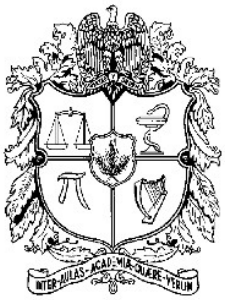
Selección de características en el análisis acústico de voces

Jesús Francisco Vargas Bonilla



Universidad Nacional de Colombia Sede Manizales
Facultad de Ingeniería y Arquitectura
Departamento de Electricidad, Electrónica y Computación
Grupo Control y Procesamiento Digital de Señales
Manizales
2003

Selección de características en el análisis acústico de voces

Jesús Francisco Vargas Bonilla

Trabajo de Grado como requerimiento parcial para optar al título de
Magister en Automatización Industrial

Director
César Germán Castellanos Domínguez

Universidad Nacional de Colombia Sede Manizales
Facultad de Ingeniería y Arquitectura
Departamento de Electricidad, Electrónica y Computación
Grupo Control y Procesamiento Digital de Señales
Manizales
2003

Feature Selection in Acoustic Analysis of Voice Signals

Jesús Francisco Vargas Bonilla

In partial fulfillment of the requirements
for the Degree of Master of Science

Tutor
Profesor César Germán Castellanos Domínguez

Universidad Nacional de Colombia Sede Manizales
Facultad de Ingeniería y Arquitectura
Departamento de Electricidad, Electrónica y Computación
Grupo Control y Procesamiento Digital de Señales
Manizales
2003

Este trabajo se realiza en el marco del proyecto *Identificación Automatizada de Voces Eufónicas y disfuncionales en la población adulta de la Ciudad de Manizales* financiado por el **DIMA** de la UN-Manizales, y apoyado por **COLCIENCIAS**, Código 2020100108.
Jesús Francisco Vargas Bonilla, desarrolló actividades académicas dentro de este proyecto, a través del programa JOVENES INVESTIGADORES de **COLCIENCIAS**.

Dedicatoria

A mis Padres, por todo su amor,

A mi Familia, por su apoyo incondicional,

A mis amigos, por los momentos que me alejan de la locura,

A Nico, para que tenga una opción más cuando piense que quiere ser.

F. Vargas

Agradecimientos

Teniendo en cuenta que este trabajo representa el esfuerzo e interés de un grupo de personas que de manera decidida apuntaron sus esfuerzos a lo que hoy en día se plasma en este documento, quisiera expresarles mi mas sentido agradecimiento.

Al profesor German Castellanos, director de este trabajo, por haber creído en mi y haberme encomendado la tarea de ahondar en un campo que día a día logra despertar en mi nuevas inquietudes, lo que permite pensar en un futuro de trabajo arduo y decidido para seguir logrando cosas. A la Doctora Libia Maria Botero Tobon, quien desde hace ya varios años, ha decido compartir sus esfuerzos y valiosos conocimientos con nosotros; llena de paciencia, ha sabido comprender los errores y desatinos que hemos tenido, pero que al igual que nosotros, esta convencida de que aquí, en nuestra tierra COLOMBIA, existe la posibilidad de hacer investigación, y mas importante aún, investigación aplicada socialmente, sin importar que esta se desarrolle en el mundo de la ingeniería.

A los integrantes del Grupo de Control y Procesamiento Digital de Señales, por su compañía, apoyo, soporte, ayuda, colaboración, y todo aquello que involucra el trabajo en grupo. Un agradecimiento especial, a los ingenieros Fabian Ojeda, Mauricio Orozco y Ricardo Henao, por sus

valiosos aportes, criticas, discusiones y precisiones. A Omar D. Castrillón, por toda su colaboración. A COLCIENCIAS, por el apoyo recibido a través del programa *Jovenes Investigadores*.

A Natalia G., María Andrea L., Ivonne M., Nubia M., Daniel O., Carlos O., William C., Germán R., grandes personas que me permitieron alejarme de la locura, y que me brindaron momentos bellos que siempre recordaré. Espero que pasados algunos años, algunos de ellos consulten este documento y recuerden que donde quiera que esté, siempre tendré una sonrisa en el rostro cada vez que me pregunte por ellos. Gracias, *AMIGOS*.

A todos ellos y aquellos que se me olvidan, **MUCHAS GRACIAS**, por aguantarse a este loco, enamorado de la vida y de las señales.

Resumen

Se presenta una metodología de selección de características basada en el análisis de independencia estadística y en el análisis de componentes principales (PCA). Se emplean pruebas de hipótesis, y se analizan las variantes lineal y no lineal de PCA; para el caso no lineal se utiliza un kernel RBF. La metodología está orientada al análisis acústico de voces con el fin de determinar la presencia de algún grado de disfonía en registros de señales de voz tomados de personas adultas de la población urbana de la ciudad de Manizales, Colombia. Para la prueba de la metodología, se comparan los porcentajes de clasificación obtenidos con el conjunto completo de características y para el conjunto reducido. Se utiliza una maquina de soporte vectorial (SVM) como clasificador.

Abstract

A feature selection methodology based on statistical independence and principal components analysis (PCA) is presented. Hypothesis tests are employed, and linear and no linear PCA are analyzed; in no linear analysis, Kernel RBF was used. Methodology is oriented towards acoustic analysis of voices in orden to detect any disphony degree in voice signals records from adult people of urban population of Manizales, Colombia. To prove the methodology, classification percentages obtained for complete and reduced feature sets are compared. A Support Vector Machine (SVM) is used as classifier.

Índice General

Índice de Figuras

Índice de Tablas

Introducción

El presente trabajo, recopila las actividades realizadas entorno a la Tesis de Maestría en Automatización Industrial realizada en el área de procesamiento de señales de voz, específicamente en selección de características. Se desarrolla una metodología de selección orientada al análisis acústico de voces, y tiene como bases el análisis de independencia estadística y el análisis de componentes principales, como herramientas para decidir qué características permiten determinar de mejor manera, la presencia de algún grado de disfonía a partir de una muestra población de los adultos en la ciudad de Manizales, Colombia.

