4.9: Resultados de las características clásicas y los MFCC en habla continua.

	Características	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	MFCC	69,00 ± 13,70	66,00 ± 22,19	72,00 ± 13,98	C=10, γ=0,1
	Clásicas	69,00 ± 8,76	68,00 ± 21,50	70,00 ± 17,00	C=100, γ=0,0001
Frase 2	MFCC	64,00 ± 12,65	58,00 ± 18,43	70,00 ± 14,14	C=1, γ=0,1
	Clásicas	69,00 ± 14,49	62,00 ± 30,48	76,00 ± 18,38	C=100, γ=0,001
Frase 3	MFCC	69,00 ± 12,87	70,00 ± 23,57	68,00 ± 13,98	C=1, γ=0,1
	Clásicas	63,00 ± 18,29	58,00 ± 27,41	68,00 ± 21,50	C=10, γ=0,1
Frase 4	**MFCC**	**68,00 ± 12,29**	60,00 ± 24,94	76,00 ± 12,65	C=1000, γ=0,1
	Clásicas	**65,00 ± 10,80**	62,00 ± 14,76	68,00 ± 13,98	C=1, γ=0,001
Frase 5	MFCC	63,00 ± 13,37	46,00 ± 23,19	80,00 ± 13,33	C=10, γ=0,1
	Clásicas	58,00 ± 13,17	42,00 ± 19,89	74,00 ± 13,50	C=100, γ=0,0001
Frase 6	MFCC	61,00 ± 15,24	48,00 ± 27,00	74,00 ± 9,66	C=100, γ=0,1
	Clásicas	61,00 ± 11,01	56,00 ± 18,38	66,00 ± 21,19	C=1, γ=0,0001

Sin embargo, es importante dejar claro que no se realizó ningún tipo de segmentación sonora/sorda (*voiced/unvoiced*), ya que el objetivo de este trabajo es precisamente no utilizar este tipo de métodos, para poder analizar longitudes de señal suficientemente largas, que permitan detectar no sólo cambios en alta frecuencia sino también cambios en baja frecuencia que

puedan ser asociados al tremor. Características como los MFCC sólo tienen sentido aplicarlas en períodos de vibración de las cuerdas vocales, en los cuales hay contenido armónico de la señal, por lo que su uso en el escenario de voz continua se debe limitar a segmentos sonoros.

Luego, se llevaron a cabo los primeros experimentos en habla continua utilizando las técnicas TF, inicialmente sobre la frase 1 (ver tabla 4.1) por medio de tres representaciones TF: MS, WVD y WPT. El conjunto de características extraídas de los espectros de las representaciones incluye, los centroides y el contenido de energía de diferentes bandas de frecuencia, junto con las medidas de entropía y operadores de energía no lineales. Durante los experimentos se consideró la caracterización dinámica y estática, en la primera cada registro de voz fue segmentado en ventanas de análisis de 262 ms con un incremento de 64 ms, formando vectores de características por cada ventana y cada uno de ellos fue utilizado durante la etapa de clasificación. Mientras que, para la caracterización estática se estiman 4 estadísticos a partir de cada vector de características: valor medio, la desviación estándar, curtosis y asimetría; formando un conjunto de características por cada registro de voz usado en la etapa de clasificación.

La capacidad de discriminación de las características estimadas se evalúa usando tres estrategias diferentes de clasificación: GMM, GMM-UBM y SVM. Para el ajuste de los parámetros, en el caso de los clasificadores basados en GMM se utilizó un número de Gaussianas M, entre 2 y 6; mientras que para la SVM, se realizó una variación de los parámetros $C = [0.1, 1, 10, 100, 1000]$ y $\gamma = [0.0001, 0.001, 0.01, 0.1, 1, 10, 100]$.

Uno de los objetivos de este experimento fue determinar cuál de los dos tipos de caracterizaciones (dinámica o estática) presenta mejor desempeño durante la etapa de clasificación. El rendimiento de los sistemas se resume en la tabla 4.10, mostrando los porcentajes de eficiencia de acuerdo a las tres estrategias de clasificación: GMM, GMM-UBM y SVM.

Tabla 4.10: % Eficiencia de los diferentes tipos de caracterización en la frase 1.

	Caracterización	GMM	GMM-UBM	SVM
MS	Estática	59.00 ± 9.43	60.00 ± 13.42	66.00 ± 15.62
	Dinámica	64.00 ± 16.25	60.00 ± 10.95	**68.00 ± 10.77**
WVD	Estática	60.00 ± 11.83	59.00 ± 15.13	69.00 ± 10.44
	Dinámica	70.00 ± 16.12	71.00 ± 16.61	**71.00 ± 10.44**
WPT	Estática	62.00 ± 15.36	69.00 ± 12.21	66.00 ± 13.56
	Dinámica	63.00 ± 14.18	63.00 ± 14.18	**73.00 ± 9.00**

De acuerdo con los resultados obtenidos, puede inferirse que la caracterización dinámica presenta el mejor desempeño, es decir, cuando se utilizaron las características de todas las ventanas pertenecientes a cada registro de voz en la etapa de clasificación. En general, con los tres clasificadores se presentó la tendencia a que la mejor eficiencia fuera obtenida por la caracterización dinámica, siendo la SVM la que alcanzó las tasas de clasificación más altas para cada una de las técnicas TF, como se muestra en la tabla. Luego de haber observado en este experimento, que el mejor rendimiento fue obtenido por medio de la caracterización dinámica, se determinó utilizarla con el resto de frases.

En la tabla 4.11 se muestran las medidas de rendimiento obtenidas con las tres técnicas

TF (MS, WVD y WPT) y los mejores parámetros de los tres clasificadores, donde cada conjunto de características se considera por separado durante el proceso de clasificación.

Tabla 4.11: Resultados de las técnicas TF para la frase 1.

Características	Clasificador	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
MS	GMM	64.00 ± 16.25	58.00 ± 22.01	60.10 ± 24.43	M=2
	GMM-UBM	60.00 ± 10.95	64.00 ± 18.38	63.50 ± 18.43	M=2
	SVM	**68.00 ± 10.77**	86.00 ± 16.47	50.00 ± 17.00	C=10, γ=0.1
WVD	GMM	70.00 ± 16.12	62.00 ± 28.98	71.30 ± 20.23	M=4
	GMM-UBM	71.00 ± 16.61	62.00 ± 31.90	72.09 ± 18.48	M=6
	SVM	**71.00 ± 10.44**	82.00 ± 14.76	60.00 ± 21.08	C=0.1, γ=0.1
WPT	GMM	63.00 ± 14.18	70.00 ± 19.44	60.52 ± 25.89	M=3
	GMM-UBM	63.00 ± 14.18	68.00 ± 21.50	64.04 ± 16.56	M=3
	SVM	**73.00 ± 9.00**	72.00 ± 23.48	74.00 ± 21.19	C=100, γ=0.0001

Los mejores rendimientos se presentan utilizando el clasificador SVM, la máxima eficiencia obtenida con los MS fue de 68 %, mientras que, para la WVD la tasa de acierto fue de 71 % y para la WPT fue de 73 %. Además, se puede observar que la WPT es la técnica TF con mejor rendimiento. En general la eficiencia obtenida a partir de los tres métodos de clasificación es muy similar, aunque en el caso de la WPT se evidencia unas bajas tasas de acierto obtenidas a partir de los clasificadores GMM y GMM-UBM, mientras que la tasa obtenida con la SVM se encuentra 10 puntos porcentuales respecto a éstos; este comportamiento no se evidencia en los MS y la WVD, ya que los resultados conseguidos por las tres técnicas de clasificación para la misma TFD son muy similares, más aún WVD alcanza tasas por encima del 70 % en los tres casos.

Aunque la WPT haya obtenido la eficiencia más alta, es importante resaltar el desempeño de la WVD. Adicionalmente, la eficiencia obtenida por las tres técnicas sobre la frase 1, es mejor que la eficiencia obtenida por MS sobre las cinco vocales, indicando que la información contenida en la señal de voz continua, tiene mayor capacidad discriminante para el problema de detección de Parkinson, que los registros de vocales sostenidas.

Por medio de estos experimentos se determinó que el clasificador más estable y con mejor desempeño fue la SVM, motivo por el cual, todos los experimentos que serán presentados en adelante fueron realizados mediante la SVM.

A continuación, se presentan los experimentos realizados con las 6 frases que contiene la base de datos de Parkinson, mediante la implementación de las 6 representaciones TF propuestas en este trabajo: MS, WVD, PWVD, SPWVD, CWD y WPT. Es importante resaltar que para la implementación de las TFD exceptuando la WPT, fue necesario remuestrear las señales de voz de las frases, ya que éstas tienen un tiempo de duración entre 2 y 5 segundos, dependiendo de la frase, por lo que la longitud de las muestras es muy grande para el cálculo de las representaciones TF. El costo computacional que requieren es elevado, especialmente en el caso de las TFD de la clase de Cohen, ya que el almacenamiento de la función de ambigüedad requiere una elevada capacidad de memoria para su almacenamiento. Por este motivo, las señales de voz se remuestrearon a 11025 Hz, lo cual se puede realizar, ya que el objetivo de este trabajo es analizar componentes de energía a baja frecuencia que no

deberían verse afectadas si se disminuye la frecuencia de muestreo a este valor.

En primer lugar, se explora la información proporcionada por 5 TFD: MS, WVD, PWVD, SPWVD y CWD, mediante la caracterización dinámica de los espectros obtenidos de cada una de las representaciones. Para empezar, el conjunto de características incluye los centroides y el contenido energético de diferentes bandas de frecuencia. Al igual que en las vocales sostenidas, fue necesario realizar un ajuste de los parámetros de los centroides, en este caso utilizando como clasificador la SVM, y se seleccionaron los mejores de acuerdo a los criterios de error de clasificación.

Es importante resaltar que, no sólo se realizaron experimentos con los mejores parámetros de los centroides sobre el espectro completo y la energía a su alrededor, sino que también se realizó un análisis de cada una de las subbandas de los centroides, eligiendo la mejor de acuerdo a su rendimiento en términos del porcentaje de clasificación. En la tabla 4.12, se presenta un resumen de la eficiencia de las cinco TFD para cada una de las 6 frases, con la implementación de los centroides sobre todo el espectro y los centroides calculados sólo con la mejor subbanda. Se encuentra en negrilla las mejores tasas de acierto para cada frase, y se resaltan las TFD y frases con mejor rendimiento.

Tabla 4.12: % Eficiencia obtenida con los centroides.

		MS	WVD	PWVD	SPWVD	CWD
Frase 1	Centroides	70.00 ± 10.00	72.00 ± 9.80	**72.00 ± 8.72**	72.00 ± 10.77	70.00 ± 13.42
	Csub	70.00 ± 11.83	72.00 ± 11.66	69.00 ± 7.00	71.00 ± 9.43	68.00 ± 10.77
Frase 2	Centroides	71.00 ± 13.75	72.00 ± 16.00	**72.00 ± 9.80**	71.00 ± 10.44	71.00 ± 10.44
	Csub	68.00 ± 10.77	70.00 ± 13.41	71.00 ± 7.00	69.00 ± 10.44	69.00 ± 15.13
Frase 3	Centroides	71.00 ± 13.75	69.00 ± 8.31	**73.00 ± 10.05**	69.00 ± 11.36	70.00 ± 7.75
	Csub	70.00 ± 10.95	68.00 ± 12.49	66.00 ± 12.81	67.00 ± 11.00	67.00 ± 15.52
Frase 4	Centroides	**78.00 ± 10.77**	70.00 ± 11.83	71.00 ± 11.36	73.00 ± 11.87	71.00 ± 10.44
	Csub	71.00 ± 13.74	71.00 ± 15.13	73.00 ± 9.00	70.00 ± 17.89	72.00 ± 15.36
Frase 5	Centroides	70.00 ± 13.42	69.00 ± 10.44	69.00 ± 13.00	71.00 ± 10.44	69.00 ± 13.00
	Csub	**72.00 ± 15.36**	71.00 ± 13.00	71.00 ± 13.00	71.00 ± 12.21	70.00 ± 10.95
Frase 6	Centroides	**72.00 ± 16.00**	71.00 ± 13.00	70.00 ± 12.65	67.00 ± 13.45	68.00 ± 13.27
	Csub	72.00 ± 16.61	69.00 ± 10.40	70.00 ± 10.95	71.00 ± 17.00	64.00 ± 13.56

Csub: centroides calculados sólo con la mejor subbanda

De acuerdo a los resultados, los MS presentan la mejor tasa de acierto con un 78 % para la frase 4. En el caso de la WVD, se obtuvo un 72 % para la frase 1, para la PWVD 73 % con la frase 3, para la PWVD y SPWVD 73 % con las frases 3 y 4, respectivamente; y por último, la CWD presenta un 71 % con la frase 2. Por otro lado, al comparar cada uno de los resultados se puede determinar que, la PWVD con las tres primeras frases y los MS con las tres últimas frases, como se aprecia en negrilla; mientras que el mejor rendimiento global fue con la frase 4. Sin embargo, en promedio la técnica TF que obtuvo un mejor desempeño mediante los centroides fue MS.

Luego de haber identificado las transformaciones que aportan mayor información en cuanto a las características obtenidas por los centroides, en las tablas 4.13 y 4.14, se presentan los resultados en términos de eficiencia, sensibilidad, especificidad y los mejores parámetros del clasificador de los MS y PWVD. Resaltando el mejor desempeño obtenido con los centroides calculados sobre todo el espectro y los centroides calculados sólo con la mejor subbanda.

Tabla 4.13: Resultados obtenidos con los centroides de los MS.

		% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	Centroides	70.00 ± 10.00	64.00 ± 18.38	76.00 ± 20.66	C=10, γ=1
	Csub	70.00 ± 11.83	74.00 ± 16.47	66.00 ± 18.97	C=100
Frase 2	Centroides	71.00 ± 13.75	64.00 ± 18.38	78.00 ± 19.89	C=1, γ=0.1
	Csub	68.00 ± 10.77	56.00 ± 18.38	80.00 ± 18.86	C=1, γ= 100
Frase 3	Centroides	71.00 ± 13.75	72.00 ± 25.30	70.00 ± 21.60	C=1000, γ=0.01
	Csub	70.00 ± 10.95	74.00 ± 26.74	66.00 ± 16.46	C=10, γ=1
Frase 4	**Centroides**	**78.00 ± 10.77**	80.00 ± 13.33	76.00 ± 15.78	C=10, γ=0.01
	Csub	71.00 ± 13.74	66.00 ± 25.03	76.00 ± 20.66	C=10, γ=10
Frase 5	Centroides	70.00 ± 13.42	52.00 ± 28.60	88.00 ± 16.87	C=1000, γ=0.1
	Csub	**72.00 ± 15.36**	64.00 ± 22.70	80.00 ± 18.85	C=1, γ=0.1
Frase 6	Centroides	72.00 ± 16.00	70.00 ± 17.00	74.00 ± 21.19	C=10, γ=0.01
	Csub	72.00 ± 16.61	72.00 ± 16.86	72.00 ± 26.99	C=1000, γ=100

Csub: centroides calculados sólo con la mejor subbanda

En el caso de los centroides calculados sobre todo el espectro, se alcanzó una eficiencia de 78 % y 72 % con los centroides calculados sólo con la mejor subbanda.

Tabla 4.14: Resultados obtenidos con los centroides de PWVD.

		% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	Centroides	72.00 ± 8.72	82.00 ± 14.76	62.00 ± 22.01	C=100, γ= 0.01
	Csub	69.00 ± 7.00	64.00 ± 18.38	74.00 ± 16.47	C=100, γ= 0.001
Frase 2	Centroides	72.00 ± 9.80	78.00 ± 17.51	66.00 ± 16.47	C=100, γ=0.001
	Csub	71.00 ± 7.00	62.00 ± 17.51	80.00 ± 13.33	C=0.1, γ=10
Frase 3	**Centroides**	**73.00 ± 10.05**	82.00 ± 14.76	64.00 ± 15.78	C=10, γ=0.1
	Csub	66.00 ± 12.81	70.00 ± 19.44	62.00 ± 22.01	C=1000, γ=100
Frase 4	Centroides	71.00 ± 11.36	74.00 ± 18.97	68.00 ± 16.87	C=1000, γ= 0.1
	Csub	**73.00 ± 9.00**	60.00 ± 21.08	86.00 ± 21.19	C=1000, γ=100
Frase 5	Centroides	69.00 ± 13.00	60.00 ± 23.09	78.00 ± 19.89	C=100, γ= 0.01
	Csub	71.00 ± 13.00	70.00 ± 17.00	72.00 ± 23.48	C=1000, γ=1
Frase 6	Centroides	70.00 ± 12.65	70.00 ± 19.44	70.00 ± 19.44	C=1000, γ=0.1
	Csub	70.00 ± 10.95	74.00 ± 18.97	66.00 ± 23.19	C=100, γ=100

Csub: centroides calculados sólo con la mejor subbanda

En cambio, la distribución PWVD obtuvo la mejor eficiencia del 73 %, tanto para los centroides calculados sobre todo el espectro y aquellos calculados sólo con la mejor subbanda, diferenciándose sólo en su desviación como se muestra en la tabla. En cuanto a las medidas de rendimiento, la sensibilidad y la especificidad varían de una frase a otra y no se puede determinar de manera general cuál presenta un valor mayor, pero se podría decir que hay una tendencia a tener una sensibilidad mayor que la especificidad, sobre todo en el caso de la PWVD.

La tabla 4.15 muestra la mejor subbanda de los centroides para cada una de las representaciones, de acuerdo al rendimiento en términos del porcentaje de clasificación.

Al analizar los resultados obtenidos cuando se utilizan los centroides, es evidente como la mayoría de los resultados presentan las mejores eficiencias cuando se implementan los centroides sobre todo el espectro, en lugar de los centroides calculados sólo con la mejor subbanda; aunque la eficiencia difiere sólo entre 1 y 3 puntos porcentuales. Es importante resaltar que, la subbanda con mayor capacidad discriminante en la mayoría de los casos

Tabla 4.15: Mejores subbandas de los centroides para las TFD.

	Frase 1	Frase 2	Frase 3	Frase 4	Frase 5	Frase 6
MS	1	5	2	1	1	20
WV	7	1	1	1	4	3
PWV	1	4	4	3	5	6
SPWV	1	4	11	9	4	6
CW	1	1	1	1	6	1

corresponde a la primera banda.

Por otra parte, con el objetivo de presentar los resultados en una forma más compacta, en la figura 4.3 se encuentran las curvas ROC de los mejores resultados obtenidos con los centroides de los MS y de PWVD, con sus correspondientes AUC dentro del recuadro.

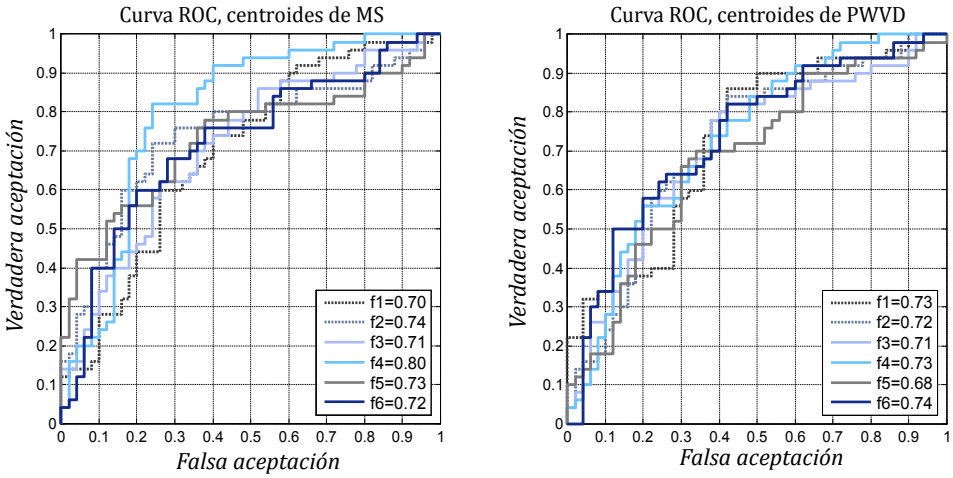

Figura 4.3: Curvas ROC y AUC estimadas a partir de los centroides de cada una de las frases (f).

Las curvas ROC presentan el mismo comportamiento que se evidencia en las tablas, en el caso de los MS la frase 4 obtuvo el mejor valor de AUC con 0.80, y en el caso de la PWVD el mejor resultado fue para la frase 6 con 0.74, diferenciándose del resultado presentado en la tabla. Se observa que los MS presentan un mejor rendimiento durante la detección y, en general, los resultados de las AUC mantienen un valor similar a tasas de acierto.

Posteriormente, se procede a presentar los experimentos con las demás características dinámicas propuestas en este trabajo, como son los marginales de tiempo (mgt) y de frecuencia (mgf), la energía, la FI, el EBW, los LFCC y MFCC; con el objetivo de disminuir el volumen inicial de datos y conseguir la mayor concentración de información de los espectros. A diferencia de las características anteriores, estas se estiman sobre todo el espectro y no sobre subbandas.

A continuación, en las tablas se presenta un resumen de todos los experimentos realizados con cada conjunto de características dinámicas y por cada TFD en las 6 frases, con el objetivo de realizar una comparación directa de todas las estimaciones de características

dinámicas implementadas y poder determinar cuáles aportan mayor información durante el proceso de clasificación. En cada una de las tablas se encuentra en negrilla los mejores resultados obtenidos en cada frase y son resaltadas las representaciones TF con mejor desempeño.

Los marginales de tiempo y frecuencia que se extraen de las TFD, tienen información de la energía de la señal a medida que transcurre el tiempo (o la frecuencia). En la tabla 4.16, se presenta un resumen de las eficiencias obtenidas mediante la implementación de los marginales en cada una de las frases, calculados sobre todo el espectro de las diferentes TFD.

Tabla 4.16: % Eficiencia obtenida con los marginales de tiempo y de frecuencia.

		MS	WVD	PWVD	SPWVD	CWD
Frase 1	mgt	54,00 ± 11,14	60,00 ± 10,95	62,00 ± 11,66	61,00 ± 16,40	**66,00 ± 15,62**
	mgf	**66,00 ± 14,97**	65,00 ± 14,32	65,00 ± 9,22	59,00 ± 10,44	64,00 ± 11,14
Frase 2	mgt	55,00 ± 6,71	67,00 ± 12,69	**68,00 ± 14,00**	55,00 ± 9,22	66,00 ± 14,28
	mgf	**68,00 ± 14,00**	66,00 ± 15,62	64,00 ± 11,14	54,00 ± 4,90	62,00 ± 8,72
Frase 3	mgt	61,00 ± 7,00	**68,00 ± 9,80**	67,00 ± 11,87	59,00 ± 17,00	64,00 ± 12,81
	mgf	**68,00 ± 13,27**	66,00 ± 13,56	61,00 ± 8,31	56,00 ± 21,81	61,00 ± 9,43
Frase 4	mgt	62,00 ± 14,00	58,00 ± 11,66	62,00 ± 14,00	60,00 ± 10,00	**65,00 ± 15,65**
	mgf	**66,00 ± 13,56**	60,00 ± 11,83	62,00 ± 14,00	60,00 ± 6,32	63,00 ± 16,76
Frase 5	mgt	57,00 ± 12,69	**69,00 ± 9,43**	58,00 ± 9,80	56,00 ± 4,90	64,00 ± 8,00
	mgf	**67,00 ± 11,87**	66,00 ± 13,56	61,00 ± 8,31	63,00 ± 6,40	62,00 ± 13,27
Frase 6	mgt	53,00 ± 13,45	55,00 ± 10,25	**57,00 ± 11,00**	55,00 ± 6,71	55,00 ± 12,04
	mgf	**62,00 ± 15,36**	58,00 ± 11,66	57,00 ± 15,52	53,00 ± 4,58	53,00 ± 12,69

mgt=marginales de tiempo. mgf=marginales de frecuencia

De acuerdo a los resultados, los marginales de frecuencia presentan los mejores resultados cuando son implementados sobre los espectros de los MS, con la mejor tasa de acierto del 68 % en la frase 3. En el caso de los marginales de tiempo, la WVD fue la representación TF que mejor eficiencia tuvo, con un 69 % en la frase 5. Se puede observar que las eficiencias obtenidas con los marginales son más bajas que las obtenidas con los centroides e incluso la mayoría se encuentran por debajo de los valores obtenidos con las vocales, por lo que se puede inferir que esta característica dinámica no tiene una gran capacidad discriminante en la tarea de detección.

En la tabla 4.17, se presentan las medidas de rendimiento y los mejores parámetros del clasificador, de los marginales de frecuencia obtenidos de los MS, ya que fue la transformación que mejores resultados alcanzó.

Tabla 4.17: Resultados obtenidos mediante los marginales de frecuencia de los MS.

	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	66.00 ± 14.97	62.00 ± 19.89	70.00 ± 23.57	C=1, γ=10
Frase 2	68.00 ± 14.00	74.00 ± 16.47	62.00 ± 23.94	C=100, γ=1
Frase 3	**68.00 ± 13.27**	68.00 ± 19.32	68.00 ± 19.32	C=1000, γ=1
Frase 4	66.00 ± 13.56	52.00 ± 21.50	80.00 ± 16.33	C=0.1, γ=100
Frase 5	67.00 ± 11.87	46.00 ± 25.03	88.00 ± 19.32	C=0.1, γ=100
Frase 6	62.00 ± 15.36	52.00 ± 28.60	72.00 ± 19.32	C=1, γ=100

Los marginales de frecuencia en los MS alcanzaron un 68 % en la frase 3, y en cuanto a las

medidas de rendimiento, se puede decir que la especificidad obtuvo un mejor rendimiento que la sensibilidad.

Con el fin de medir la concentración de energía presente en cada una de las TFD, se calculó la energía a partir del marginal de tiempo de cada TFD. En la tabla 4.18, se presentan las eficiencias obtenidas en las diferentes frases.

Tabla 4.18: % Eficiencia obtenida con la energía.

	MS	WVD	PWVD	SPWVD	CWD
Frase 1	55,00 ± 12,85	**72,00 ± 14,70**	69,00 ± 15,13	68,00 ± 14,70	68,00 ± 17,78
Frase 2	55,00 ± 10,25	68,00 ± 15,36	67,00 ± 15,52	68,00 ± 14,00	**69,00 ± 16,40**
Frase 3	59,00 ± 11,36	70,00 ± 17,32	67,00 ± 17,92	**72,00 ± 12,49**	72,00 ± 17,78
Frase 4	67,00 ± 15,52	69,00 ± 17,00	**70,00 ± 14,83**	65,00 ± 16,28	**70,00 ± 14,83**
Frase 5	55,00 ± 12,85	65,00 ± 10,25	66,00 ± 8,00	64,00 ± 11,14	**69,00 ± 13,75**
Frase 6	56,00 ± 14,28	68,00 ± 20,88	**69,00 ± 13,75**	63,00 ± 11,87	68,00 ± 19,90

La representación CWD fue la distribución que en promedio presentó el mejor desempeño, aunque la mejor eficiencia fue alcanzada por la SPWVD en la frase 3 con 72 %. Además, se puede observar que el desempeño de la energía es mejor que el de los marginales, sin embargo, sus eficiencias siguen siendo más bajas que las obtenidas con los centroides.

Luego de determinar la TFD que mejor rendimiento tuvo, en la tabla 4.19 se presentan los resultados en términos de eficiencia, sensibilidad, especificidad y los mejores parámetros del clasificador para la distribución CWD.

Tabla 4.19: Resultados obtenidos mediante la energía de CWD.

	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	68.00 ± 17.78	58.00 ± 22.01	78.00 ± 28.98	C=10, γ=0.001
Frase 2	69.00 ± 16.40	62.00 ± 17.51	76.00 ± 27.97	C=10, γ=0.01
Frase 3	72.00 ± 17.78	64.00 ± 24.59	80.00 ± 28.28	C=1, γ=0.1
Frase 4	**70.00 ± 14.83**	60.00 ± 24.94	80.00 ± 24.94	C=10, γ=1
Frase 5	69.00 ± 13.75	50.00 ± 19.44	88.00 ± 21.50	C=1, γ=1
Frase 6	68.00 ± 19.90	62.00 ± 25.73	74.00 ± 31.34	C=1000, γ=0.001

La energía de CWD obtuvo la mejor eficiencia en la frase 4 con un 70 %, y la especificidad alcanzó porcentajes más altos que la sensibilidad.

Por lo que se refiere a EBW, en la tabla 4.20 se presentan los porcentajes de eficiencia obtenidos para cada frase con las diferentes TFD.

Tabla 4.20: % Eficiencia obtenida con EBW.

	MS	WVD	PWVD	SPWVD	CWD
Frase 1	**68,00 ± 9,80**	63,00 ± 11,00	57,00 ± 11,87	63,00 ± 13,45	62,00 ± 16,00
Frase 2	**70,00 ± 10,95**	64,00 ± 12,00	60,00 ± 14,14	56,00 ± 9,17	65,00 ± 9,22
Frase 3	**66,00 ± 16,85**	56,00 ± 11,14	57,00 ± 13,45	61,00 ± 15,78	67,00 ± 11,87
Frase 4	**68,00 ± 8,72**	56,00 ± 9,17	56,00 ± 15,62	63,00 ± 7,81	61,00 ± 13,00
Frase 5	67,00 ± 17,35	54,00 ± 6,63	60,00 ± 6,32	58,00 ± 15,36	**70,00 ± 11,83**
Frase 6	65,00 ± 15,00	61,00 ± 13,75	59,00 ± 10,44	60,00 ± 10,95	**65,00 ± 13,60**

Se puede evidenciar que EBW al igual que los marginales y la energía, no presentan un buen

rendimiento en la tarea de detección de la EP, en habla continua.

La representación TF que en general obtuvo las tasas más altas es MS, la tabla 4.21 presenta las medidas de rendimiento y los mejores parámetros del clasificador, obtenidos con los anchos de banda.

Tabla 4.21: Resultados obtenidos mediante EBW de los MS.

	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	68.00 ± 9.80	64.00 ± 15.78	72.00 ± 13.98	C=1000, γ=1
Frase 2	**70.00 ± 10.95**	70.00 ± 19.44	70.00 ± 17.00	C=100, γ=0.1
Frase 3	66.00 ± 16.85	76.00 ± 20.66	56.00 ± 27.97	C=1, γ=0.0001
Frase 4	68.00 ± 8.72	52.00 ± 16.87	84.00 ± 12.65	C=10, γ=1
Frase 5	67.00 ± 17.35	60.00 ± 23.09	74.00 ± 28.36	C=1000, γ=1
Frase 6	61.00 ± 13.75	62.00 ± 25.71	59.00 ± 23.30	C=1, γ=100

La frase 2 alcanzó la mejor eficiencia con un 70 %, y del mismo modo que con la energía, la especificidad obtuvo porcentajes más altos que la sensibilidad.

En relación con la FI, los resultados obtenidos mediante esta característica dinámica se presentan en la tabla 4.22 con cada una de las frases.

Tabla 4.22: % Eficiencia obtenida con FI.

	MS	WVD	PWVD	SPWVD	CWD
Frase 1	71,00 ± 15,13	71,00 ± 12,21	70,00 ± 12,65	**71,00 ± 10,44**	71,00 ± 12,21
Frase 2	66,00 ± 14,97	72,00 ± 13,27	69,00 ± 13,75	**73,00 ± 13,45**	70,00 ± 15,49
Frase 3	**72,00 ± 12,49**	68,00 ± 10,77	71,00 ± 9,43	70,00 ± 14,14	69,00 ± 11,36
Frase 4	64,00 ± 14,28	**70,00 ± 12,65**	67,00 ± 9,00	67,00 ± 17,35	67,00 ± 12,69
Frase 5	68,00 ± 11,66	**70,00 ± 10,00**	68,00 ± 10,77	66,00 ± 12,81	70,00 ± 12,65
Frase 6	61,00 ± 12,21	**69,00 ± 13,75**	66,00 ± 14,28	68,00 ± 13,27	66,00 ± 15,62

La representación WVD en promedio presentó el mejor desempeño, sin embargo, la mejor eficiencia fue obtenida por la SPWVD en la frase 2, con un 73 %. Se observa que las eficiencias obtenidas con FI son comparables con las obtenidas con los centroides, aunque en general siguen siendo mejores los resultados de los centroides.

En la tabla 4.23, se presentan los resultados obtenidos con la FI de WVD, en términos de eficiencia, sensibilidad, especificidad y los mejores parámetros del clasificador.

Tabla 4.23: Resultados obtenidos mediante la FI de WVD.

	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	71.00 ± 12.21	56.00 ± 20.66	86.00 ± 13.50	C=100, γ=10
Frase 2	**72.00 ± 13.27**	56.00 ± 22.71	88.00 ± 13.98	C=10, γ=0.0001
Frase 3	68.00 ± 10.77	66.00 ± 18.97	70.00 ± 14.14	C=0.1, γ=0.001
Frase 4	70.00 ± 12.65	60.00 ± 21.08	80.00 ± 9.43	C=0.1, γ=0.01
Frase 5	68.00 ± 10.77	64.00 ± 13.29	72.00 ± 13.98	C=10, γ=0.001
Frase 6	69.00 ± 13.75	54.00 ± 21.19	84.00 ± 15.78	C=1, γ=10

La frase 2 alcanzó la mejor eficiencia con un 72 %, y del mismo modo que con las características dinámicas anteriores, la especificidad presenta porcentajes más altos que la sensibilidad.

Para finalizar, en la tabla 4.24 se presentan los experimentos realizados con los coeficientes cepstrales, donde se encuentran las eficiencias alcanzadas con los LFCC y MFCC

en cada una de las frases. Se debe resaltar que para la implementación de los coeficientes cepstrales, se seleccionó el mejor número de coeficientes de acuerdo a los criterios de error de clasificación.

Tabla 4.24: % Eficiencia obtenida con los coeficientes cepstrales.

		MS	WVD	PWVD	SPWVD	CWD
Frase 1	LFCC	73,00 ± 9,00	**74,00 ± 6,63**	71,00 ± 14,46	72,00 ± 14,00	74,00 ± 8,00
	MFCC	73,00 ± 11,00	75,00 ± 13,60	70,00 ± 12,65	74,00 ± 9,17	**75,00 ± 10,25**
Frase 2	LFCC	73,00 ± 6,40	70,00 ± 10,00	71,00 ± 9,43	67,00 ± 13,45	**74,00 ± 14,97**
	MFCC	73,00 ± 14,18	71,00 ± 7,00	73,00 ± 10,05	69,00 ± 11,36	**75,00 ± 10,25**
Frase 3	LFCC	70,00 ± 14,83	71,00 ± 15,13	73,00 ± 10,05	65,00 ± 12,04	**75,00 ± 10,25**
	MFCC	69,00 ± 10,44	71,00 ± 10,44	70,00 ± 12,65	65,00 ± 12,04	**74,00 ± 12,81**
Frase 4	LFCC	74,00 ± 10,20	73,00 ± 11,87	73,00 ± 12,69	67,00 ± 17,35	**77,00 ± 12,69**
	MFCC	**75,00 ± 10,25**	70,00 ± 7,75	69,00 ± 10,44	68,00 ± 17,20	72,00 ± 11,66
Frase 5	LFCC	**74,00 ± 12,00**	72,00 ± 14,70	69,00 ± 9,43	72,00 ± 11,66	73,00 ± 13,45
	MFCC	75,00 ± 16,88	71,00 ± 14,46	69,00 ± 13,00	72,00 ± 11,66	**75,00 ± 15,65**
Frase 6	LFCC	71,00 ± 8,31	75,00 ± 10,25	64,00 ± 12,00	70,00 ± 14,14	**73,00 ± 6,40**
	MFCC	72,00 ± 10,77	72,00 ± 10,77	68,00 ± 12,49	68,00 ± 10,77	**76,00 ± 9,17**

Se puede apreciar que la representación CWD fue la que en general presentó el mejor desempeño. En el caso de los LFCC la tasa de acierto más alta fue de 77 % obtenida en la frase 4, mientras que los MFCC alcanzaron un 76 % en la frase 4. Los resultados muestran que la capacidad discriminante de los coeficientes cepstrales es superior que la mayoría de características dinámicas, con porcentajes de eficiencia por encima del 70 % en la mayoría de frases.