Los análisis fueron aplicados sobre diferentes conjuntos de características, que fueron conformados teniendo en cuenta características de dos naturalezas: *características de origen acústico*, las cuales poseen una interpretación física, y *características de representación*, las cuales corresponden a la utilización de la transformada Ondita, y no tienen una interpretación física relacionada.

La metodología contempla la utilización de pruebas de hipótesis, correlación de Spearman, y análisis de componentes principales, para determinar si determinada características está en capacidad de aportar información necesaria para la tarea de clasificación. En cuanto al análisis de

componentes principales, se analizan las variantes de tipo lineal y no lineal, con el fin de determinar la mejor representación de los datos.

En el *primer capítulo* se expone la fisiología del aparato fonador, aquí se brindan los conceptos teóricos a cerca del proceso de producción vocal por parte del aparato fonador humano, complementándolos con las clasificaciones fisiológicas y linguisticas de los fonemas que conforman el idioma castellano. Se presenta el concepto de patología vocal, se resumen sus efectos, y se describe en especial la disfonía. De igual manera se describe el procedimiento de análisis acústico de voces y las utilidades de este en la valoración de las señales de voz.

En el *capítulo dos*, se presentan los conceptos del procesamiento digital de señales que sirvieron de soporte para el desarrollo del presente trabajo, se revisan las etapas de adquisición y adecuamiento de la señal, su preprocesamiento, y finalmente se expone la representación de la señal partir de conceptos matemáticos como la transformada de Fourier y la transformada Ondita.

El *capítulo tres* trata la estimación de las características de la señal de voz, presenta la base teórica de su cálculo y analiza los algoritmos implementados para el presente trabajo.El capítulo cuatro, presenta los procedimientos utilizados en la metodología de selección de características propuesta. Se describen los procedimientos de independencia estadística y análisis de componentes principales.

Finalmente, el capítulo cinco describe el marco experimental realizado, se presentan las prue-

bas realizadas y sus correspondientes resultados. Se presentan tablas comparativas, que permiten analizar los resultados obtenidos para las diferentes pruebas realizadas.

Capítulo 1

Fisiología del aparato Fonador

1.1 Aparato de voz

En general, se considera la voz como el sonido producido mediante la interacción debida a la vibración de las cuerdas vocales y la articulación ejecutada en las cavidades superiores a la laringe; la vibración de las cuerdas se debe gracias a la intervención del aire espirado, que al pasar a través del tracto vocal amplifica el sonido y, así el producto es enriquecido en timbre y sonoridad, generando un factor específico de identificación personal. Las cualidades sensoriales del sonido están dadas por su frecuencia e intensidad.

1.1.1 Mecanismo de producción de voz

La producción de la voz involucra tres procesos básicos: [1]

- *Fuente*, encargada de la generación de los sonidos (pulmones, cuerdas vocales).
- *Articulación (modulador)*, le da forma a los sonidos que se están generando y define la capacidad de entonación. Comprende el tracto vocal(principalmente, las cavidades oral y

nasal).

– *Emisión,* corresponde a la parte final de las cavidades oral y nasal, por donde se expulsa el sonido en forma de ondas de presión sonora.

El órgano vocal humano está compuesto por los pulmones, tráquea, laringe, faringe y las cavidades oral y nasal. La parte superior que comienza con la laringe, conocida como el *tracto vocal*, es modificable en varias formas mediante el movimiento de la mandíbula, lengua y labios. La cavidad nasal está separada de la faringe y de la cavidad oral por la elevación del velo o paladar blando (Figura 1.1). Cuando los músculos abdominales elevan el diafragma, el aire es expulsado desde los pulmones hacia la laringe, pasando a través de la tráquea y el espacio entre las cuerdas vocales (*la glotis*). Esta última separa las cuerdas vocales y se mantiene abierta durante la respiración, pero en el momento de producir sonidos se va estrechando, de manera intermitente, cerrando el paso del aire. Este movimiento de apertura y cierre de la glotis (acercamiento, alejamiento y tensión de las cuerdas vocales) está asociado con la *entonación* que se da al habla. La velocidad con la que las cuerdas vocales vibran se conoce como la *frecuencia fundamental* [1]. Tras superar la glotis, el aire se acerca al tracto vocal, que va variando su forma de manera rápida, en función de los sonidos que se desee producir. Los articuladores de la cavidad oral (lengua, labios, mandíbulas, velo del paladar) actúan como elementos variables de resonancia, los cuales amplifican o atenúan selectivamente los componentes espectrales de la onda de presión que hasta aquí haya llegado. Cada una de las resonancias tiene su energía concentrada alrededor de cierta frecuencia, conocidas como *formantes*.

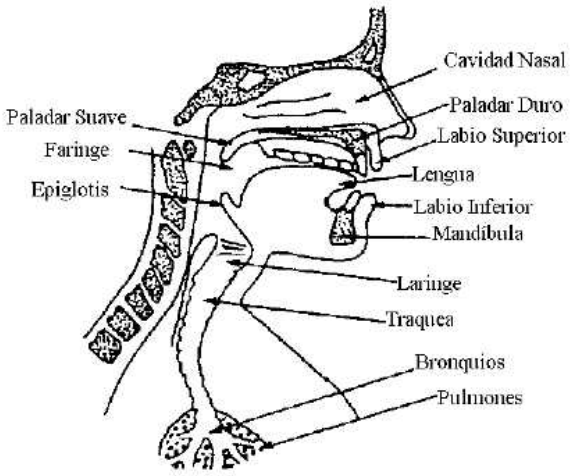


Figura 1.1: Esquema del Aparato Fonador Humano

1.1.2 Clasificación fisiológica

Se determina la voz *normal* como la emisión coordinada, armónica con todas las cualidades y buen manejo de los niveles anatómico-fisiológicos que participan en la producción vocal, entre ellos el respiratorio, de resonancia, auditivo, emisor, hormonal y de comando [2]. La voz de cualquier persona está condicionada por sus características anatomo-fisiológicas particulares. En general, la estructura laríngea y las características de la voz reflejan la edad, el género y el estado de salud y anímico. De otra parte, la voz puede ser del tipo *hablada*, cuando se pone en marcha el uso habitual de la voz conversacional y, si bien es preciso un funcionamiento adecuado, tanto de la respiración como de la fonación, no requiere la especial adaptación y acomodación de los órganos y músculos que participan, ya sea en la emisión de la voz, en el apoyo respiratorio y en las estructuras resonantes supralaríngeas. Otro tipo de voz es la *proyectada*, en la cual a diferencia del caso anterior, se da una sutil acomodación y conjunción de un complejo juego posturo-muscular

y propioceptivo que hace posible la proyección de la voz en situaciones como presentación en público, interpelación a alguien a cierta distancia, intervención en situaciones con niveles fuertes de ruido de fondo, etc. Sin embargo, para una mayor precisión en el estudio de la voz, ésta puede ser clasificada de acuerdo al género, edad y niveles de empleo vocal [3], así:

Género

Las características físicas del tejido que compone las cuerdas vocales difieren en cada sexo, en el caso de la mujer, presenta medidas que oscilan entre 3.6 cm de altura, 4.3 cm de anchura y con diámetro antero-posterior estimado en los 2.6 cm; la longitud de las cuerdas vocales se sitúa entre los 1.5 a 2 cm. Respecto al hombre, su laringe es de mayor tamaño con valores medios de altura y anchura de 4.9 cm, diámetro antero-posterior de 3.5 cm. Las cuerdas vocales masculinas tienen una longitud entre 2 y 2.5 cm. Estas diferencias marcadas generan discriminación en el valor promedio de los parámetros de emisión acústica para cada género.

Edad

La laringe es un órgano con características sexuales secundarias, cuya maduración corre paralela a la del diencéfalo [4]. Esta maduración se prolonga a largo de los distintos períodos vitales que determinan las modificaciones estructurales y fónicas notables. Los grupos de edad que se consideran son los siguientes [3]:

- *Neonatal*:Se caracteriza por las altas frecuencias. El ataque del sonido es brusco, de fuerte intensidad y modulación muy reducida.

- *Primera Infancia*: El ataque se hace menos brusco. A los 18 meses aparece la modulación

vocal.

- *Segunda Infancia*: Las variaciones vocales llegan hasta una octava y media de extensión.

- *Pubertad*: La mutación vocal se produce en el varón entre los 13 y 14 años, y en la mujer, entre los 14 y 15 años. Al cumplirse el descenso laríngeo se hace notable la disminución de las frecuencias de los sonidos producidos.

- *La senilidad vocal*: es más precoz en la mujer que en el hombre y se presenta más marcada en la voz cantada que en la hablada (60 - 70 años). En la mujer el tono de la voz se hace grave.

En general, se considera que las modificaciones más marcadas de la voz se dan en la pubertad para el varón, mientras en la senectud para la mujer.

Uso de la voz

En la práctica se consideran los siguientes niveles de empleo vocal [4]:

- *Usuario selecto:* Corresponde al usuario selecto o especial, en quien aún una ligera aberración vocal le genera consecuencias desastrosas (por ejemplo, la mayoría de cantantes y actores).

  Profesional de la voz: Se refiere a personas en quienes una moderada disfunción vocal impide el adecuado desempeño de sus labores, (por ejemplo, la mayoría de sacerdotes, conferencistas y operadores de teléfonos, fonoaudiólogos, profesores, locutores, entre otros).

- *No profesionales y no vocales*: Se refiere al trabajador que no da a su voz un uso profe-

sional, como ejemplo se tiene los obreros, oficinistas, etc. Si bien algunas personas de este grupo pueden sufrir morbilidad significativa como resultado de un trastorno vocal, éste no les impide realizar sus labores regulares de trabajo.

Paralelamente, orientados a la voz cantada, se ha desarrollado como criterio de clasificación, el timbre de la voz, el cual se define como la cualidad que permite diferenciar dos sonidos, que presenten la misma intensidad y frecuencia. El timbre corresponde al número de armónicos que conforman el sonido y, en parte depende del grado de tersura de las cuerdas vocales, de su modo de vibración, y de las medidas de las cavidades de resonancia (senos paranasales, cavidades supralaríngeas, cavidad orofaríngea). Se han distinguido dos timbres en la voz humana: Timbre vocálico y timbre extravocálico. El timbre vocálico corresponde a circunstancias fisiológicas condicionables, incluyendo aquí todas las técnicas de aprendizaje; mientras, el timbre extravocálico depende exclusivamente de la conformación laríngea y sus cuerdas, y es el que caracteriza la voz común de cada individuo.

1.1.3 Clasificación lingüistíca

El estudio de los sonidos del habla está relacionado, tanto con la *fonética*, como con la *fonología*. En el caso de la fonética, se estudia el inventario de los sonidos de una lengua respecto a las diferencias articulatorias perceptibles [5]. Mientras, la fonología por su parte, organiza estos sonidos dentro de un sistema y establece unidades de sonido o *fonemas*, que son las unidades fonológicas más pequeñas en que se puede dividir un conjunto fónico, su característica principal es la capacidad para diferenciar significados. En la mayoría de los lenguajes, los fonemas pueden ser clasificados en dos clases:

- *consonantes*, que corresponden a aquellos sonidos producidos como consecuencia del choque o el roce del aire con alguno de los órganos fonatorios. Su clasificación se realiza atendiendo al lugar o punto de articulación, a su modo de articulación, a la vibración de las cuerdas vocales y a la acción del velo del paladar.

- *vocales*, o sonidos producidos por la vibración de las cuerdas vocales de la laringe. Su diferente timbre se debe a la variación del volumen de la cavidad bucal que actúa en calidad de elemento de resonancia, y a las diferentes posiciones de la lengua, que puede levantarse por la parte anterior o posterior de la cavidad oral, acercándose al paladar duro. Por lo tanto, no existen obstáculos significativos al pronunciarlas, tienen el máximo de *sonoridad* y *perceptibilidad*, siempre son centro y no margen de sílabas, y son sonidos sonoros. Las vocales se pueden clasificar según las variaciones espaciales en la cavidad oral, producidas por el movimiento y ubicación de la lengua.