En la tabla 4.25, se presenta los resultados obtenidos por los coeficientes cepstrales de la representación CWD, en términos de eficiencia, sensibilidad, especificidad y los mejores parámetros del clasificador.

Tabla 4.25: Resultados obtenidos con los coeficientes cepstrales de CWD.

		% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	LFCC	74.00 ± 8.00	74.00 ± 16.47	74.00 ± 18.97	C=1000, γ=0.01
	MFCC	75.00 ± 10.25	66.00 ± 16.47	84.00 ± 15.78	C=100, γ=0.1
Frase 2	LFCC	74.00 ± 14.97	78.00 ± 22.01	70.00 ± 19.44	C=1000, γ=0.1
	MFCC	75.00 ± 10.25	76.00 ± 18.38	74.00 ± 18.97	C=1000, γ=0.01
Frase 3	LFCC	75.00 ± 10.25	74.00 ± 16.47	76.00 ± 18.38	C=10, γ=1
	MFCC	74.00 ± 12.81	64.00 ± 18.38	84.00 ± 15.78	C=100, γ=0.1
Frase 4	LFCC	**77.00 ± 12.69**	72.00 ± 21.50	82.00 ± 22.01	C=1000, γ=0.1
	MFCC	72.00 ± 11.66	68.00 ± 21.50	76.00 ± 18.38	C=100, γ=0.1
Frase 5	LFCC	73.00 ± 13.45	74.00 ± 16.47	72.00 ± 19.32	C=100, γ=1
	MFCC	75.00 ± 15.65	70.00 ± 21.60	80.00 ± 18.86	C=100, γ=0.1
Frase 6	LFCC	73.00 ± 6.40	62.00 ± 14.76	84.00 ± 12.65	C=100, γ=0.1
	MFCC	**76.00 ± 9.17**	68.00 ± 25.30	84.00 ± 20.66	C=1000, γ=0.1

Después de haber presentado los experimentos y analizado sus respectivos resultados, nuevamente se reafirma el hecho de que la información que contiene la señal de voz continua tiene mayor capacidad discriminante en la detección de Parkinson. En donde las características que mejor desempeño obtuvieron fueron los coeficientes cepstrales (CWD), los centroides (MS) y la FI (CWD). Mientras que, los marginales, la energía y EBW presentan

los desempeños más bajos.

Adicionalmente, se presentan en la figura 4.4 las curvas ROC de los coeficientes cepstrales (LFCC y MFCC) obtenidos mediante la representación CWD, que fue la representación con uno de los mejores resultados, como se indicó en las tablas anteriores. Además, se incluyen las correspondientes AUC dentro del recuadro.

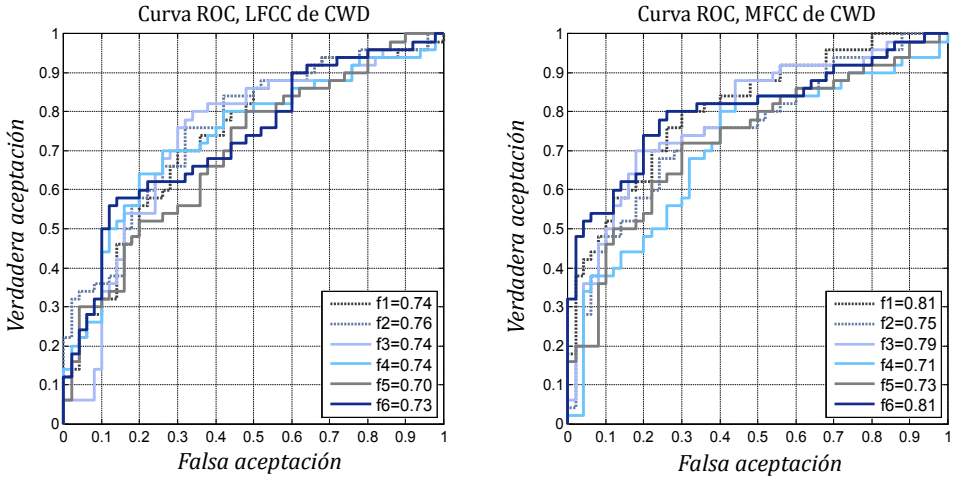

Figura 4.4: Curvas ROC y AUC estimadas a partir de los LFCC y MFCC para cada una de las frases (f).

Las curvas ROC de los LFCC presentan un buen desempeño, con la mejor AUC de 0.76 para la frase 2 y se puede decir que, los resultados obtenidos con las AUC son similares a las eficiencias presentadas. En cuanto a los MFCC, se observa un gran desempeño, con un rendimiento mayor que con las demás características, en donde los valores de AUC para las frases 1 y 6 fue de 0.81.

En relación con la WPT, se realizó una descomposición del nivel 2 al 5, en donde se realizaron experimentos para cada uno de los niveles de manera independiente y también se consideraron todos los coeficientes de los diferentes niveles. En los cuales, se aplicaron un conjunto de características basadas en energía y medidas de información, como son la energía normalizada, la energía de TEO y la energía de Shannon, además de sus respectivos logaritmos. En la tabla 4.26, se presentan los resultados obtenidos por la mejor descomposición, debido a que se hizo una clasificación por cada nivel.

Tabla 4.26: Resultados obtenidos por la WPT.

	% Eficiencia	**% Precisión**	**% Sensibilidad**	**% Especificidad**	**Parámetros**
Frase 1	73.00 ± 9.00	77.61 ± 13.82	72.00 ± 23.48	74.00 ± 21.19	C=100, γ=0.001
Frase 2	76.00 ± 15.62	75.98 ± 19.13	74.00 ± 26.75	78.00 ± 14.76	C=1000, γ=0.001
Frase 3	77.00 ± 9.00	76.48 ± 10.97	80.00 ± 18.86	74.00 ± 13.50	C=100, γ=0.001
Frase 4	**77.00 ± 6.40**	79.87 ± 12.08	78.00 ± 22.01	76.00 ± 18.38	C=100, γ=0.001
Frase 5	71.00 ± 11.36	79.17 ± 16.78	60.00 ± 21.08	82.00 ± 14.76	C=10, γ=0.001
Frase 6	75.00 ± 8.06	74.45 ± 5.59	76.00 ± 20.66	74.00 ± 9.66	C=10, γ=0.01

La mejor eficiencia presentada es para la frase 4 con 77 %, se evidencia cómo todos los resultados obtenidos son altos con tasas por encima del 70 %, resultados similares a los obtenidos con los MFCC y LFCC con las demás TFD. En cuanto a la sensibilidad y especificidad, sus valores son muy similares, presentando un buen equilibrio entre la detección de los pacientes con EP y aquellos que no la padecen.

Fusión de información

En esta sección se encuentran los resultados relacionados con la combinación de información a través de las tres estrategias de fusión presentadas en el capítulo anterior, con el fin de establecer si la información obtenida a partir de las diferentes técnicas es complementaria, y cuál es el nivel de eficiencia que puede ser obtenido a partir de su combinación.

- **Fusión a nivel de características**

Primero se presentarán los resultados obtenidos mediante la fusión a nivel de características, la cual fue realizada con diferentes conjuntos de características dinámicas obtenidas de los espectros de las 5 TFD: MS, WVD, PWVD, SPWVD y CWD. Es importante aclarar que, la información de la WPT no fue utilizada en esta estrategia, debido a que como no fue implementada bajo las mismas condiciones, los vectores de características no son compatibles. Se realizaron varios experimentos (Exp), en los cuales se hizo una concatenación de los conjuntos de características para cada TFD de manera independiente, de la siguiente manera:

- Exp 1 = FI, EBW, LFCC y MFCC.
- Exp 2 = mejor subbanda de los centroides, FI, EBW, energía, marginales, LFCC y MFCC.
- Exp 3 (“todas”) = centroides, FI, EBW, energía, marginales, LFCC y MFCC.
- Exp 4 (“mejores”) = centroides, FI, LFCC y MFCC.

A continuación, se presenta un resumen de los mejores resultados obtenidos con dos de los experimentos propuestos: el Exp 3, utilizando “todas” las características para cada representación TF de manera independiente; y el Exp 4, utilizando sólo las “mejores” características, es decir, aquellas que obtuvieron mejores tasas de aciertos de acuerdo a las tablas 4.16 a la 4.24. En la tabla 4.27, se presentan los porcentajes de eficiencia de la fusión a nivel de características, para cada una de las TFD y las 6 frases. Se encuentran en negrilla las eficiencias más altas en cada frase y se resaltan las TFD Y Frases con mejor desempeño.
La tasa de acierto más alta fue de 74 % cuando se utilizaron “todas” las características y las “mejores” características, con SPWVD y WVD, respectivamente en la frase 6. En general, las eficiencias de las TFD tienen valores muy similares, pero en promedio las dos distribuciones que mejor comportamiento presentaron fueron la SPWVD y la CWD. Además, se observa que en la mayoría de TFD y frases, los mejores resultados se obtuvieron cuando se concatenaron las “mejores” características, con valores de eficiencia por encima de 1 y hasta 6 puntos porcentuales de los obtenidos con “todas” las características. Sin embargo, es evidente que

Tabla 4.27: % Eficiencia de la fusión a nivel de características.

	Características	MS	WVD	PWVD	SPWVD	CWD
Frase 1	Todas	69.00 ± 10.44	70.00 ± 13.42	67.00 ± 14.87	71.00 ± 9.43	**71.00 ± 10.44**
	Mejores	69.00 ± 10.44	70.00 ± 13.42	68.00 ± 14.70	70.00 ± 14.83	72.00 ± 11.66
Frase 2	Todas	**71.00 ± 10.44**	65.00 ± 11.18	70.00 ± 14.14	69.00 ± 12.21	68.00 ± 8.72
	Mejores	71.00 ± 13.75	66.00 ± 12.00	70.00 ± 12.65	69.00 ± 12.21	71.00 ± 14.46
Frase 3	Todas	66.00 ± 15.62	70.00 ± 10.95	**71.00 ± 10.44**	70.00 ± 8.94	**71.00 ± 10.44**
	Mejores	68.00 ± 11.66	71.00 ± 16.40	**70.00 ± 10.95**	71.00 ± 9.43	**73.00 ± 10.05**
Frase 4	Todas	70.00 ± 12.65	65.00 ± 11.18	66.00 ± 11.14	71.00 ± 17.00	70.00 ± 17.32
	Mejores	72.00 ± 8.72	65.00 ± 12.04	66.00 ± 11.14	68.00 ± 13.27	71.00 ± 13.75
Frase 5	Todas	69.00 ± 12.21	**71.00 ± 12.21**	68.00 ± 15.36	72.00 ± 16.00	69.00 ± 15.78
	Mejores	**72.00 ± 12.49**	**74.00 ± 11.14**	67.00 ± 11.00	73.00 ± 15.52	69.00 ± 14.46
Frase 6	Todas	69.00 ± 11.36	68.00 ± 12.49	67.00 ± 10.05	**74.00 ± 11.14**	70.00 ± 10.00
	Mejores	72.00 ± 15.36	**74.00 ± 11.14**	69.00 ± 14.46	**73.00 ± 14.18**	71.00 ± 11.36

los resultados no fueron más altos que los obtenidos al clasificar de manera independiente cada una de las características dinámicas.

Después de haber identificado las dos TFD que aportan mayor información de acuerdo a las tasas de clasificación. En la tabla 4.28 se presentan los resultados obtenidos por SPWVD cuando se realiza fusión de "todas" las características, y en la tabla 4.29, se presentan los resultados obtenidos por CWD cuando se realiza la fusión de las "mejores" características.

Tabla 4.28: Resultados de la fusión a nivel de todas características para la SPWVD.

	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	71.00 ± 9.43	52.00 ± 19.32	90.00 ± 10.54	C=0.1, γ=0.01
Frase 2	69.00 ± 12.21	58.00 ± 19.89	80.00 ± 16.33	C=1, γ=0.001
Frase 3	70.00 ± 8.94	58.00 ± 19.89	82.00 ± 17.51	C=1, γ=0.001
Frase 4	71.00 ± 17.00	70.00 ± 25.39	72.00 ± 27.00	C=10, γ=0.001
Frase 5	72.00 ± 16.00	72.00 ± 19.32	72.00 ± 23.48	C=0.1, γ=0.0001
Frase 6	**74.00 ± 11.14**	72.00 ± 25.30	76.00 ± 12.65	C=1, γ=0.01

En el caso de SPWVD la eficiencia más alta es de 74 % en la frase 6, y en cuanto a las demás medidas de rendimiento, se puede observar que en la mayoría de los casos, la especificidad tiene porcentajes más altos que los de sensibilidad, lo cual indica la capacidad del sistema para detectar correctamente las personas sanas.

Tabla 4.29: Resultados de la fusión a nivel de las mejores características para la CWD.

	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Frase 1	72.00 ± 11.66	86.00 ± 16.47	58.00 ± 22.01	C=1, γ=0.0001
Frase 2	71.00 ± 14.46	64.00 ± 22.71	78.00 ± 14.76	C=10, γ=0.01
Frase 3	**73.00 ± 10.05**	64.00 ± 18.38	82.00 ± 17.51	C=10, γ=0.0001
Frase 4	71.00 ± 13.75	62.00 ± 23.94	80.00 ± 16.33	C=1, γ= 0.01
Frase 5	69.00 ± 14.46	58.00 ± 22.01	80.00 ± 18.86	C=1, γ= 0.01
Frase 6	71.00 ± 11.36	74.00 ± 18.97	68.00 ± 19.32	C=1, γ=0.001

Mientras que, para CWD la eficiencia más alta es de 73 % en la frase 3, y los porcentajes de especificidad tienden a ser más altos respecto a la sensibilidad.

Al igual que los experimentos anteriores, también se presentan las curvas ROC. En la figura

4.5 se encuentran las curvas ROC de las distribuciones SPWVD y CWD, con sus correspondientes AUC dentro del recuadro.

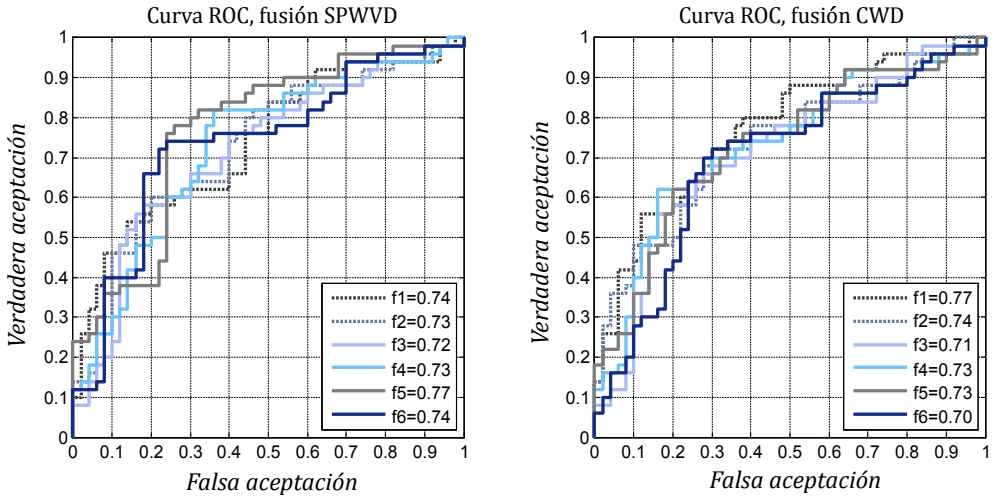

Figura 4.5: Curvas ROC y AUC estimadas a partir de la fusión a nivel de características de cada una de las frases (f).

Se observa en las curvas ROC que, la SPWVD tiene un mejor desempeño que el de CWD, con el mejor valor de AUC de 0.77 en la frase 5. A pesar de que las curvas ROC obtenidas con CWD no muestran un gran desempeño, las AUC alcanzan valores muy similares que los de SPWVD, con la mejor AUC de 0.77 en la frase 1. Se puede evidenciar que las AUC no presentan un comportamiento similar al de las tablas.

- **Fusión a nivel de *scores***

En segundo lugar, se presentarán los resultados obtenidos mediante la fusión a nivel de *scores*, en donde se utilizaron los conjuntos de características dinámicas obtenidas de las 6 TFD, para alimentar una segunda etapa de clasificación.

Se llevaron a cabo diferentes experimentos, primero se realizó la fusión a nivel de *scores* de las 6 frases para cada una de las representaciones TF, teniendo en cuenta las características dinámicas que mejor desempeño presentaron en los experimentos anteriores, eligiendo como características los coeficientes cepstrales (MFCC y LFCC), los centroides y la FI, además se utilizaron los conjuntos de características conformados por la unión de “todas” características y las “mejores” características.

La tabla 4.30, presentan los porcentajes de eficiencia de la fusión a nivel de *scores*, con los seis conjuntos de características elegidos y para cada una de las TFD. Se encuentran en negrilla las eficiencias más altas para cada TFD, y se resaltan los conjuntos de características y TFD con mejor desempeño.
Se observa que la mejor tasa de acierto fue del 79 %, la cual se obtiene con los MFCC de la distribución CWD; seguida por la fusión de las “mejores” características que alcanzan una

Tabla 4.30: % Eficiencia de la fusión de las frases a nivel de *scores*.

Características	MS	WVD	PWVD	SPWVD	CWD
Centroides	72.00 ± 15.36	70.00 ± 10.00	71.00 ± 11.36	70.00 ± 13.42	70.00 ± 10.95
FI	70.00 ± 12.65	**72.00 ± 10.77**	71.00 ± 12.21	**72.00 ± 10.77**	76.00 ± 8.00
LFCC	72.00 ± 14.00	70.00 ± 7.75	67.00 ± 15.52	69.00 ± 16.40	72.00 ± 9.80
MFCC	71.00 ± 13.00	71.00 ± 12.21	71.00 ± 11.36	68.00 ± 15.36	**79.00 ± 12.21**
Todas	73.00 ± 10.05	69.00 ± 17.00	**71.00 ± 10.44**	71.00 ± 12.21	69.00 ± 13.00
Mejores	**74.00 ± 10.20**	71.00 ± 12.21	70.00 ± 11.83	74.00 ± 10.20	70.00 ± 14.83

tasa de acierto del 74 % con los MS, y en tercer lugar, se encuentra la FI con un 72 % tanto para la distribución WVD como para SPWVD. Los resultados obtenidos son superiores a los obtenidos con la estrategia de fusión de características, ya la mayoría de resultados presentan tasas por encima del 70 %, más aún presentan un desempeño similar a los resultados obtenidos de manera individual por cada conjunto de características e incluso un poco mejor en algunos casos.

En la tabla 4.31 se presentan las medidas de rendimiento de las características con mejor desempeño para cada representación TF, además se incluye la fusión de *scores* de las frases con la información obtenida mediante la WPT. Es importante aclarar que, en el caso de la

Tabla 4.31: Resultados de la fusión a nivel de *scores*.

	Características	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
MS	FI	70.00 ± 12.65	74.00 ± 23.19	66.00 ± 9.66	C=0.1, γ=10
	MFCC	71.00 ± 13.00	74.00 ± 18.97	68.00 ± 19.32	C=100, γ=0.001
WVD	FI	**72.00 ± 10.77**	68.00 ± 16.87	76.00 ± 15.78	C=1000, γ=0.0001
	MFCC	71.00 ± 12.21	74.00 ± 21.19	68.00 ± 21.50	C=1, γ=10
PWVD	FI	71.00 ± 12.21	68.00 ± 27.00	74.00 ± 13.5	C=10, γ=0.1
	MFCC	**71.00 ± 11.36**	86.00 ± 13.50	56.00 ± 22.71	C=1000, γ=1
SPWVD	FI	**72.00 ± 10.77**	64.00 ± 18.38	80.00 ± 16.33	C=1000, γ=0.0001
	MFCC	68.00 ± 15.36	70.00 ± 21.6	66.00 ± 21.19	C=0.1, γ=10
CWD	FI	76.00 ± 8.00	76.00 ± 15.78	76.00 ± 15.78	C=1000, γ=0.001
	MFCC	**79.00 ± 12.21**	70.00 ± 21.60	88.00 ± 13.98	C=100, γ=0.1
WPT	**"Energías"**	**82,00 ± 7,48**	84,00 ± 15,78	80,00 ± 16,33	C=1000, γ=0.001

WPT la energía como característica dinámica, se refiere a un con conjunto de características basadas en energía y medidas de información, como son la energía normalizada, la energía de TEO y la energía de Shannon, además de sus respectivos logaritmos.

Adicionalmente, la tabla 4.32 presenta la fusión de información de las seis frases con cada técnica TF, pero en este caso utilizando el mejor conjunto de características, este conjunto corresponde a la unión de la FI, los coeficientes cepstrales y los centroides; excepto por la WPT que se caracterizó solamente con niveles de energía. La WPT logró el mejor resultado con 82 %, el rendimiento más alto hasta el momento y que es de 8 puntos porcentuales por encima de otras tasas de acierto obtenidas durante la fusión de las frases, mientras que la fusión de las 6 distribuciones sólo alcanzó un 79 %.

Los resultados presentan una mejoría en el desempeño respecto a los obtenidos con la en la fusión a nivel de características, con valores de eficiencia por encima del 70 %. Por otra parte, la sensibilidad y especificidad de todo el sistema muestra un equilibrio en la detección

Tabla 4.32: Resultados de la fusión fusión a nivel de *scores* con el mejor conjunto de características.

	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
MS	74.0 ± 10.2	68.0 ± 13.9	80.0 ± 18.8	C=10, γ=0,01
WVD	71.0 ± 12.2	68.0 ± 25.3	74.0 ± 16.5	C=100, γ=0,01
PWVD	70.0 ± 11.8	52.0 ± 16.8	88.0 ± 16.8	C=0.1, γ=0,01
SPWVD	72.0 ± 10.8	64.0 ± 18.3	80.0 ± 16.3	C=1000, γ=0.001
CWD	70.0 ± 14.8	64.0 ± 26.3	76.0 ± 15.8	C=100, γ=0,0001
WPT	**82.0 ± 7.5**	84.0 ± 15.8	80.0 ± 16.3	C=1000, γ=0.01
Fusión	**79.0 ± 10.4**	80.0 ± 18.8	78.0 ± 14.7	C=100, γ=0.001

de personas con EP y las sanas. Además, hay que mencionar que la fusión de la información de las 6 frases mediante la WPT, presenta un resultado que supera cualquiera de los obtenidos durante la fusión, con una tasa de acierto del 82 % y una disminución en el porcentaje de varianza. En cuanto a los porcentajes de sensibilidad y especificidad, ambos superan el 80 %, manteniendo el mismo equilibro para la detección de personas con EP y las sanas.

Por otra parte, en la figura 4.6 se encuentran las curvas ROC del conjunto de las "mejores" características de los MS y los MFCC de la CWD, para cada una de las frases, además de la fusión de las 6 frases. Así mismo, en la figura 4.7 se presentan las curvas ROC de las características de la WPT para cada frase y para la fusión. Además, se incluyen las correspondientes AUC dentro del recuadro.

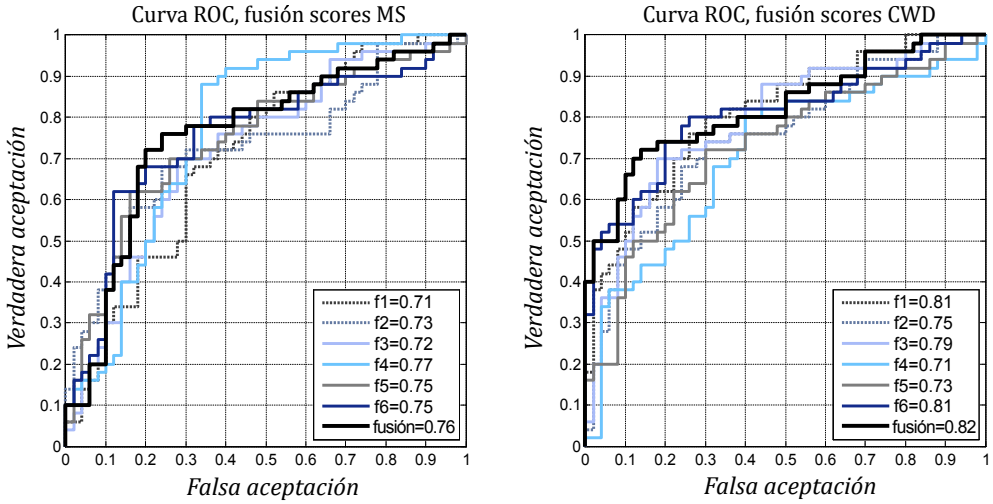

Figura 4.6: Curvas ROC y AUC estimadas a partir las "mejores" características de los MS y los MFCC de CWD, además de la fusión de las 6 frases (f).

Las curvas ROC, presentan el mismo comportamiento que se evidencia en los resultados de las tablas. En el caso de los MS, la fusión alcanza un valor de AUC de 0.76 y en el caso de la CWD, el valor de la AUC de la fusión fue de 0.82. Las curvas de ambas representaciones tienen un buen desempeño en cuanto a la fusión, aunque la CWD presenta un mejor desempeño y en la figura se evidencia como la fusión mejora el rendimiento al unir la información de las 6 frases. En el caso de la WPT, la fusión tiene un valor de AUC de 0.85, superando los valores

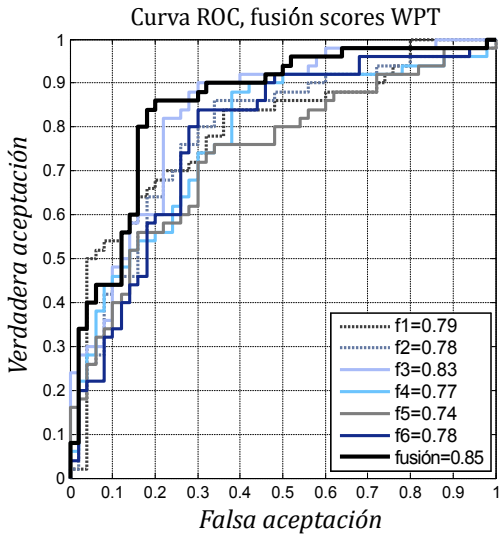

Figura 4.7: Curvas ROC y AUC estimadas a partir la WPT y la fusión de las 6 frases (f).

obtenidos anteriormente, y además, se observa como la fusión combina información valiosa de las 6 frases.

Por otro lado, se realizó una fusión a nivel de *scores* de 5 TFD (MS, WVD, PWVD, SPWVD y CWD) con cada conjunto de características dinámicas. En la tabla 4.33, se presenta un resumen de las eficiencias obtenidas para cada frase.

Tabla 4.33: % Eficiencia de la fusión de las TFD a nivel de *scores*.

Características	Frase 1	Frase 2	Frase 3	Frase 4	Frase 5	Frase 6
Centroides	72.00 ± 10.77	71.00 ± 15.13	68.00 ± 10.77	**77.00 ± 9.00**	75.00 ± 12.85	70.00 ± 7.75
FI	71.00 ± 11.36	69.00 ± 12.21	72.00 ± 8.72	68.00 ± 13.27	70.00 ± 11.83	68.00 ± 14.00
LFCC	71.00 ± 9.43	69.00 ± 10.44	**75.00 ± 9.22**	70.00 ± 10.95	70.00 ± 13.42	**75.00 ± 12.04**
MFCC	**73.00 ± 10.05**	**73.00 ± 11.87**	72.00 ± 10.77	72.00 ± 15.36	72.00 ± 14.70	74.00 ± 8.00
Todas	70.00 ± 10.00	71.00 ± 13.75	72.00 ± 9.80	70.00 ± 12.65	72.00 ± 14.00	69.00 ± 12.21
Mejores	70.00 ± 10.00	71.00 ± 12.21	71.00 ± 10.44	70.00 ± 10.00	**76.00 ± 12.81**	71.00 ± 14.46

Los resultados alcanzan valores de eficiencia por encima del 70 %, presentando mejoras respecto a los resultados obtenidos con la fusión de las a frases a nivel de *scores*. Los centroides obtuvieron la mejor tasa de acierto con un 77 % para la frase 4, y el conjunto de las “mejores” características con 76 % para las frases 5. Después de analizar el promedio de las eficiencias, las características que presentan el mejor comportamiento son los centroides, los MFCC y el conjunto de las “mejores” características; mientras que, la frase 5 presenta la mejor fusión de información, como se muestra en la tabla 4.34.

Las tasas más altas fueron obtenidas con el conjunto de las “mejores” características, con una tasa de acierto del 76 % y cuando se utilizan los centroides con un 75 %. En cuanto a la sensibilidad y especificidad, se muestran un equilibrio entre en la detección de personas con EP y las sanas, aunque hay una tendencia a que la especificidad sea mayor.

Tabla 4.34: Resultados de la fusión de la frase 5 a nivel de *scores*.

Características	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Centroides	75.00 ± 12.85	64.00 ± 18.38	86.00 ± 16.47	C=10, γ=0.1
FI	70.00 ± 11.83	68.00 ± 21.50	72.00 ± 13.98	C=0.1, γ=1
LFCC	70.00 ± 13.42	74.00 ± 16.47	66.00 ± 18.97	C=1000, γ=0.0001
MFCC	72.00 ± 14.70	72.00 ± 21.50	72.00 ± 27.00	C=1000, γ=0.001
Todas	72.00 ± 14.00	66.00 ± 16.47	78.00 ± 22.01	C=100, γ=0.001
Mejores	**76.00 ± 12.81**	74.00 ± 18.97	78.00 ± 22.01	C=100, γ=0.001

La figura 4.8 presenta las curvas ROC de los centroides obtenidos con la frase 4 y de las "mejores" características para la frase 5, y sus respectivas fusiones.

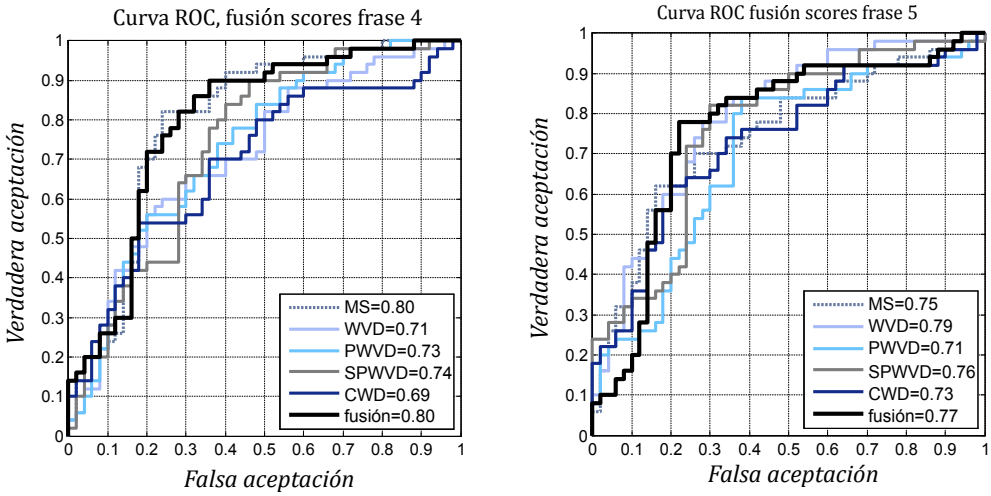

Figura 4.8: Curvas ROC y AUC estimadas utilizando los centroides para la frase 4 y las "mejores" características para la frase 5.

Las curvas ROC presentan el mismo comportamiento que se evidencia en las tablas, en el caso de la frase 4, la fusión tiene un valor de AUC de 0.80 y para la frase 5, la fusión fue de 0.77, las curvas de ambas representaciones presentan una mejora evidente en su desempeño durante la fusión.

Se debe agregar que, también se realizó una fusión a nivel de *scores* para cada conjunto de características, pero incluyendo la información obtenida por la WPT para cada frase, con el fin de realizar una comparación directa con la fusión anterior y determinar si la información entregada por la WPT puede aportar información valiosa, ya que los resultados obtenidos de manera individual para cada frase, fueron unos de los más altos.

En la tabla 4.35 se presentan los resultados de la fusión de *scores* junto con la información de la WPT. Los resultados que presentan el mejor comportamiento, son obtenidos por los LFCC y MFCC; con una tasa de acierto de 78 % para los LFCC+WPT de la frase 3 y 79 % para los MFCC+WPT de la frase 1. Es importante resaltar que, la influencia de la WPT mejora bastante la eficiencia de la fusión de cada conjunto de características, en promedio 4 %

Tabla 4.35: % Eficiencia de la fusión a nivel de *scores* de las TFD + la WPT.

Características	Frase 1	Frase 2	Frase 3	Frase 4	Frase 5	Frase 6
Centroides+WPT	77.00 ± 10.05	71.00 ± 9.43	78.00 ± 9.80	73.00 ± 13.45	**74.00 ± 14.28**	76.00 ± 12.81
FI+WPT	76.00 ± 9.17	75.00 ± 15.00	76.00 ± 6.63	70.00 ± 10.95	71.00 ± 9.43	**77.00 ± 11.87**
LFCC+WPT	77.00 ± 11.00	72.00 ± 14.00	**78.00 ± 8.72**	72.00 ± 12.49	68.00 ± 12.49	75.00 ± 12.04
MFCC+WPT	**79.00 ± 7.00**	72.00 ± 14.70	76.00 ± 11.14	71.00 ± 15.13	71.00 ± 10.44	75.00 ± 10.25
Todas+WPT	75.00 ± 8.06	**73.00 ± 10.05**	78.00 ± 10.77	73.00 ± 12.69	72.00 ± 11.66	72.00 ± 11.66
Mejores+WPT	77.00 ± 9.00	72.00 ± 13.27	77.00 ± 13.45	**74.00 ± 11.14**	72.00 ± 11.66	73.00 ± 15.52

puntos porcentuales por encima del valor inicial.

Después de analizar los resultados, la frase 1 presenta la mejor fusión de información en cada uno de los conjuntos de características como se muestra en la 4.36.

Tabla 4.36: Resultados de la fusión a nivel de *scores* + WPT para la frase 1.

Características	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
Centroides+WPT	77.00 ± 10.05	80.00 ± 18.86	74.00 ± 18.97	C=10, γ0.01
FI+WPT	76.00 ± 9.17	78.00 ± 19.89	74.00 ± 16.47	C=10, γ0.001
LFCC+WPT	77.00 ± 11.00	76.00 ± 12.65	78.00 ± 23.94	C=1000, γ0.0001
MFCC+WPT	**79.00 ± 7.00**	76.00 ± 18.38	82.00 ± 17.51	C=10, γ0.01
Todas+WPT	75.00 ± 8.06	74.00 ± 18.97	76.00 ± 20.66	C=1000, γ0.0001
Mejores+WPT	77.00 ± 9.00	74.00 ± 21.19	80.00 ± 16.33	C=10, γ0.01

Las eficiencias obtenidas por la mayoría de las frases se mantuvieron por encima del 75 %, con el mejor resultado al utilizar los MFCC+WPT alcanzando una tasa de acierto del 79 %, y también se evidencia una disminución en las varianzas; mientras que la sensibilidad y la especificidad, presentan un equilibrio entre en la detección de personas con EP y las sanas.