1.2 Patologías en la especificación de la voz

Las particularidades anatómicas y fisiológicas del tracto vocal, en general, están determinadas por diferentes características cuyas variaciones definen la tipicidad (en el sentido de la normalidad del hablante referida a condiciones dadas) o atipicidad de la voz. Las patologías están definidas para cambios o variaciones fuera de los limites determinados como normales en la producción de voz. Se conoce una gran cantidad de alteraciones de la voz y el habla, que de manera significativa se reflejan en la naturaleza física de la señal. Por cuanto, la presencia de patologías en los pliegues vocales puede causar cambios significativos en los patrones de vibración normales de los

mismos, y que a su vez desmejoran la calidad de la producción vocal, en la práctica, tiene importancia el análisis de las anomalías congénitas, o aquellas que hayan sido adquiridas, pero que de manera evidente influyan en la particularidad de la voz y el habla, un ejemplo de esto puede ser las alteraciones de la voz debidas a la patología del aparato de voz periférico, así, cualquier cambio en la laringe condiciona fuertes perturbaciones en las funciones de la producción vocal, las cuales pueden ser clasificadas en dos grupos: aquellas que generan la *disfonía* y las que llevan a la *afonía* [6].

1.2.1 Voz eufónica o normal

Se determina como la emisión coordinada, armónica con todas las cualidades y buen manejo de los niveles anatómico-fisiológicos que participan en la producción vocal como es el nivel respiratorio, resonancial, auditivo, emisor, hormonal y de comando [2].

1.2.2 Disfonía

Se define como la alteración de una o varias de las cualidades y características normales de la voz. Sin embargo, existen por una parte, voces *alteradas* no patológicas y, por otra, dificultades vocales que carecen de traducción acústica. En [7], se define la disfonía como un trastorno momentáneo o duradero de la función vocal considerado tal, por la propia persona o por su entorno. Frecuentemente, se manifiesta por la alteración de uno o varios parámetros de la voz, que son, por orden de frecuencia, el timbre, la intensidad, y altura tonal [7]. Algunas de las causas de la disfonía son

- *Laringitis aguda.* Ocurre por la inflamación de las cuerdas vocales debido a infección viral o al uso excesivo de la voz.

- *Nódulos de cuerdas vocales.* Aparecen en personas con mal uso vocal; que hablan muy alto, durante demasiado tiempo, o con mala técnica de emisión vocal.
- *Reflujo gastroesofágico.* El reflujo de material gástrico, sobre todo durante la noche, puede producir irritación de las cuerdas vocales y disfonía.
- *Parálisis de cuerdas vocales.* Por afectación del nervio recurrente debido a cirugía del tiroides o compresión, consecuencia de tumoraciones, o sin causa aparente.
- Otras causas pueden ser también alergias o traumas de la laringe.

1.2.3 Afonía

las razones de su aparición se deben a problemas en el sistema nervioso central, así como en patologías de la laringe. En este caso la voz no se genera, y el habla es posible solo en forma de susurro. La afonía se debe a la parálisis y/o cortes en los músculos de la laringe, debido a la afección de la corteza cerebral o del cerebelo, además, en caso de problemas de infecciones y traumas del nervio inferior de la laringe o en alguna de sus ramificaciones. Como resultado de la parálisis de los músculos que sirven para la contracción y dilatación de la laringe, las cuerdas vocales no se cierran completamente y la voz desaparece.

Se ha encontrado que en muchos pacientes, las disfunciones en la laringe están caracterizadas por [8]:

- Incremento en el grado de ronquera, debido a que la voz de estos pacientes contiene componentes de ruido
- Grandes variaciones en el pitch y las amplitudes pico del mismo.

- Quebrantos en la generación del pitch durante la emisión de vocales sostenidas.
- Presencia de componentes subarmónicos en el espectro de la vocal.
- Distorsión en la forma de los pulsos del pitch.
- Presencia de componentes de ruido de alta frecuencia.

1.3 Evaluación funcional de la voz

Con el fin de dar una adecuada orientación en el trabajo de entrenamiento ó de re-educación, se debe llevar a cabo la observación directa de las estructuras que intervienen en la producción vocal, para determinar la influencia orgánica sobre el producto de la emisión que se está desarrollando y el estado en que se encuentra su comportamiento vocal.

1.3.1 Evaluación clínica

Los principales objetivos del examen clínico de la voz son [9]:

1. Realizar el diagnóstico etiológico, en orden a determinar el grado y extensión de la enfermedad etiológica,
2. Evaluar el grado y la naturaleza de la disfonía,
3. Determinar el pronóstico y monitorear sus cambios.

El estudio de la voz, además de realizar el diagnóstico de cualquier enfermedad etiológica, obtiene información sobre el estado en que se encuentra cada uno de los aspectos evaluados que intervienen

en la emisión vocal, precisando así, las cualidades de la voz: intensidad, altura tonal, timbre, duración. Se considera que para obtener un estudio completo de la voz se requiere [10]:

- Valoración del Otorrinolaringólogo
- Entrevista estructurada
- Análisis acústico de la voz
- Examen respiratorio y de órganos fonoarticuladores
- Examen de la postura corporal
- Prueba subjetiva de la voz.

Es importante medir de forma objetiva el rendimiento de la función vocal de una persona, así como la desviación en su posible deterioro con relación a la norma. Entre los métodos existentes para el análisis y diagnostico del tracto vocal se encuentran: la glotografía, laringoscopia, electromiografía, y análisis acústico de emisión de señales de voz. Recientemente los investigadores han venido aumentando su interés por el análisis acústico de las voces normales y patológicas. Una de las razones para esta tendencia es que los métodos acústicos tienen el potencial de las técnicas cuantitativas para la valoración clínica del funcionamiento del tracto vocal y la laringe.