Además, en la figura 4.9 se presentan las curvas ROC de los resultados obtenidos con los coeficientes cepstrales, los MFCC para las frases 1 y los LFCC para la frase 3, con sus correspondientes AUC dentro del recuadro.

Las curvas ROC para la frase 1 presenta uno de los mejores rendimientos hasta el momento, con una AUC de 0.82 en la fusión, y en el caso de la frase 3 del 0.78.

- **Fusión mediante MKL**

Para terminar, se realizó la fusión de datos mediante MKL, en esta oportunidad no se realiza una fusión individual para cada conjunto de características con las diferentes TFD, sino que se combina la información entregada por todas las TFD al mismo tiempo. Para este procedimiento fue necesario, primero correr el algoritmo con la norma-ℓ1, ya que permite seleccionar los conjuntos de características y *kernels*, que realmente aportan información valiosa al proceso de detección de la EP. Luego de esta selección una vez más se corre el algoritmo, pero esta vez utilizando la norma-ℓ2, con el propósito de obtener los resultados finales de la clasificación.

Se realizaron varios experimentos, utilizando los centroides y dos de los conjuntos de características que se usó en la primera técnica de fusión de información, “todas” las características y las “mejores” características.

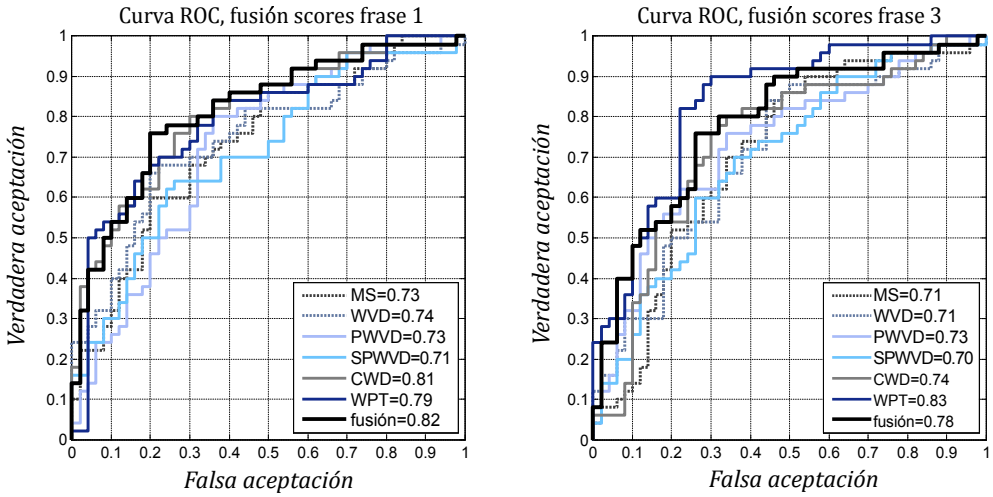

Figura 4.9: Curvas ROC y AUC estimadas utilizando los MFCC para la frase 1 y los LFCC para la frase 3, junto con la fusión.

A continuación, en la tabla 4.37 se presentan los resultados obtenidos con la fusión mediante MKL, con la combinación de la información de 5 TFD al mismo tiempo, es decir, se usaron MS, WVD, PWVD, SPWVD y CWD, como entrada al clasificador y cada TFD tuvo su conjunto de *kernels*. Se resalta en negrilla el mejor desempeño para cada frase en términos de eficiencia.

Tabla 4.37: Resultados de la fusión mediante MKL de MS, WVD, PWVD, SPWVD y CWD.

	Características	% Eficiencia	% Sensibilidad	% Especificidad	C
	Centroides	59.80 ± 6.42	56.75 ± 9.59	62.81 ± 11.05	0.1
Frase 1	Todas	**60.94 ± 6.90**	52.87 ± 10.70	69.02 ± 9.27	0.1
	Mejores	60.91 ± 6.70	53.06 ± 10.44	68.74 ± 9.10	0.1
	Centroides	60.57 ± 7.69	56.25 ± 12.73	64.31 ± 11.03	0.1
Frase 2	Todas	**61.13 ± 7.45**	56.00 ± 12.79	65.35 ± 9.42	1
	Mejores	59.96 ± 8.24	52.04 ± 13.77	67.04 ± 11.07	0.1
	Centroides	59.33 ± 6.49	55.53 ± 11.05	62.89 ± 7.65	0.1
Frase 3	Todas	59.63 ± 6.63	56.54 ± 9.85	62.44 ± 8.51	0.1
	Mejores	**60.22 ± 6.95**	58.39 ± 11.73	61.57 ± 7.92	0.1
	Centroides	**61.85 ± 7.69**	52.37 ± 12.27	70.69 ± 13.34	0.1
Frase 4	Todas	60.41 ± 5.11	48.37 ± 8.50	71.33 ± 12.36	10
	Mejores	60.58 ± 5.33	54.13 ± 8.96	66.55 ± 12.20	1
	Centroides	58.25 ± 4.68	37.05 ± 6.51	77.89 ± 10.35	0.1
Frase 5	Todas	**59.60 ± 5.25**	43.01 ± 6.71	75.03 ± 10.15	1
	Mejores	58.73 ± 5.46	42.90 ± 7.47	73.19 ± 10.81	0.1
	Centroides	57.57 ± 4.94	52.76 ± 4.16	62.49 ± 10.40	10
Frase 6	Todas	**59.90 ± 4.65**	53.95 ± 6.90	65.63 ± 8.92	1
	Mejores	58.84 ± 4.93	51.67 ± 7.01	65.66 ± 8.27	1

Las frases 1, 2, 5 y 6, presentaron los mejores resultados cuando se usaron “todas” las características, aunque las diferencias con los demás conjuntos en cuanto a eficiencia no son muy altas. La frase 4 alcanza la mejor tasa de acierto con 61.85 %.

A continuación, en la tabla 4.38 se presentan los resultados obtenidos con la combinación

de información de las 6 TFD (MS, WVD, PWVD, SPWVD, CWD y WPT) mediante MKL. Se resalta en negrilla el mejor desempeño para cada frase en términos de eficiencia.

Tabla 4.38: Resultados de la fusión mediante MKL utilizando las 6 TFD.

	Características	% Eficiencia	% Sensibilidad	% Especificidad	C
Frase 1	Centroides	**63.11 ± 4.74**	59.08 ± 10.74	66.81 ± 9.78	1
	Todas	62.98 ± 3.99	58.91 ± 7.36	66.67 ± 9.18	1
	Mejores	63.22 ± 4.50	64.49 ± 7.11	67.43 ± 9.51	1
Frase 2	Centroides	**61.47 ± 6.62**	57.63 ± 9.74	64.91 ± 8.73	10
	Todas	60.00 ± 8.50	52.47 ± 14.81	66.61 ± 10.14	0.1
	Mejores	60.00 ± 8.50	52.47 ± 14.81	66.61 ± 10.14	0.1
Frase 3	Centroides	59.69 ± 6.54	55.63 ± 12.44	63.16 ± 7.57	0.1
	Todas	**60.17 ± 10.19**	58.22 ± 10.78	61.78 ± 8.76	0.1
	Mejores	60.46 ± 6.63	58.73 ± 10.98	61.68 ± 8.29	0.1
Frase 4	Centroides	61.28 ± 7.73	51.67 ± 11.58	70.22 ± 14.37	0.1
	Todas	**61.31 ± 5.39**	63.75 ± 12.13	48.89 ± 9.59	0.1
	Mejores	60.12 ± 5.72	54.49 ± 10.73	65.39 ± 13.22	1
Frase 5	Centroides	**58.21 ± 5.42**	39.17 ± 7.55	75.97 ± 10.55	0.1
	Todas	57.95 ± 6.31	48.66 ± 9.41	66.81 ± 11.57	1
	Mejores	58.10 ± 5.58	43.55 ± 7.66	71.65 ± 11.21	0.1
Frase 6	Centroides	59.80 ± 5.12	56.56 ± 8.20	63.21 ± 7.58	10
	Todas	60.88 ± 4.16	56.83 ± 8.77	64.63 ± 7.12	10
	Mejores	**60.36 ± 4.39**	55.57 ± 8.26	64.84 ± 6.66	1

Las frases 1 presenta el mejor resultado con la mejor tasa de acierto de 63.115¿%, cuando se usaron los centroides, aunque las diferencias con los demás conjuntos en cuanto a eficiencia no son muy altas.

Estos resultados son los más bajos obtenidos mediante las estrategias de fusión de información, y se encuentran por debajo de todos los resultados presentados en frases a lo largo de este trabajo, además del costo computacional tan elevado que conlleva. Sin embargo, cabe destacar el hecho que presenta porcentajes de varianzas mucho menor al resto de los experimentos, que en su mayoría se encuentran por encima del 10 %. En relación con la sensibilidad y especificidad, no son muy altos los porcentajes, pero la especificidad es levemente mayor en la mayoría de los casos.

Aunque se puede observar que los resultados en términos de eficiencia de la técnica de fusión mediante MKL, comparativamente están por debajo de la técnica de fusión de *scores*, analizando los pesos asociados a los *kernels* de las diferentes TFD, como se observa en la figura 4.10, los mayores pesos están asociados a los MS seguidos por WPT y CWD, lo cual está relacionado con los resultados obtenidos mediante la fusión a nivel de características, puesto que al promediar la eficiencias obtenidas por cada una de las frases cuando se utilizaron "todas" las características y las "mejores", se obtienen resultados similares a los obtenidos por los pesos de los *kernels*.

Adicionalmente, en la figura 4.11 se presentan las curvas ROC de los resultados obtenidos mediante la fusión MKL, cuando se utilizaron "todas" las características y las "mejores", con sus correspondientes AUC dentro del recuadro.

Las curvas ROC presentan el mejor desempeño cuando se realiza la fusión "todas" las características, mientras que los valores de AUC son muy similares en ambos conjuntos de características. La AUC fue de 0.75 para la frase 6 cuando se utilizan "todas" las características

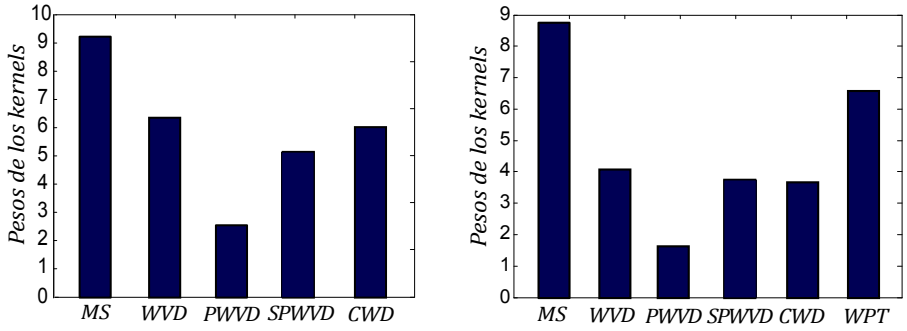

Figura 4.10: Gráficas de la suma de los pesos de los *kernels* obtenidos con el mejor conjunto de características mediante la fusión MKL.

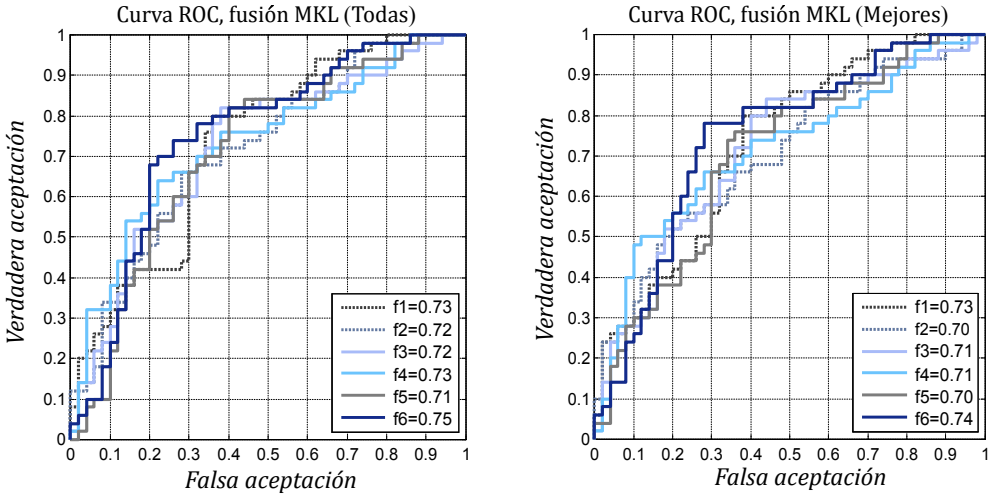

Figura 4.11: Curvas ROC y AUC estimadas a partir de la fusión mediante MKL de cada una de las frases (f).

y de 0.74 cuando se utilizan "todas" las características, también con la frase 6. A pesar de que las tasas de acierto no fueron muy altas, las AUC presentan un desempeño mucho mayor, en todos los casos por encima de 0.70.

Por último, luego de haber determinado las 3 mejores TFD y sus respectivos conjuntos de características, la tabla 4.39 presenta la fusión de información a través de las dos mejores estrategias de fusión, fusión a nivel de características y fusión a nivel de *scores*, pero esta vez utilizando sólo las mejores técnicas TF y características dinámicas, es decir, los MS se caracterizaron a través de los centroides, CWD se caracteriza con MFCC y la WPT mediante energía.

En este caso las dos estrategias de fusión funcionan con un desempeño similar, se puede deducir que el uso de este pequeño conjunto puede lograr tasas de acierto similares que la de los experimentos anteriores. La frase 1 logra el mejor rendimiento con un 78 % mediante

Tabla 4.39: Resultados de la fusión de información a través de la s 3 mejores técnicas TF (MS, CWD and WPT).

Frase	Estrategia de fusión	% Eficiencia	% Sensibilidad	% Especificidad	Parámetros
1	**Características**	**78.0 ± 11.7**	76.0 ± 18.4	80.0 ± 13.3	C=1000, γ=0.01
	Scores	76.0 ± 10.2	72.0 ± 21.5	80.0 ± 13.3	C=1000, γ=0.001
2	**Características**	74.0 ± 18.0	68.0 ± 25.3	80.0 ± 21.1	C=1000, γ=0.01
	Scores	75.0 ± 12.0	64.0 ± 20.6	86.0 ± 13.5	C=10, γ=0.1
3	**Características**	75.0 ± 14.3	78.8 ± 15.9	72.0 ± 25.3	C=100, γ=0.01
	Scores	**77.0 ± 12.7**	80.0 ± 21.1	74.0 ± 18.9	C=1000, γ=0.01
4	**Características**	73.0 ± 11.0	72.0 ± 23.5	74.0 ± 28.4	C=1000, γ=0.01
	Scores	71.0 ± 13.0	72.0 ± 21.5	70.0 ± 23.6	C=1, γ=0.01
5	**Características**	72.0 ± 8.7	64.0 ± 15.8	80.0 ± 18.9	C=1000, γ=0.01
	Scores	73.0 ± 14.9	58.0 ± 19.8	88.0 ± 16.8	C=10, γ=0.1
6	**Características**	77.0 ± 11.8	70.0 ± 23.6	84.0 ± 12.7	C=100, γ=0,01
	Scores	75.0 ± 13.6	72.0 ± 25.3	78.0 ± 22.0	C=1000, γ=0.1
Fusión	***Scores***	**77.0 ± 14.2**	76.0 ± 22.7	78.0 ± 11.4	C=100, γ=0.001

la fusión nivel de características y la frase 3, un 77 % mediante la fusión a nivel de *scores*. La fusión de estas 6 frases alcanzó un 77 % en términos de eficiencia.

Adicionalmente, en la figura 4.12 de la izquierda, se presenta los resultados de fusión a nivel de *scores* de la tabla anterior; mientras que la figura de la derecha presenta la curva ROC de la frase 3, ya que fue la frase con el mejor rendimiento, se muestran los resultados de las tres técnicas TF individualmente y la fusión. Cabe señalar que, de acuerdo con las curvas

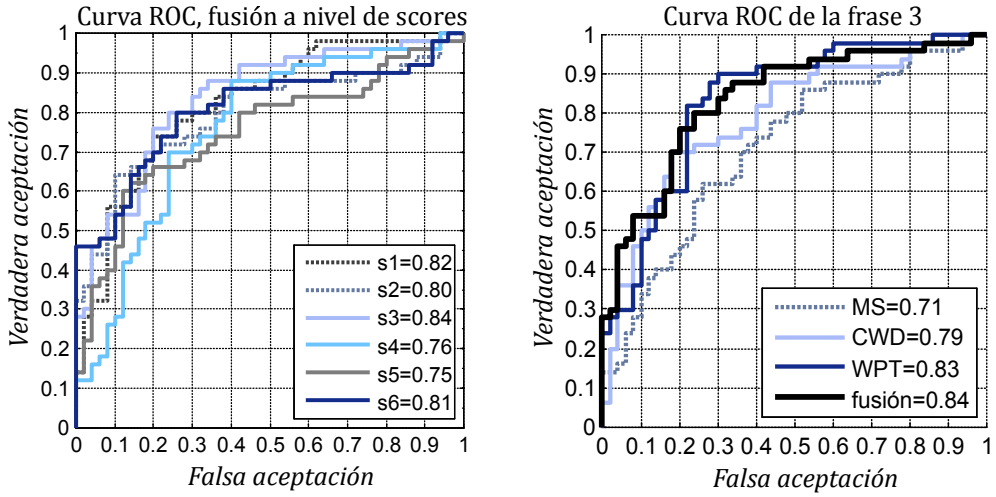

Figura 4.12: Curvas ROC de la fusión a nivel de *scores* y de la frase 3 con las diferentes técnicas TF (MS, CWD y EPT).

ROC, WPT presenta mayor precisión que MS y CWD; y como se esperaba, la mejor AUC se obtuvo con la fusión de las tres técnicas TF, lo que confirma los beneficios de la fusión de la información proporcionada por las diferentes representaciones TF.