1.3.2 Análisis acústico de la voz

El análisis acústico de un sonido articulado (comprendido entre los 16Hz y los 20.000Hz) consiste en determinar los indicadores físicos de las vibraciones que lo constituyen tales como la frecuencia fundamental, la intensidad, la composición espectral, y las variaciones del sonido modificadas

por la resonancia que actúan originando el producto sonoro percibido [11]. Los procedimientos de análisis y síntesis del espectro acústico del habla, mediante técnicas de proceso digital, han permitido avanzar en la investigación sobre el análisis de los componentes físicos de la voz normal [2]. En los últimos años ha crecido el interés por el análisis acústico automatizado de voces normales y patológicas como un método alternativo para el diagnóstico. Este tipo de análisis demuestra grandes ventajas sobre los métodos tradicionales debido a su naturaleza no invasiva y a su potencial para proveer una medida cuantitativa acerca del estado clínico del funcionamiento de la laringe y el tracto vocal. De este modo, un sistema automático, confiable, preciso y no invasivo para el reconocimiento y monitoreo de anormalidades del habla es una herramienta necesaria en su valoración y evaluación. Actualmente, existe la tecnología que permite evaluar de manera objetiva la acústica y fisiología del fenómeno, además, provee la retroalimentación visual de los mecanismos de producción vocal, para comprobar el diagnóstico realizado con pruebas subjetivas. El empleo de sistemas computarizados en la caracterización acústica y representación de la voz, provee la posibilidad de analizar indicadores imperceptibles al oído humano, lo que ha permitido adoptarlos como una herramienta de apoyo al diagnóstico con una amplia y creciente aceptación. En este sentido, se han diseñado varios procedimientos, basados fundamentalmente en la medición de los parámetros o *características acústicas* (CA) de la voz y de los fenómenos aerodinámicos que intervienen en la emisión vocal (*Análisis acústico de Voz AAV*). Estos procedimientos permiten establecer el diagnóstico de la alteración vocal, pero son interesantes en múltiples aspectos: proporcionan una imagen inicial de algunas deficiencias que manifiesta el malestar vocal y que permiten que la persona comprenda mejor su trastorno; en ocasiones orientan la reeducación al sugerir la aplicación de técnicas especializadas, según las deficiencias que se hayan encontrado, y

facilitan, asimismo, el seguimiento de la evolución durante el tratamiento demostrando, por ejemplo, la existencia de la mejoría de un parámetro cuya valoración subjetiva por parte del paciente o del terapeuta puede ponerse en tela de juicio; por último, estos métodos pueden utilizarse para detectar a personas con riesgo, a las que podría aplicarse provechosamente una pedagogía preventiva [7].

Capítulo 2

Proceso digital de señales de voz

Las aplicaciones del procesamiento digital de las señales de voz son amplias, entre las cuales cabe destacar: la síntesis, el reconocimiento, la compresión, la transmisión y el mejoramiento de la calidad, entre otras. En el caso particular del reconocimiento de voz, el procesamiento digital se inicia con la adquisición y adecuación de la señal que incluye los procesos de conversión electro-acústica, amplificación, filtrado pasabajo inicial, así como la conversión A/D; todos los anteriores procesos, típicamente, realizados a nivel de hardware. La etapa siguiente consiste en el preproceso e incluye los procedimientos de segmentación, pre-énfasis, filtración y remoción de perturbaciones. Por último, está la etapa de análisis de información que incluye la estimación y selección de los parámetros y características de acuerdo al tipo de aplicación, al cual está orientado el proceso digital de señales de voz.

2.1 Adquisición y adecuación de señales de voz

El objetivo de esta etapa es el acondicionamiento de las señales salidas de los sensores a una forma adecuada para su posterior análisis, y consta de los siguientes procedimientos:

2.1.1 Conversión electro-acústica

El primer paso en la recolección de señales es la transformación, que en el caso concreto se realiza mediante el micrófono, el cual ejecuta la *conversión* a energía eléctrica, de los desplazamientos del aire debido a cambios de presión, correspondientes a la forma natural de generación de la voz.

La conversión electro-acústica es del tipo análoga, esto es,

$$\Gamma\{x(t)\} = y(t) \sim kx(t) \tag{2.1}$$

siendo $x(t)$ la señal original de voz y k la constante de linealidad de conversión. En la práctica, hay distorsiones de amplitud y fase que conllevan a la dependencia no lineal entre la entrada y salida del conversor, haciendo que este sea una fuente potencial de errores en el registro de las señales. En este sentido, la selección del micrófono es importante durante el registro de voz, el cual se sustenta básicamente en las siguientes características técnicas:

- *Respuesta de frecuencia.* Se debe garantizar el que la curva de respuesta de frecuencia sea constante con el menor rizado o clase de variación, conjuntamente con el mayor ancho de banda posible.

- *Direccionalidad del patrón.* El patrón es entendido como la forma de concentración de la energía recibida por el micrófono con respecto al ángulo cero desde la fuente de emisión sonora. Los patrones de recepción pueden ser omnidireccionales (patrón circular) o direccionales (cardioide, elípticos, etc.). En el reconocimiento de voz es preferible el uso de micrófonos con patrón direccional, comúnmente del tipo cardioide, a fin de orientar al máximo la emisión del hablante sobre el sistema (*close-talk*) y, así mismo, reducir las emisiones de ruido de fondo [12].

2.1.2 Conversión A/D y compresión

Discretización

En diferentes tareas de reconocimiento, la frecuencia máxima nominal de la señal de voz es tomada igual a 4kHz, debido a que la mayoría de sonidos vocálicos tienen energía espectral significativa hasta este valor de frecuencia como se puede observar en la figura 2.1. Sin embargo, en las tareas asociadas al análisis acústico de voz, deben ser considerados los armónicos superiores, con frecuencias del orden de 5kHz. Por lo tanto y acorde con el teorema del muestreo, la señal de voz debe ser digitalizada con una frecuencia de 10 kHz o superior.