4.4. Discusión

En este capítulo se ha presentado un estudio experimental de la metodología propuesta sobre el uso del análisis TF en habla continua, para la detección automática de la EP.

En primer lugar, se realizó un estudio en vocales sostenidas mediante el uso de características clásicas como los MFCC, medidas de ruido y perturbación, en contraste con una de las técnicas TF propuestas en este trabajo, los MS sin embargo, no presentaron tasas de acierto muy altas. Estos resultados eran de esperarse, teniendo en cuenta que en el caso de los pacientes con EP, se ve afectada no solo la fonación sino también los movimientos articulatorios durante el habla, y a pesar de que en las vocales sostenidas interviene la posición de la lengua y otros elementos articulatorios, no es tan rico como en el caso de la producción de la voz continua, ya que todo el tiempo se está modulando y moviendo el aparato fonador. Es por esto que las vocales sostenidas por si solas no son adecuadas para la evaluación de las habilidades de calidad de la voz y comunicación de los pacientes, ya que no incorporan aspectos dinámicos de la voz continua como son las variaciones de periodos de *pitch*, el ritmo, la entonación, etc.

Los resultados obtenidos en el análisis de vocales se tomaron como base para determinar que la SVM fue la estrategia de clasificación más estable, presentando eficiencias muy similares cuando se utilizó técnicas clásicas de extracción de características que cuando se usó todo el espacio de características. Por lo que, pudo determinarse que no era necesario realizar una reducción de dimensión mediante técnicas de extracción, para alcanzar los mejores rendimientos de las características. Además, de que implican un costo computacional bastante elevado durante el proceso de validación no por la técnica misma, sino debido a que se implementó un método tipo *wrapper*, en donde se aumentan los parámetros que deben ser escogidos durante el entrenamiento, y deben de escogerse el número de componentes a seleccionar, mediante la variación diferentes porcentajes de varianza acumulada, que en este caso se encuentran entre 65 % y 95 %, lo que significa que las simulaciones realizadas se incrementan en el número de variaciones realizadas por cada una de las técnicas de clasificación usadas, las cuales requieren a su vez el ajuste de los parámetros propios de cada técnica.

Posteriormente se continuo con el estudio en habla continua para la detección de la EP que permite un análisis más completo sobre la enfermedad al analizar las diferentes dimensiones del habla. Se comenzó realizando un experimento con las características clásicas sobre las frases, aunque los resultados son incluso inferiores que los obtenidos mediante las técnicas TF. Sin embargo, es importante dejar claro que no se realizó ningún tipo de segmentación sonora/sorda (*voiced/unvoiced*), ya que el objetivo de este trabajo es precisamente no utilizar este tipo de métodos, para poder analizar longitudes de señal suficientemente largas, que permitan detectar no sólo cambios en alta frecuencia sino también cambios en baja frecuencia que puedan ser asociados al tremor. Es importante aclarar, que características como los MFCC sólo tienen sentido aplicarlas en períodos de vibración de las cuerdas vocales, en los cuales hay contenido armónico de la señal, por lo que su uso en el escenario de voz continua se debe limitar a segmentos sonoros.

Mediante el uso de diferentes técnicas TF se logra realizar un análisis no estacionario, al

utilizar segmentos de voz más largos que los convencionales (262 ms), sin utilizar técnicas de segmentación. Los primeros experimentos realizados, inicialmente confirman que la SVM es el clasificador que presenta el mejor desempeño para el propósito de este trabajo, y que además la caracterización dinámica alcanza mejores tasas de acierto para todas las TFD propuestas, durante la etapa de clasificación. Se debe tener en cuenta que el contenido espectral de las señales que se están analizando, se encuentra directamente relacionado con el desempeño de la TFD, y en este trabajo las representaciones que mejor rendimiento presentan son la CWD y la WPT, seguidas por los MS.

Los resultados evidencian que el mejor desempeño se obtiene con las características extraídas a partir de CWD, alcanzando las tasas de acierto más altas de todo el sistema mediante la caracterización dinámica de los coeficientes cepstrales, en particular con los MFCC. Esta distribución sobresalió entre las demás no sólo por sus altas tasas de acierto, sino por el comportamiento estable y consistente durante todos los experimentos.

Hay que mencionar, además, que la WPT también obtiene unos de los mejores rendimientos, por encima de técnicas TF como los MS y la SPWVD, utilizando sólo un conjunto de características basado en energía y medidas de información. Aunque, hay que tener en cuenta que para la implementación de todas las trasformaciones TF exceptuando la WPT, fue necesario submuestrear las señales de voz de las frases, debido al costo computacional que implicaba trabajar con la frecuencia original, especialmente en el caso de las TFD de la clase de Cohen, ya que el almacenamiento de la función de ambigüedad requiere una capacidad de memoria bastante grande para su almacenamiento. Mientras que, la WPT si puede trabajar con un costo computacional más razonable utilizando la señal sobre todo el espectro.

Dicho lo anterior, es importante resaltar que entre los experimentos realizados se implementó la WPT no sólo con las señales originales sino también con las remuestreadas, y sus resultados tuvieron una caída drástica, entre 7 y 10 puntos por debajo de los resultados obtenidos, utilizando todo el espectro, motivo por el cual los resultados presentados son con las señales sin remuestrear. Al comparar la WPT en las mismas condiciones que las demás transformaciones, aquellas basadas en la WVD logran mejores resultados, esto lo que indica es que en términos de costo-beneficio está siendo mejor la WPT, pero la información de distribuciones como WVD puede ser mejor, sin embargo, este es el sacrificio que hay que hacer para reducir el costo computacional, lo cual hace que decrezcan un poco los resultados.

En cuanto a las características dinámicas implementadas sobre los espectros de las TFD, se obtuvo la mejor tasa de acierto con los coeficientes cepstrales, tanto LFCC como MFCC, aunque en general los MFCC fueron mejores para la mayoría de representaciones. En segundo lugar, se encuentran los centroides y la FI, que también alcanzaron altas tasas de acierto, pero no fueron las mejores. A diferencia de los marginales, la energía y los EBW, que presentan las tasas más bajas con todas las transformaciones.

Por lo que se refiere a la metodología de fusión de información, se logró mejoras en algunos casos, sobre todo con las estrategias de fusión a nivel de características y *scores*, ya que mediante MKL los resultados no fueron muy alentadores e implican un costo computacional elevado en términos costo-beneficio y un trabajo dispendioso. En el caso de la fusión a nivel de características, los resultados mantuvieron una tendencia al mejor valor alcanzando por

el conjunto de características clasificado de manera individual. Mientras que, en la fusión de *scores*, en particular cuando se combinaron las 6 TFD, se lograron las mejores tasas de acierto en cuanto a fusión de información para la detección automática de la EP.

Es importante resaltar que la detección de la EP es una tarea compleja, que involucra una gran cantidad de aspectos relacionados con el cuerpo entero; por ejemplo la escala UPDRS, es una de las escalas que suelen usar los neurólogos para el diagnóstico y análisis de pacientes con Parkinson, tiene alrededor de 130 parámetros a evaluar y sólo uno de ellos está asociado con la voz, por lo que en este contexto, un 80 % de exactitud a partir de información extraída únicamente de la señal de voz es de hecho un nivel de precisión muy considerable y comparable con test médicos que tienen niveles de significancia muy similares, que se utilizan en la práctica.

Capítulo 5

Conclusiones

La metodología desarrollada en este trabajo explora la caracterización de señales de voz a partir de diferentes representaciones TF, las cuales permiten evaluar los cambios en la riqueza espectral de las señales de voz, sin supuestos de estacionariedad. De esta manera, se establecen las técnicas y características que ofrecen información relevante en la detección automática de pacientes con EP en habla continua, con el fin de proporcionar medidas y estrategias de caracterización que proporcionen información adicional y complementaria a las medidas convencionales de análisis.

Se utilizaron diferentes técnicas TF y para su implementación, se tuvo en cuenta un tamaño de ventana lo suficientemente grande con el fin de poder caracterizar los cambios espectrales introducidos en la señal de voz a causa de la EP, entre ellos el tremor, ya que al ser una perturbación en baja frecuencia no puede ser detectada mediante segmentos cortos. El contenido espectral de las señales que se están analizando, es el que determina el desempeño del sistema, en este caso las representaciones TF que mejor rendimiento presentan son la CWD y la WPT, además de los MS. Lo cual va acorde con lo presentado a lo largo del trabajo, debido a que CWD es una de las mejores distribuciones suavizadas de la WVD y permite representar los cambios dependientes del tiempo, lo cual es ideal para la descripción de los fenómenos no estacionarios, como es el caso de la señal de voz, ya que alcanza un compromiso entre los beneficios de la WD y del espectrograma.

Por lo que se refiere a la WPT, fue diseñada específicamente para tratar con señales variantes en el tiempo y se evidencia a través de los altos resultados obtenidos. Aunque los resultados de la WPT reflejan uno de los mejores comportamientos, se debe tener en cuenta que para las representaciones TF, a diferencia de la WPT, fue necesario submuestrearlas debido al costo computacional que implicaba su implementación. Al comparar la WPT bajo las mismas condiciones que las demás TFD, incluso transformaciones como WVD presentan mejores resultados, por lo que el sacrificio que hay que hacer en la frecuencia de muestreo para reducir el costo computacional hace que decrezcan los resultados. Lo que indica que en términos de costo-beneficio está siendo mejor la WPT, pero la información de las demás transformaciones TF puede ser mejor y que, además, no sólo las componentes de baja frecuencia son importantes, sino que también hay componentes de alta frecuencia que pueden aportar información valiosa para la detección de la EP y se están perdiendo durante

el submuestreo.

Por otro lado, los MS ofrecen una forma compacta para fusionar los fenómenos que se presentan durante la producción del habla, ya que proporcionan información dinámica importante y complementaria para la detección de voces patológicas; lo cual se ve reflejado en los resultados obtenidos mediante la implementación de los centroides y el contenido de energía alrededor de éstos. En los resultados obtenidos con los centroides, es evidente cómo la mayoría presentan las mejores eficiencias cuando se utilizan los centroides calculados con el espectro completo que aquellos calculados con la mejor subbanda, aunque los intervalos de confianza permiten observar que las diferencias entre ambos no son estadísticamente significativas. Así que se debe resaltar, el hecho de que tan sólo con una subbanda se logra concentrar la mayor cantidad de información obtenida de la representación TF, y además puede evidenciarse que la mayoría corresponden a la primera subbanda, lo que podría ser un indicador del contenido de baja frecuencia, asociado a la presencia de temblor en el discurso.

De esta manera, se puede concluir que la mayoría de información se encuentra alrededor de las bandas de baja frecuencia, de donde puede deducirse que el contenido de baja frecuencia refleja tener la mayor capacidad discriminante en los pacientes con EP. Así, el análisis de las TFD mediante bandas de frecuencia se puede considerar como una herramienta robusta y viable para la extracción de características, pese a que no tuvo las eficiencias más altas, puede entregar información valiosa para la detección de la enfermedad.

Adicionalmente, se puede decir que estos resultados van de la mano y están bien representados con las imágenes de los espectros presentadas en el capítulo 2, en las cuales se evidencia como las concentraciones de energía presentes en el espectro de las voces con EP se localizan en frecuencias más bajas que en las sanas, para las técnicas basadas en WVD. Y en el caso de los MS, aunque su representación no se maneja en el mismo dominio, se presenta un comportamiento similar. En cuanto a las demás características dinámicas, en la mayoría de transformaciones TF se obtuvo la mejor tasa de acierto con los coeficientes cepstrales, en su implementación se aplica la transformada homomórfica, la cual es fundamental para concentrar de forma efectiva la mayor información en una cantidad de coeficientes menor al número de filtros utilizados. También se tuvo rendimiento aceptable con la frecuencia instantánea, lo cual concuerda bien con las señales de voz, que presentan más características discriminantes en el dominio de la frecuencia que en el dominio del tiempo.

Por otra parte, los marginales, la energía instantánea y el ancho de banda equivalente, son las variables dinámicas con las cuales se obtiene el rendimiento más bajo, y no tienen buen desempeño para caracterizar las señales bajo estudio. En lo que se refiere a la energía instantánea, es una característica que tiene solamente en cuenta la variabilidad de la señal en el dominio del tiempo y se calcula a través del marginal de tiempo de la TFD, pero la señal debe tener una estructura bien definida en el dominio del tiempo para que se pueda obtener buenos resultados, y en este caso las señales de voz tienen más características discriminantes en el dominio de la frecuencia que en el dominio del tiempo, y por este motivo la energía instantánea no tiene un alto desempeño.

La metodología de fusión de información logró mejoras en algunos casos. En la fusión a nivel de características, los resultados mantuvieron una tendencia al mejor valor alcanzado por los conjuntos de características clasificados de forma individual, sin superar este valor. Y

en la combinación de información a nivel de scores mediante el clasificador SVM, funcionó mejor que las otras estrategias logrando mejoras, en especial cuando se combinaron las 6 TFD propuestas en este trabajo, ya que al incluir la información que aporta la WPT, se logró las tasas de acierto más altas en cuanto a fusión de información. Este es un resultado muy importante, pues confirma que existe información complementaria entre las diferentes representaciones, además de que el contenido de baja frecuencia puede ser complementado con el de alta frecuencia. Es posible concluir que el uso de metodologías basadas en la combinación de clasificadores, proporciona buenos resultados para la detección de la EP. Además, un punto importante a tener en cuenta es que el objetivo fundamental de este trabajo es proporcionar medidas y estrategias de caracterización que proporcionen información adicional y complementaria a las medidas convencionales de análisis, no reemplazar una medida por otra.

Adicionalmente, el área bajo la curva ROC es proporcional al rendimiento del clasificador, lo que se evidencia en los resultados obtenidos, en donde los valores más altos mediante el área bajo la curva corresponden a las características estimadas a partir de CWD y WPT, coincidiendo con las tasas de aciertos de los clasificadores. En cuanto a las curvas ROC obtenidas para las características estimadas con diferentes TFD, permitieron mostrar el comportamiento de la sensibilidad y especificidad del sistema, y cómo varían a medida que se toman diferentes valores para el umbral de decisión. A lo largo del trabajo se observó que, en general tuvieron un desempeño variante, en algunos casos la sensibilidad presentaba porcentajes un poco mayores que los de especificidad, y viceversa; aunque en general la mayoría de resultados tuvieron una tendencia a tener un balance entre la detección de las personas con EP y las sanas. Se debe tener en cuenta que, siendo por supuesto importante la sensibilidad como medida del nivel de detección que tiene el sistema con respecto a la enfermedad, en sistemas reales para el diagnóstico automático de enfermedades es importante tener balanceados tanto la sensibilidad como la especificidad, porque un alto número de falsos positivos impone una carga mayor al sistema de salud, que puede hacer que el uso de la herramienta desarrollada no sea viable en la práctica.

Hay que mencionar, que este trabajo está enmarcado en un proyecto de investigación financiado por Colciencias y la universidad de Antioquia, con participación de diferentes profesionales del área de la salud, como médicos neurólogos y neurocientíficos, con quienes se discuten regularmente los resultados. Es importante resaltar que la detección de la EP es una tarea que involucra una gran cantidad de aspectos relacionados con el cuerpo entero y sólo uno de ellos está asociado con la voz, de esta manera, los porcentajes de acierto obtenidos a partir de información extraída únicamente de la señal de voz son muy considerables. Por lo que, aportar en el diagnóstico de los pacientes con Parkinson, es una tarea compleja desde el punto de vista clínico y que regularmente se puede hacer cuando el estado de enfermedad del paciente es muy avanzado, por lo tanto, cualquier esfuerzo que permita la consecución de un sistema de apoyo al diagnóstico y al seguimiento del tratamiento es bien valorado por los médicos especialistas.

Como trabajo futuro, se plantea encontrar métodos que sean capaces de implementar las metodologías basadas en técnicas TF, especialmente para la distribución CWD, sin la necesidad de remuestrear la voz, o dado el caso, remuestrearla a una frecuencia no muy baja

con respecto a la original. Ya que, si se pudieran procesar las TFD con su frecuencia original se estaría incluyendo la información de alta frecuencia, que como se evidenció con la WPT, también contiene información relevante en el problema de detección de la EP. Esto podría hacerse por medio de la implementación de algoritmos mejorados de las TFD o incluso mediante programación paralela.

Apéndice I

Técnicas de extracción de características

I.a. Análisis de componentes principales (PCA)

PCA es una técnica estadística cuyo propósito es condensar la información de un gran conjunto de variables correlacionadas, en otro conjunto con menos variables "las componentes principales", reteniendo tanto como sea posible la variación presente en el conjunto inicial de datos, es decir, busca las direcciones en el espacio original que tienen mayor variación. El procedimiento utilizado es el propuesto en [176] y [163]. Tomando $\tau_x = \{x_1, \ldots, x_l\}$, como un conjunto de vectores de entrenamiento de entrada $n-$dimensional en el espacio R^n. El conjunto de vectores $\tau_z = \{z_1, \ldots, z_l\}$ es una representación de más baja dimensionalidad que el conjunto de entrada τ_x , con $m-$dimensional en el espacio R^m [176]. Los vectores τ_z son obtenidos mediante la proyección ortonormal lineal.

$$z = W^T x + b, \tag{I.1}$$

Donde la matriz $W[n \times m]$ y el vector $b[m \times 1]$, son parámetros de la proyección. El vector reconstruido $\tau_{\tilde{x}} = \{\tilde{x}_1, \ldots, \tilde{x}_l\}$, es computado por la proyección posterior lineal. Derivado de obtener la inversa de la ecuación I.1, se obtiene la ecuación I.2

$$\tilde{x} = W(z - b), \tag{I.2}$$

La reconstrucción del error cuadrático medio es una función de los parámetros de la proyección lineal de las ecuaciones I.1 y I.2:

$$\varepsilon_{MS}(W, b) = \frac{1}{l} \sum_{i=1}^{l} \| x_i - \tilde{x}_i \|^2, \tag{I.3}$$

El Análisis PCA es la proyección ortonormal lineal de las ecuaciones I.1 y I.2, que permite la mínima reconstrucción del error cuadrático medio I.3, para los datos de entrenamiento τ_x. Los parámetros (W, b), de la proyección lineal son la solución del problema de optimización,

ecuación I.4

$$(W,b) = \underset{W',b'}{argmin}\,\varepsilon_{MS}(W',b'), \tag{I.4}$$

Sujeto a la ecuación I.5:

$$\langle w_i \cdot w_j \rangle = \delta(i,j), \quad \forall i,j \tag{I.5}$$

Donde w_i, $i = 1,\ldots,m$, son vectores columna de la matriz $W = [w_1,\ldots,w_m]$, que contiene los m vectores propios de la muestra en la matriz de covarianza, la cual contiene los mayores valores propios. El vector b es igual a $W^T\mu$, donde μ es la media de la muestra de los datos de entrenamiento [176]. Al final se logra una transformación lineal, que proyecta los datos en las máximas direcciones de varianza, buscando reunir la mayor cantidad de información posible del fenómeno de interés [176], [163].