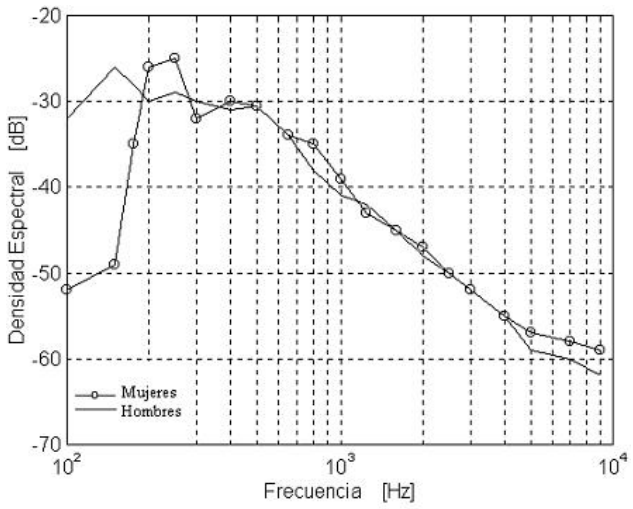


Figura 2.1: Espectro medio de energía de una señal de voz calculado

De otra parte, la voz exhibe un rango dinámico entre los 50 y 60 dB, por lo cual sería suficiente codificar la señal con 8 bits, sin embargo, para el caso de sistemas de procesamiento de voz de alta calidad, se utilizan entre 11 y 20 bits, ya que como es conocido, cada bit adicional contribuye a

mejorar la relación señal a ruido en 6 dB aproximadamente.

Compresión

La compresión tiene como objetivo la reducción del tamaño efectivo de la señal de voz para su proceso en tiempo real, o bien para su almacenamiento, cumpliendo con los requisitos de calidad que imponga el sistema específico de aplicación. La clasificación de los métodos de compresión es relativa a la naturaleza de su diseño y comprende dos clases:

a. *Compresión entrópica*. Considera los flujos de datos sin importar el contenido de la señal útil que representan. Es una técnica sin pérdidas y completamente reversible. Dentro de esta clasificación se pueden considerar la codificación por longitud de series, que aprovecha el hecho de que en múltiples tipos de datos son comunes las cadenas de igual contenido, las cuales pueden reemplazarse por un marcador especial no permitido en los datos en otras condiciones, seguido del símbolo que indica la serie y luego la cantidad de veces que se repite. Si el marcador especial ocurre entre los datos, se duplica (como en el relleno de caracteres) [13]]. Otro tipo importante de codificación entrópica es la codificación estadística, donde se emplea un código corto para representar símbolos comunes y un código largo para los símbolos menos frecuentes. Como ejemplos se tienen la codificación *Huffman* y el algoritmo *Ziv - Lempel*.

b. *Compresión por fuente*. Se basa en modelos aproximativos que buscan una representación de la estructura de las señales de voz, lo que, generalmente, presenta pérdidas. Ejemplos de este tipo, son la codificación diferencial ADPCM, y LPC [14], entre otros. Los casos más difun-

didos en compresión de audio por fuente, son los que hacen uso de transformaciones como la de Fourier, las Wavelet, etc. Aún cuando se eliminen algunos componentes de frecuencia (dejando por ejemplo, solo los n componentes más importantes), se podrá reconstruir la señal con la fidelidad necesaria. El valor n estará determinado por los requerimientos de calidad del sistema. La eliminación de componentes espectrales implica una reducción en el tamaño de los datos. Los criterios para esta eliminación dependen del proceso. Por ejemplo, en la codificación MP3, se efectúa el filtrado mediante el modelo psicoacústico del oído humano (además de efectuarse la compresión empleando la Transformada Discreta del Coseno DCT [15]), con el que se eliminan los elementos redundantes en la muestra.

2.1.3 Fuentes de degradación de señales acústicas

En general, un sistema robusto de reconocimiento automatizado de voz se ve afectado por una serie de factores que degradan su rendimiento, originados por diferentes fuentes, entre ellas las siguientes: características electro-acústicas del hardware de conversión (parlantes, conectores, micrófonos, etc.), adquisición y acondicionamiento de señales, representación y proceso digital, cambios de las condiciones ambientales, etc. Entre los factores que causan mayor degradación en la precisión del reconocimiento de voz, están el ruido aditivo y la filtración lineal. Otras fuentes de degradación incluyen los efectos de articulación inducidos por la influencia ambiental (*efecto Lombard*), el ruido transiente con alta energía e interferencias producidas por señales de voz de personas hablantes ubicadas cercanamente (*efecto fiesta*).

El ruido aditivo

El rendimiento de un sistema de reconocimiento de voz depende significativamente de los niveles de ruido que se hayan tenido durante el registro de señales en el proceso de entrenamiento y prueba [16]. En la mayoría de los casos el ruido ambiental o de fondo, es considerado aditivo y, por tanto, puede ser modelado como un proceso $\eta(t)$ estacionario gaussiano aditivo con media m_η y varianza σ_η^2, que no posee ninguna naturaleza de correlación con la señal de voz $x(t)$, de tal manera, que la señal resultante es:

$$y(t) = x(t) + \eta(t), \tag{2.2}$$

Filtración lineal

La señal puede de voz puede sufrir una serie distorsiones espectrales durante su producción, registro y proceso electrónico. El espacio físico donde se dispone el sistema de recolección de señales puede presentar un nivel cambiante de reverberación y, esto genera influencia sobre el espectro de la señal. Así mismo, cuando la configuración de los micrófonos difiere en los procesos de entrenamiento con los empleados en la adquisición de las señales a identificar, el rendimiento del procedimiento de reconocimiento, en general, es peor. El modelo de ambas formas analizadas de

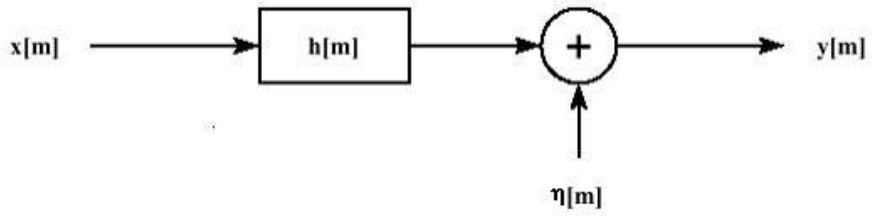


Figura 2.2: Modelo de degradación de la voz debida a la filtración lineal y al ruido aditivo

degradación de la señal de voz es representado en la figura 2.2, en el cual la señal de voz es pasada por un filtro lineal con respuesta a impulso $h(t)$ desconocida, cuya salida es distorsionada por el ruido aditivo $\eta(t)$. Las densidades espectrales de potencia (*Power Spectral Density* - PSD) de cada uno de los procesos en 2.2 serán [17]:

$$\hat{s}_y(\omega) = \hat{s}_x(\omega) \left|\hat{h}(\omega)\right|^2 + \hat{s}_\eta(\omega)$$

Donde $\hat{s}_x(\omega)$ es la PSD de la función $x(t)$ y $\hat{h}(\omega)$ la función de transferencia del canal. Las técnicas de reducción (compensación) de ruido de fondo intentan disminuir la influencia de la perturbación aditiva, mientras las técnicas de ecualización tratan de disminuir las distorsiones producidas por la filtración lineal [18].