El mérito de PCA está en que las variables extraídas tienen la mínima correlación a lo largo de los ejes principales. Sin embargo, existen algunos defectos que se encuentran en PCA. Primero, como se menciona en [177], PCA es un método sensitivo a la escala, es decir, las componentes principales pueden ser dominadas por elementos con grandes varianzas. Otro problema que presenta PCA es que las direcciones de máxima varianza no son necesariamente las direcciones de máxima discriminación dado que no utiliza la información de etiquetado de las clases.

I.b. Análisis discriminante lineal (LDA)

LDA es una técnica que busca la máxima relación entre la dispersión entre−clases y la dispersión intra−clases, proyectando las muestras de las clases de un espacio−dimensional dentro de una linea. Al tomar $\tau_(XY) = \{(x_1,y_1),\ldots,(x_l,y_l)\}$, $x_i\ \epsilon\ R^n$, $y\ \epsilon\ Y = \{1,2,\ldots,c\}$, son etiquetados como un conjunto de vectores de entrenamiento. Dentro de la clase $S-W$ y entre la clase $S-B$, las matrices de dispersión se describen mediante las ecuaciones I.6 y I.7:

$$S_W = \sum_{y\varepsilon Y} S_y, \quad S_y = \sum_{i\varepsilon Iy} (x_i - \mu_i)(x_i - \mu_i)^T, \tag{I.6}$$

$$S_B = \sum_{y\varepsilon Y} (\mu_y - \mu)(\mu_y - \mu)^T, \tag{I.7}$$

Donde μ es el vector total de media y μ_y es el vector de medias de clase, $y\ \epsilon\ Y$ están definidos en las ecuaciones I.8 y I.9.

$$\mu = \frac{1}{l}\sum_{i=1}^{l} x_i, \tag{I.8}$$

$$\mu = \frac{1}{|I_y|} \sum_{i \epsilon i y} x_i, y \epsilon Y, \tag{I.9}$$

El éxito de la técnica LDA es entrenar los datos de proyección lineal, $z = W^T x$. Tales que el criterio de separación de clases está definido en la siguiente ecuación, el cual es maximizado.

$$F(W) = \frac{det(\widetilde{S}_B)}{det(\widetilde{S}_W)} = \frac{det(S_B)}{det(S_W)}$$

$\tilde{S}_B$ esta entre las clases, mientras que $\tilde{S}_W$ está contenido en las clases, de la matriz de dispersión de la proyección de los datos, lo cual es similarmente definido en las anteriores técnicas de selección [176], [163].

Bibliografía

[1] J. Godino-Llorente, R. Fraile, N. Sáenz-Lechón, V. Osma-Ruiz, and P. Gómez-Vilda, "Automatic detection of voice impairments from text-dependent running speech," *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 176–182, 2009.

[2] A. L. Merati, Y. D. Heman-Ackah, M. Abaza, K. W. Altman, L. Sulica, and S. Belamowicz, "Common movement disorders affecting the larynx: a report from the neurolaryngology committee of the aao-hns," *Otolaryngology-Head and Neck Surgery*, vol. 133, no. 5, pp. 654–665, 2005.

[3] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *Biomedical Engineering, IEEE Transactions on*, vol. 52, no. 3, pp. 421–430, 2005.

[4] M. Sun, M. Scheuer, and R. Sclabassi, "Decomposition of biomedical signals for enhancement of their time-frequency distributions," *Journal of the Franklin Institute*, vol. 337, no. 4, p. 453–467, 2000.

[5] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *DIgital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[6] O. Hornykiewicz, "Biochemical aspects of parkinson's disease," *Neurology*, vol. 51, pp. S2–S9, 1998.

[7] A. Alm, "Stuttering and the basal ganglia circuits: a critical review of possible relations," *Journal of Communication Disorders*, vol. 37, no. 4, pp. 325–369, 2004.

[8] M. de Rijk, L. Launer, K. Berger, M. Breteler, J. F. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, and et al, "Prevalence of parkinson's disease in europe: A collaborative study of population-based cohorts," *Neurology*, vol. 54, no. 11, pp. 21–23, 2000.

[9] Sanchez J.L. and Buritica O. and Pineda D. and Uribe C.S. and Palacio L.G., "Prevalence of parkinson's disease and parkinsonism in a colombian population using the capture-recapture method," *International Journal of Neuroscience*, vol. 114, no. 2, pp. 175–82, 2004.

[10] L. Ramig, C. Fox, and S. Shimon, "Speech treatment for parkinson's disease," *Expert Review Neurotherapeutics*, vol. 8, no. 2, pp. 297–309, 2008.

[11] J. Rusz, R. Cmejla, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Acoustic analysis of voice and speech characteristics in early untreated parkinson's disease." in *MAVEBA*, 2011, pp. 181–184.

[12] T. Khan, J. Westin, and M. Dougherty, "Classification of speech intelligibility in parkinson's disease," *Biocybernetics and Biomedical Engineering*, vol. 34, no. 1, pp. 35–45, 2014.

[13] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka, "Objectification of dysarthria in parkinson's disease using bayes theorem," *Age (year)*, vol. 61, no. 12.60, pp. 58–08, 2011.

[14] G. Schultz and M. Grant, "Effects of speech therapy and pharmacologic and surgical treatments on voice and speech in parkinson's disease: a review of the literature," *Journal of Communication Disorders*, vol. 33, pp. 59–88, 2000.

[15] K. Perez, L. Ramig, M. Smith, and C. Dromey, "The parkinson larynx: tremor and videostroboscopic findings," *Journal of Voice*, vol. 10, no. 4, pp. 354–361, 1996.

[16] G. Deuschl, J. Raethjen, R. Baron, M. Lindemann, H. Wilms, and P. Krack, "The pathophysiology of parkinsonian tremor: a review," *Journal of neurology*, vol. 247, no. 5, pp. V33–V48, 2000.

[17] N. Sáenz-Lechón, J. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, pp. 120–128, 2006.

[18] J. Proakis, J. Deller, and J. Hansen, "Discrete-time processing of speech signals," *New York, Macrnillan Pub. Co*, 1993.

[19] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[20] J. B. Alonso, F. Díaz-de María, C. M. Travieso, and M. A. Ferrer, "Using nonlinear features for voice disorder detection," in *ISCA Tutorial and Research Workshop (ITRW) on Non-Linear Speech Processing*, 2005.

[21] P. Jain and R. B. Pachori, "Marginal energy density over the low frequency range as a feature for voiced/non-voiced detection in noisy speech signals," *Journal of the Franklin Institute*, vol. 350, no. 4, pp. 698 – 716, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0016003213000069

[22] J. Orozco-Arroyave, F. Hönig, J. Arias-Londoño, J. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of parkinson's disease from words uttered in three different languages," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[23] V. Parsa and D. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of Speech, Language and Hearing Research*, vol. 43, no. 2, pp. 469–485, 2000.

[24] Massachusetts Eye and E. Infirmary, *Voice disorders database*, 1st ed. Lincoln Park. N.J. Kay Elemetrics Corp, 1994.

[25] S. Hadjitodorov and P. Mitev, "A computer system for acoustic analysis of pathological voices and laryngeal diseases screening," *Medical Engineering & Physics*, vol. 24, no. 6, pp. 419–429, 2002.

[26] K. Shama, A. Krishna, and N. Cholayya, "Study of harmonics to noise ratio and critical-band energy spectrum of speech as acoustic indicators of laryngeal and voice pathology," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. ID 85286, p. 9, 2007.

[27] J. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. BlancoVelasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *Journal of Voice*, vol. 24, no. 1, pp. 47–56, 2008.

[28] M. Vasilakis and Y. Stylianou, "Voice pathology detection based eon short-term jitter estimations in running speech," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 153–170, 2009.

[29] R. J. Baken and R. F. Orlikoff, *Clinical measurement of speech and voice*. Cengage Learning, 2000.

[30] R. J. Moran, R. B. Reilly, P. De Chazal, and P. D. Lacy, "Telephony-based voice pathology assessment using automated speech analysis," *Biomedical Engineering, IEEE Transactions on*, vol. 53, no. 3, pp. 468–477, 2006.

[31] E.P.-M Ma and E.M.-L Yiu and, "Suitability of acoustic perturbation measures in analysing periodic and nearly periodic voice signals," *Folia Phoniatrica et Logopaedica*, vol. 57, no. 1, pp. 38–47, 2005.

[32] H. T. Cordeiro, J. M. Fonseca, and C. M. Ribeiro, "Reinke's edema and nodules identification in vowels using spectral features and pitch jitter," *Procedia Technology*, vol. 17, pp. 202–208, 2014.

[33] Y. Heman-Ackah, D. Michael, and G. Goding, "The relationship between cepstral peak prominence and selected parameters of dysphonia," *Journal of Voice*, vol. 16, no. 1, pp. 20–27, 2002.

[34] P. Murphy, "Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals," *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 978–988, 2000.

[35] A. Dibazar, S. Narayanan, and T. Berger, "Feature analysis for automatic detection of pathological speech," in *In Proceedings of the 24th IEEE EMBS Annual International Conference*, 2002, pp. 182–183.

[36] J. Godino-Llorente and P. Gómez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 2, pp. 380–384, 2004.

[37] J. Godino-Llorente, P. Gómez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality asessment system based on gaussian mixture models and short-term cepstral parameters," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 10, pp. 1943–1953, 2006.

[38] R. Fraile, N. Sáenz-Lechón, J. Godino-Llorente, V. Osma-Ruiz, and C. Freudouille, "Automatic detection of laryngeal pahologies in records of sustained vowels by means of mel-frecuency cepstral coeffient parametes and differentiation of patients by sex," *International journal of phoniatrics, speech therapy and communication pathology*, vol. 61, no. 3, 2009.

[39] J. Alonso, J. de Leon, I. Alonso, and M. Ferrer, "Automatic detection of pathologies in the voice by HOS based parameters," *EURASIP Journal on Advanced Signal Processing*, vol. 2001, no. 4, pp. 275–284, 2001.

[40] J. H. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *Biomedical Engineering, IEEE Transactions on*, vol. 45, no. 3, pp. 300–313, 1998.

[41] J. Jiang and Y. Zhang, "Nonlinear dynamic analysis of speech from pathological subjects," *Electronics Letters*, vol. 38, no. 6, pp. 294–295, 2002.

[42] G. Vaziri, F. Almasganj, and R. Behroozmand, "Pathological assessment of patients' speech signals using nonlinear dynamical analysis," *Computers in biology and medicine*, vol. 40, no. 1, pp. 54–63, 2010.

[43] B. S. Aghazadeh, H. Khadivi, and M. Nikkhah-Bahrami, "Nonlinear analysis and classification of vocal disorders," in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. IEEE, 2007, pp. 6199–6202.

[44] J. Orozco-Arroyave, E. Belalcazar-Bolaños, J. Arias-Londoño, J. Vargas-Bonilla, S. Skodda, J. Rusz, F. Honig, K. Daqrouq, and E. Nöth, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and organic diseases," *Biomedical and Health Informatics, IEEE Journal of*, vol. PP, no. 99, pp. 1–1, 2015.

[45] M. A. Kiliç, F. Öğüt, G. Dursun, E. Okur, I. Yildirim, and R. Midilli, "The effects of vowels on voice perturbation measures," *Journal of Voice*, vol. 18, no. 3, pp. 318–324, 2004.

[46] E. J. Wallen and J. H. Hansen, "A screening test for speech pathology assessment using objective quality measures," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2. IEEE, 1996, pp. 776–779.

[47] F. Klingholtz, "Acoustic recognition of voice disorders: A comparative study of running speech versus sustained vowels," *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2218–2224, 1990.

[48] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voicesustained vowels versus continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 2, pp. 327–339, 2001.

[49] J. R. Orozco, J. F. Vargas, J. B. Alonso, M. Ferrer, C. M. Travieso, P. Henríquez *et al.*, "Voice pathology detection in continuous speech using nonlinear dynamics," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*. IEEE, 2012, pp. 1030–1033.

[50] C. M. Travieso, J. B. Alonso, J. R. Orozco-Arroyave, J. Solé-Casals, and E. Gallego-Jutglà, "Automatic detection of laryngeal pathologies in running speech based on the hmm transformation of the nonlinear dynamics," in *Advances in Nonlinear Speech Processing*. Springer, 2013, pp. 136–143.

[51] R. Fraile, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and J. M. Gutiérrez-Arriola, "Characterization of dysphonic voices by means of a filterbank-based spectral analysis: sustained vowels and running speech," *Journal of Voice*, vol. 27, no. 1, pp. 11–23, 2013.

[52] H. Cordeiro, C. Meneses, and J. Fonseca, "Continuous speech classification systems for voice pathologies identification," in *Technological Innovation for Cloud-Based Engineering Systems*. Springer, 2015, pp. 217–224.

[53] L. Ramig, C. Bonitati, J. Lemke, and Y. Horii, "Voice therapy for patients with parkinson's disease: development of an approach and preliminary efficacy data," *Journal of Medical Speech-Language Pathology*, vol. 2, pp. 191–210, 1994.

[54] Y. Yunusova, G. Weismer, J. R. Westbury, and M. J. Lindstrom, "Articulatory movements during vowels in speakers with dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 51, no. 3, pp. 596–611, 2008.

[55] A. Goberman and M. Blomgren, "Fundamental frequency change during offset and onset of voicing in individuals with parkinson disease," *Journal of Voice*, vol. 22, no. 2, pp. 178–191, 2008.

[56] M. Little, P. McSharry, J. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *IEEE transactions on biomedical engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[57] S. Sapirand, L. Raming, J. Spielman, and C. Fox, "Formant centralization ratio (FCR): A proposal for a new acoustic measure of dysarthric speech," *Journal of Speech Language and Hearing Research*, vol. 53, no. 1, pp. 1–20, 2010.

[58] M. Novotny, J. Rusz, R. Cmejla, and E. Ruzicka, "Automatic evaluation of articulatory disorders in parkinson's disease," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 9, pp. 1366–1378, 2014.

[59] B. Walsh and A. Smith, "Basic parameters of articulatory movements and acoustics in individuals with parkinson's disease," *Movement Disorders*, vol. 27, no. 7, pp. 843–850, 2012.

[60] S. Skodda, V. Wenke, and S. Uwe, "Short- and long-term dopaminergic effects on dysarthria in early parkinson's disease," *Journal of Neural Transmission*, vol. 117, no. 2, pp. 197–205, 2010.

[61] S. Skodda, W. Grönheit, and U. Schlegel, "Intonation and speech rate in parkinson's disease: General and dynamic aspects and responsiveness to levodopa admission," *Journal of Voice*, vol. 25, no. 4, pp. e199–e205, 2011.

[62] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues." in *INTERSPEECH*, 2013, pp. 1149–1153.

[63] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[64] T. Bocklet, E. Nöth, G. Stemmer, H. Ruzickova, and J. Rusz, "Detection of persons with parkinson's disease by acoustic, vocal, and prosodic analysis," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, 2011, pp. 478–483.

[65] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of parkinson's speech - acoustic, prosodic and voice related cues," in *Proceedings of the INTERSPEECH*, 2013, pp. 1149–1153.

[66] A. Bayestehtashk, M. Asgari, I. Shafran, and J. McNames, "Fully automated assessment of the severity of parkinson's disease from speech," *Computer speech & language*, vol. 29, no. 1, pp. 172–185, 2015.

[67] A. Papandreou-Suppappola, *Applications in Time-Frequency Signal Processing*. Abingdon: CRC Press, 2002.

[68] L. Cohen, "Time-frequency distributions-a review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, Jul 1989.

[69] E. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev.*, vol. 40, pp. 749–759, Jun 1932.

[70] J. Ville, "Theorie et Applications de la Notion de Signal Analytique," *Cables et Transmission*, vol. 1, pp. 61–74, 1948.

[71] F. Auger, P. Flandrin, P. Goncalves, and O. Lemoine, "Time-frequency toolbox for use with matlab," *Centre National de la Recherche Scientifique and Rice University*, 1996.

[72] F. Hlawatsch and F. Auger, *Time-Frequency Analysis: Concepts and Methods*, 1st ed. ISTE, 2008.

[73] S. Kadambe and G. F. Boudreaux-Bartels, "A comparison of the existence of 'cross terms' in the Wigner distribution and the squared magnitude of the wavelet transform and the short-time Fourier transform," *Signal Processing, IEEE Transactions on*, vol. 40, no. 10, pp. 2498–2517, 1992.

[74] D. Gabor, "Theory of communication. part 1: The analysis of information," *Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of*, vol. 93, no. 26, pp. 429–441, November 1946.

[75] L. Cohen, "Generalized phase-space distribution functions," *Journal of Mathematical Physics*, vol. 7, pp. 781–786, 1966.

[76] Claasen, T. A. C. M., and W. F. G. Mecklenbrauker, "The Wigner distribution: A tool for time-frequency signal analysis. Part I–Continuous-time signals," *Philips Journal of Research*, vol. 35, no. 3, pp. 217–250, 1980.

[77] Claasen, T. A. C. M., and W. F. G. Mecklenbrauker , "The Wigner distribution: A tool for time-frequency signal analysis. Part II–Discrete-time signals," *Philips Journal of Research*, vol. 35, no. 4/5, pp. 276–300, 1980.

[78] M. Born, W. Heisenberg, and P. Jordan, "Zur quantenmechanik. ii." *Zeitschrift für Physik*, vol. 35, no. 8-9, pp. 557–615, 1926.

[79] Y. Zhao, L. E. Atlas, and R. J. Marks, "The use of cone-shaped kernels for generalized time-frequency representations of nonstationary signals," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 7, pp. 1084–1091, 1990.

[80] A. Rihaczek, "Signal energy distribution in time and frequency," *IEEE Trans. Inf. Theor.*, vol. 14, no. 3, pp. 369–374, Sep. 2006.

[81] H. Margenau and R. Hill, "Correlation between measurements in quantum theory," *Progress of Theoretical Physics*, vol. 26, no. 5, Nov 1961.

[82] H.-I. Choi and W. J. Williams, "Improved time-frequency representation of multicomponent signals using exponential kernels," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 6, pp. 862–871, 1989.

[83] B. Boashash, *Time-frequency signal analysis–methods and applications*. Longman Cheshire, 1992.

[84] S. Kadambe, R. Murray, and G. Boudreaux-Bartels, "The dyadic wavelet transform based qrs detector [ECG analysis]," in *Signals, Systems and Computers, 1992. 1992 Conference Record of The Twenty-Sixth Asilomar Conference on*, Oct 1992, pp. 130–134 vol.1.

[85] B. Dugnol, C. Fernández, G. Galiano, and J. Velasco, "Implementation of a diffusive differential reassignment method for signal enhancement: An application to wolf population counting," *Applied Mathematics and Computation*, vol. 193, no. 2, pp. 374 – 384, 2007.

[86] D. Mandic, N. Rehman, Z. Wu, and N. Huang, "Empirical mode decomposition-based time-frequency analysis of multivariate signals: The power of adaptive data analysis," *Signal Processing Magazine, IEEE*, vol. 30, no. 6, pp. 74–86, Nov 2013.

[87] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation ofspeech," in *In Proceedings ICASSP, 1997*, 1997, pp. 1647–1650.

[88] A. Belouchrani, M. Amin, N. Thirion-Moreau, and Y. Zhang, "Source separation and localization using time-frequency distributions: An overview," *Signal Processing Magazine, IEEE*, vol. 30, no. 6, pp. 97–107, Nov 2013.

[89] T. S. Lee, "Image representation using 2d gabor wavelets," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 10, pp. 959–971, 1996.

[90] A. Gramfort, D. Strohmeier, J. Haueisen, M. Hamalainen, and M. Kowalski, "Functional brain imaging with m/eeg using structured sparsity in time-frequency dictionaries," in *Information Processing in Medical Imaging*. Springer, 2011, pp. 600–611.

[91] F. Auger, P. Flandrin, Y.-T. Lin, S. McLaughlin, S. Meignen, T. Oberlin, and H.-T. Wu, "Time-frequency reassignment and synchrosqueezing: An overview," *Signal Processing Magazine, IEEE*, vol. 30, no. 6, pp. 32–41, Nov 2013.

[92] C. Li and C. Zheng, "QRS detection by wavelet transform," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1993.

[93] J. Jessop and J. Jones, "Evaluation of the actions of general anaesthetics in the human brain," *General Pharmacology: The Vascular System*, vol. 23, no. 6, pp. 927–935, 1992.

[94] J. Bulgrin, B. Rubal, C. Thompson, and J. Moody, "Comparison of short-time fourier, wavelet and time-domain analyses of intracardiac sounds." *Biomedical sciences instrumentation*, vol. 29, pp. 465–472, 1992.

[95] Z.-Y. Lin and Z. Chen, "Time-frequency representation of the electrogastrogram-application of the exponential distribution," *Biomedical Engineering, IEEE Transactions on*, vol. 41, no. 3, pp. 267–275, 1994.

[96] J. J. Eggermont and G. M. Smith, "Characterizing auditory neurons using the wigner and rihacek distributions: A comparison," *The Journal of the Acoustical Society of America*, vol. 87, no. 1, pp. 246–259, 1990.

[97] H. P. Zaveri, W. J. Williams, L. D. Iasemidis, and J. C. Sackellares, "Time-frequency representation of electrocorticograms in temporal lobe epilepsy," vol. 39, no. 5. IEEE, 1992, pp. 502–509.

[98] S. J. Schiff and J. G. Milton, "Wavelet transforms for electroencephalographic spike and seizure detection," pp. 50–56, 1993.

[99] N. V. Thakor, G. Xin-Rong, S. Yi-Chun, and D. F. Hanley, "Multiresolution wavelet analysis of evoked potentials," *Biomedical Engineering, IEEE Transactions on*, vol. 40, no. 11, pp. 1085–1094, 1993.

[100] A. A. Petrosian, D. V. Prokhorov, W. Lajara-Nanson, and R. B. Schiffer, "Recurrent neural network-based approach for early recognition of Alzheimer's disease in EEG," *Clinical Neurophysiology*, vol. 112, pp. 1378–1387, 2001.

[101] P. E. O'Suilleabhain and J. Y. Matsumoto, "Time-frequency analysis of tremors." *Brain*, vol. 121, no. 11, pp. 2127–2134, 1998.

[102] B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms," *Signal Processing, IEEE Transactions on*, vol. 47, no. 6, pp. 1622–1635, 1999.

[103] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[104] X. He and M. S. Scordilis, "Psychoacoustic music analysis based on the discrete wavelet packet transform," *Journal of Electrical and Computer Engineering*, vol. 2008, 2008.

[105] K. Abid and K. Ouni, "An improved psycho-acoustic model for mpeg 1 using a morlet cambridge wavelet," in *Signals, Circuits and Systems (SCS), 2009 3rd International Conference on*. IEEE, 2009, pp. 1–4.

[106] X. He and M. S. Scordilis, "An enhanced psychoacoustic model based on the discrete wavelet packet transform," *Journal of the Franklin Institute*, vol. 343, no. 7, pp. 738–755, 2006.

[107] A. Karmakar, A. Kumar, and R. Patney, "A multiresolution model of auditory excitation pattern and its application to objective evaluation of perceived speech quality," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1912–1923, 2006.

[108] D. Surangsrirat, A. Intarapanich, C. Thanawattano, R. Bhidayasiri, S. Petchrutchatachart, and C. Anan, "Tremor assessment using spiral analysis in time-frequency domain," in *Southeastcon, 2013 Proceedings of IEEE*, April 2013, pp. 1–6.

[109] B. Boashash, G. Azemi, and J. M. O'Toole, "Time-frequency processing of nonstationary signals: Advanced tfd design to aid diagnosis with highlights from medical applications," *Signal Processing Magazine, IEEE*, vol. 30, no. 6, pp. 108–119, 2013.

[110] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, I. M. Moroz *et al.*, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 1, p. 23, 2007.

[111] J. Nayak, P. S. Bhat, R. Acharya, and U. Aithal, "Classification and analysis of speech abnormalities," *ITBM-RBM*, vol. 26, no. 5, pp. 319–327, 2005.

[112] R. Silva-Fonseca, R. Capobiano-Guido, P. Rogéiro-Scalassara, C. Dias-Maciela, and J. Carlos-Pereira, "Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders," *Computers in Biology and Medicine*, vol. 37, no. 4, pp. 571–578, 2007.

[113] M. K. Arjmandi and M. Pooyan, "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine," *Biomedical Signal Processing and Control*, vol. 7, no. 1, pp. 3–19, 2012.

[114] O. Geman and C. Zamfir, "Using wavelet for early detection of pathological tremor," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, Aug 2012, pp. 1723–1727.

[115] A. Schuck, L. Guimaraes, and J. Wisbeck, "Dysphonic voice classification using wavelet packet transform and artificial neural network," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 3, 2003, pp. 2958–2961 Vol.3.

[116] L. Cnockaert, F. Grenez, J. Schoentgen, C. Ozsancak, and P. Auzou, "Analysis of vocal tremor by means of a complex wavelet transform," in *Proc. International Conference on Voice Physiology and Biomechanics*, 2004, pp. 27–29.

[117] G. Schlotthauer, M. E. Torres, and H. L. Rufiner, "Pathological voice analysis and classification based on empirical mode decomposition," in *Development of multimodal interfaces: active listening and synchrony*. Springer, 2010, pp. 364–381.

[118] M. Kaleem, B. Ghoraani, A. Guergachi, and S. Krishnan, "Pathological speech signal analysis and classification using empirical mode decomposition," *Medical & biological engineering & computing*, vol. 51, no. 7, pp. 811–821, 2013.

[119] L. E. Atlas and S. A. Shamma, "Joint Acoustic and Modulation Frequency," *Eurasip Journal on Advances in Signal Processing*, vol. 2003, pp. 668–675, 2003.

[120] S. Greenberg and B. E. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 1997.

[121] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *Proceedings of ICASSP*, vol. 1, 2006, p. 665–668.

[122] M. Markaki and Y. Stylianou, "Voice pathology detection and classification based on modulation spectra," in *In Proceedings of 3th Advance Voice Function Assessment International Workshop, AVFA, Madrid, Spain*, 2009.

[123] M. E. Markaki and Y. Stylianou, "Normalized modulation spectral features for cross-database voice pathology detection." in *INTERSPEECH*, 2009, pp. 935–938.

[124] J. Arias-Londoño, J. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocology*, vol. 36, no. 2, pp. 60–69, 2011.

[125] M. Markaki, Y. Stylianou, J. Arias-Londoño, and J. Godino-Llorente, "Dysphonia detection based on modulation spectral features and cepstral coefficients," in *In Proceedings of ICASSP, Dallas, USA*, 2010.

[126] T. Villa-Canas, J. Orozco-Arroyave, J. Vargas-Bonilla, and J. Arias-Londoño, "Modulation spectra for automatic detection of parkinson's disease," in *XIX Symposium on Image, Signal Processing and Artificial Vision (STSIVA), 2014*. IEEE, 2014, pp. 1–5.

[127] T. Villa-Cañas, J. Arias-Londoño, J. Orozco-Arroyave, J. Vargas-Bonilla, and E. Nöth, "Low-frequency components analysis in running speech for the automatic detection of parkinson's disease," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[128] E. Song, J. Ryu, and H.-G. Kang, "Speech enhancement for pathological voice using time-frequency trajectory excitation modeling," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, Oct 2013, pp. 1–4.

[129] S. S. Mahmoud, Z. M. Hussain, I. Cosic, and Q. Fang, "Time-frequency analysis of normal and abnormal biological signals," *Biomedical Signal Processing and Control*, vol. 1, no. 1, pp. 33 – 43, 2006.

[130] K. Gröchenig, *Foundations of Time-Frequency Analysis*, ser. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 2001.

[131] E. Sejdić, I. Djurović, and J. Jiang, "Time–frequency feature representation using energy concentration: An overview of recent advances," *Digit. Signal Process.*, vol. 19, no. 1, pp. 153–183, Jan. 2009.

[132] B. Boashash, "Time frequency signal analysis and processing: Method and applications," 2003.

[133] L. Cohen, *Time-frequency analysis*. Prentice Hall PTR Englewood Cliffs, NJ:, 1995, vol. 1406.

[134] F. Hlawatsch and G. F. Boudreaux-Bartels, "Linear and quadratic time-frequency signal representations," *IEEE signal processing magazine*, vol. 9, no. 2, pp. 21–67, 1992.

[135] A. D. Poularikas, *Transforms and applications handbook*. CRC press, 2010.

[136] L. Debnath, *Wavelets and signal processing*. Springer Science & Business Media, 2012.

[137] M. Markaki and Y. Stylianou, "Voice pathology detection and discrimination based on modulation spectral features," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 7, pp. 1938–1948, 2011.

[138] P. Flandrin, "Some features of time-frequency representations of multicomponent signals," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, vol. 9. IEEE, 1984, pp. 266–269.

[139] F. Hlawatsch, T. G. Manickam, R. L. Urbanke, and W. Jones, "Smoothed pseudo-wigner distribution, choi-williams distribution, and cone-kernel representation: Ambiguity-domain analysis and experimental comparison," *Signal Processing*, vol. 43, no. 2, pp. 149–168, 1995.

[140] P. S. Addison, *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC Press, 2010.

[141] A. N. Akansu and P. R. Haddad, *Multiresolution signal decomposition: transforms, subbands, and wavelets*. Academic Press, 2000.

[142] S. Mallat, *A wavelet tour of signal processing*. Academic press, 1999.

[143] W. Alkhaldi, W. Fakhr, and N. Hamdy, "Multi-band based recognition of spoken arabic numerals using wavelet transform," in *Radio Science Conference, 2002.(NRSC 2002). Proceedings of the Nineteenth National*. IEEE, 2002, pp. 224–229.

[144] M. Markaki and Y. Stylianou, "Using modulation spectra for voice pathology detection and classification," in *In Proceedings IEEE EMBC'09, Minneapolis, Minnesota, U.S.A.*, 2009, pp. 2514–2517.

[145] T. Kinnuen, K. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification," in *In Proceedings Odyssey: Speaker Lang. Recognition Workshop*, 2008.

[146] L. Atlas, L. Owsley, J. McLaughlin, and G. Bernard, "Automatic feature-finding for time-frequency distributions," in *Time-Frequency and Time-Scale Analysis, 1996., Proceedings of the IEEE-SP International Symposium on*. IEEE, 1996, pp. 333–336.

[147] L. Atlas and J. Fang, "Quadratic detectors for general nonlinear analysis of speech," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 2. IEEE, 1992, pp. 9–12.

[148] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," in *Computers in Cardiology 1997*. IEEE, 1997, pp. 105–108.

[149] ——, "A boundary modification method for heart sound segmentation algorithm," in *Computers in Cardiology 1998*, Sep 1998, pp. 593–595.

[150] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelogram and self-organizing probabilistic model," in *Computers in Cardiology, 2005*. IEEE, 2005, pp. 957–960.

[151] C. N. Gupta, R. Palaniappan, S. Rajan, S. Swaminathan, and S. Krishnan, "Segmentation and classification of heart sounds," in *Electrical and Computer Engineering, 2005. Canadian Conference on*. IEEE, 2005, pp. 1674–1677.

[152] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, Apr 1990, pp. 381–384 vol.1.

[153] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 520–538, Apr 1992.

[154] Z. M. Hussain and B. Boashash, "Multi-component if estimation," in *Statistical Signal and Array Processing, 2000. Proceedings of the Tenth IEEE Workshop on*. IEEE, 2000, pp. 559–563.

[155] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signal. ii. algorithms and applications," *Proceedings of the IEEE*, vol. 80, no. 4, pp. 540–568, 1992.

[156] L. Rankine, M. Mesbah, and B. Boashash, "If estimation for multicomponent signals using image processing techniques in the time–frequency domain," *Signal Processing*, vol. 87, no. 6, pp. 1234–1250, 2007.

[157] S. Kikkawa and H. Yoshida, "On unification of equivalent bandwidths of a random process," *Signal Processing Letters, IEEE*, vol. 11, no. 8, pp. 670–673, Aug 2004.

[158] H. Yoshida, S. Kikkawa *et al.*, "Information theoretic equivalent bandwidths of random processes and their applications," *Methods Inf Med*, vol. 46, no. 2, pp. 110–116, 2007.

[159] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, May 1998, pp. 617–620 vol.2.

[160] K. Paliwal, "Spectral subband centroids as features for speech recognition," in *In Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 124–131.

[161] B. Gajic and K. Paliwal, "Robust feature extraction using subband spectral centroid histograms," in *In Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, 2001, pp. 85–88.

[162] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, 2001. [Online]. Available: http://books.google.com.co/books?id=reZQAAAAMAAJ

[163] R. Duda, P. Hart, and D. Stork, "Pattern classification," *John Wiley & Sons, 2nd edition*, 2001.

[164] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactios on Speech and Audio Processing*, vol. 3, no. 1, 1995.

[165] C. Bishop, "Pattern recognition and machine learning," *Springer, New York, NY, USA*, 1997.

[166] A. J. Smola and B. Schölkopf, *Learning with kernels*. Citeseer, 1998.

[167] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[168] M. He, S. J. Horng, P. Fan, R. S. Run, R. J. Chen, J. L. Lai, M. K. Khan, and K. O. Sentosa, "Performance evaluation of score level fusion in multimodal biometric systems," *Pattern Recognition*, vol. 43, no. 5, pp. 1789–1800, 2010.

[169] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *The Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[170] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *The Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.

[171] F. R. Bach, G. R. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 6.

[172] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *The Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[173] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. d. Bona, A. Binder, C. Gehl, and V. Franc, "The shogun machine learning toolbox," *The Journal of Machine Learning Research*, vol. 11, pp. 1799–1802, 2010.

[174] J. Orozco-Arroyave, J. Arias-Londoño, J. Vargas-Bonilla, M. González-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from parkinson's disease," in *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, 2014.

[175] C. G. Goetz, W. Poewe, O. Rascol, C. Sampaio, G. T. Stebbins, C. Counsell, N. Giladi, R. G. Holloway, C. G. Moore, G. K. Wenning *et al.*, "Movement disorder society task force report on the hoehn and yahr staging scale: status and recommendations the movement disorder society task force on rating scales for parkinson's disease," *Movement disorders*, vol. 19, no. 9, pp. 1020–1028, 2004.

[176] C. Bishop, "Neural networks for pattern recognition," *Clarendon Press Oxford, Great Britain, 3th edition*, 1997.

[177] D. Peña, *Análisis de datos multivariantes*. Mc Graw Hill, 2002s.