Ruido de fondo en el AAV

En los sistemas de AAV, el registro de las señales de voz es severamente degradado por ruido que puede ser distinguido de acuerdo al tipo de interacción con la señal útil de voz, así, se diferencia el ruido aditivo y el ruido convolucional. Las componentes de ruido aditivo pueden ser producidas por el ruido de fondo como máquinas y equipos eléctricos, mientras que el ruido convolucional, generalmente es relacionado con las propiedades acústicas de la sala. Sin embargo, otros hablantes y fuentes de ruido, producen campos acústicos de intensidad comparable con el hablante principal, que no pueden ser modelados como ruido aditivo. El ruido del fondo causa directamente errores en la estimación de las características acústicas de la señal, afectando su precisión de estimación [19], [20]. De otra parte, el AAV con requerimientos altos de precisión se debe desarrollar en situaciones donde exista equilibrio entre las condiciones de prueba y las condiciones de entrenamiento, así mismo, es importante asegurar homogeneidad en la señal de entrada del analizador con las de

la entrada al sistema de reconocimiento [21], [22]. El rendimiento del AAV empeora si no se toman las medidas necesarias durante el registro de señales, realizándolo bajo condiciones de ruido ambiente y con micrófonos que sean substancialmente diferentes de las usados durante la fase de entrenamiento [12], [23], por esto se buscan métodos robustos de preprocesamiento de la señal para su mejoramiento a la entrada del AAV, así como técnicas de entrenamiento que consideren las variaciones en las estimaciones de los parámetros representativos de la señal de voz [24].

2.2 Preprocesamiento

En el procesamiento de voz, es necesario llevar a cabo el acondicionamiento y normalización de las señales adquiridas, que contienen perturbaciones introducidas por el ambiente durante su registro. Así mismo, la señal de voz contiene segmentos no sonoros, en los cuales se considera la inexistencia de componentes informativas útiles. Estos segmentos de ausencia de voz, se eliminan en el proceso informativo de la señal. Finalmente, se incluye la filtración *pre-énfasis* que acentúa las cualidades espectrales de alta frecuencia de la voz.

2.2.1 Regulación de niveles

La normalización de las señales de voz, en la práctica se considera para dos parámetros: amplitud y longitud de análisis. Los resultados de esta normalización son importantes en la medida en que ambos parámetros cambian en un rango de valores muy alto. En los sistemas de reconocimiento de voz, el entrenamiento de los clasificadores típicamente se hace sobre señales con intensidad y longitud predefinidas, sin embargo, como se ha demostrado, en el momento del reconocimiento,

es prácticamente imposible hacer que el hablante sostenga la intensidad y el tiempo de voz acorde a los valores de entrenamiento. Aún en el caso de pequeñas variaciones de ambos parámetros, que para el oído pasan desapercibidas, se pueden generar serias complicaciones en el reconocimiento automatizado. Además, hay que tener en cuenta, que la intensidad de la señal no solamente depende de la actividad desarrollada por la persona, sino también por las condiciones de registro de la voz, entre ellas, la distancia entre micrófono y los labios, la ganancia de micrófono, las pérdidas en los canales de comunicación, entre otros. Lo anterior establece la necesidad de incluir procedimientos de normalización para hacer similares las señales de voz, tanto en intensidad, como en longitud.

Ajuste de intensidad

Frecuentemente, la normalización por intensidad se lleva a cabo mediante la relación de las señales en los puntos de salida en los dispositivos de preproceso, por ejemplo la salida de los filtros pasabanda se divide por la energía total de la señal. Este método es fácil de implementar desde punto de vista computacional y conlleva a la mejora de los resultados.

En los sistemas de procesamiento digital de voz en tiempo real, la normalización en intensidad se lleva a cabo antes del análisis, e inmediatamente después de la salida del micrófono, o registro electrónico de la señal. Una forma sencilla de la normalización en estos casos, consiste en el empleo de limitadores de nivel (*clipping*), sin embargo, este método es no lineal, generando distorsiones espectrales y ruidos perceptibles auditivamente.

Para evitar el efecto de aparición de componentes espurios debido al efecto no lineal del clipping, la señal de voz se transporta al rango de frecuencias de ultrasonido, donde se realiza, conjuntamente,

el clipping y la filtración, para luego ser de demodulada. La señal obtenida de estas operaciones tiene la forma:

$$z_o(t) = \cos(\varphi(t)) \tag{2.3}$$

que se caracteriza por una amplitud constante, y para la cual solamente cambia en el tiempo su fase instantánea $\varphi(t)$. Esta señal denominada *voz de nivel constante* conserva la inteligibilidad suficiente y no tiene los problemas de ruidos debidos al clipping, además, representa la señal normalizada en amplitud, sin embargo la naturalidad de la señal decae.

Un método mejorado consiste en mezclar la señal original $x(t)$ de voz con una señal de ultrasonido $v(t)$, cuya frecuencia de trabajo Ω es mucho mayor que la máxima frecuencia de la señal $y(t)$, esto es, $\Omega \gg \omega_m$, y la amplitud a de la señal $v(t)$, es tan grande que se cumple que $a > |y(t)|$, entonces, el clipping de la suma $y(t) + v(t)$ corresponde a la modulación PWM, seguidamente se emplea el promedio en intervalos $t = 2\pi/\Omega$, la señal obtenida es filtrada en las frecuencias de audio incluyendo el valor DC. Así se obtiene la siguiente señal

$$y_0(t) = \frac{c}{a} y(t)$$

donde la constante c depende del parámetro de clipping. Esta señal coincide en forma con la señal original, pero por magnitud es inversamente proporcional al valor a de la amplitud de la señal de ultrasonido. Al realizar el promedio de la señal $|y(t)|$, en el intervalo de tiempo del mismo orden de el periodo del pitch, entonces se tiene:

$$\left|\overline{y_0(t)}\right| = \frac{c}{a} \left|\overline{y(t)}\right|$$

donde c es una constante en el intervalo dado. La implementación de la última expresión, se puede llevar a cabo utilizando un modulador. De esta manera, conservando en cada momento del

tiempo la señal constante $\left|\overline{y(t)}\right|$ proporcional a la amplitud a de la señal de ultrasonido, se puede lograr que $y_o(t)$ sea constante independiente de la magnitud $y(t)$ y conservando en este caso la forma de la señal de entrada. Este método provee la normalización confiable del nivel de la señal (hasta 3 dB) para cambios de rango dinámico de la señal de entrada de voz de hasta 40 dB [6]. Otro método de ajuste, corresponde a los sistemas de control automático de ganancia (CAG) . Por cuanto estos conllevan a efectos no lineales que generan distorsiones de estructura compleja en las señales originales de voz, en algunos casos se incluyen modificaciones a las compensaciones del efecto no lineal [6].

Ajuste de tiempo

En cuanto a la normalización de la señal de voz por longitud de tiempo, típicamente, los sistemas de reconocimiento exigen la igualdad en los patrones contra las señales a comparar exigiendo la normalización de las señales de entrada. El método de ajuste de tiempo más sencillo consiste en la normalización lineal, como resultado la señal uniformemente se comprime o expande hasta la longitud del patrón. En este caso, sin embargo no siempre se garantiza la correspondencia adecuada de los intervalos de voz, por cuanto, la velocidad de pronunciación de voz de la persona no es uniforme (*tempo*), y por esto diferentes intervalos de voz con idéntica información pueden ser representados de manera desigual en el eje del tiempo.

Otros métodos consideran la normalización no lineal, por ejemplo en [6], utilizando algoritmos de programación dinámica se desarrolla la técnica de alineamiento temporal o (*Dynamic Time Warping- DTW*) para la correspondencia en el reconocimiento de las señales de voz con sus respectivos patrones de contornos de intensidad del pitch y de los formantes. Después de realizar

la compresión o expansión lineal del contorno, y haciendo coincidir los puntos de inicio y final del patrón con los de la señal a reconocer, seguidamente se realiza el proceso de selección de la función de deformación, tal que minimice, por un criterio de error dado, la diferencia entre las señales comparadas [19], [12]. Esta técnica de normalización exige su aplicación a todos los patrones de referencia en la etapa de entrenamiento, en la cual además de determinar el valor medio para cada muestra de entrenamiento, se calcula también la dispersión (desviación estándar) sobre diferentes segmentos de los contornos en un número apreciable (> 20). Los valores de dispersión obtenidos así, se emplean para la determinación de los pesos de cada segmento en función inversamente proporcional; durante el cálculo del grado de similitud de dos contornos, con mayor peso se caracterizan aquellos segmentos del contorno que tengan mayor estabilidad en el conjunto de señales de voz de entrada.

Una variación de la técnica anterior, consiste en la sincronización artificial de la señal de voz mediante la deformación lineal por segmentos. [6], los cuales corresponden a puntos de referencia en la estructura del patrón de voz claramente identificables. Tales puntos pueden ser la aparición de sonidos fricativos, los golpes glóticos, etc.

2.2.2 Detección activa de voz

Los sonidos producidos cuando los pliegues vocales están muy cerca, es decir, la glotis se estrecha y se produce una vibración, se denominan *segmentos sonoros* (*voiced*); mientras aquellos que se producen sin la vibración de los pliegues vocales se denominan *segmentos sordos*(*unvoiced*). Dado que la información de las señales de voz se considera que está concentrada en los segmentos sonoros, es entonces, fundamental la determinación de su presencia. Así, en los momentos de

ausencia de sonidos se considera únicamente la presencia del ruido. Este proceso de diferenciación de los tipos de segmentos es denominado *detección activa de voz* (*Voice Active Detection - VAD*) e incluye la estimación de los momentos de inicio y finalización de cada segmento sonoro. La segmentación puede ser realizada por diferentes algoritmos entre los cuales están:

a. *Cálculo de las densidades de energía y cruces por cero [2]*

b. *Cálculo de los momentos de cierre glótico (GCI)*

Cálculo de las densidades de energía y cruces por cero

La energía de la señal está dada por

$$E[n] = \sum_{m \in \mathbb{Z}} x^2[m] w[m-n] \tag{2.4}$$

donde w es la función ventana de análisis. Sin embargo, el cálculo de los cuadrados en la ecuación (2.4) requiere bastante tiempo de proceso, por lo tanto se usa el cálculo de la magnitud en lugar de los términos cuadráticos:

$$E[n] = \sum_{m \in \mathbb{Z}} |x[m]| w[m-n] \tag{2.5}$$

De otra parte, se determina la densidad de cruces por cero, definida como el número de veces que la señal cambia de signo, con el objetivo de validar el análisis de energía. La densidad de cruces por cero se determina usando la ecuación (2.6):

$$Z[n] = \frac{1}{N} \sum_{n \in \mathbb{Z}} |\mathrm{sgn}(x[m]) - \mathrm{sgn}(x[m-1])| \, w[m-n] \tag{2.6}$$

Donde N es la apertura de la ventana de análisis de la señal de voz y sgn es la función signo. La cantidad de energía es mayor durante la emisión vocal, mientras la densidad de cruces por cero es

mayor en su ausencia (ver figura 2.3). Con el fin de determinar el inicio y el final de cada palabra se debe tomar inicialmente una realización del ruido de fondo (el valor aceptado de apertura de ventana de análisis en tiempo para el proceso de voz es 100 ms) y determinar la energía media y la densidad de cruces por cero del ruido. Basados en las características del ruido se pueden fijar umbrales de energía y de cruces por ceros para la segmentación. En la figura 2.3 se puede observar que la energía y la densidad de cruces por cero de la señal son mucho mayores durante los intervalos sonoros que en los sordos. Existen algunos sonidos tales como */s/*, */t/*, */ch/* entre otros,

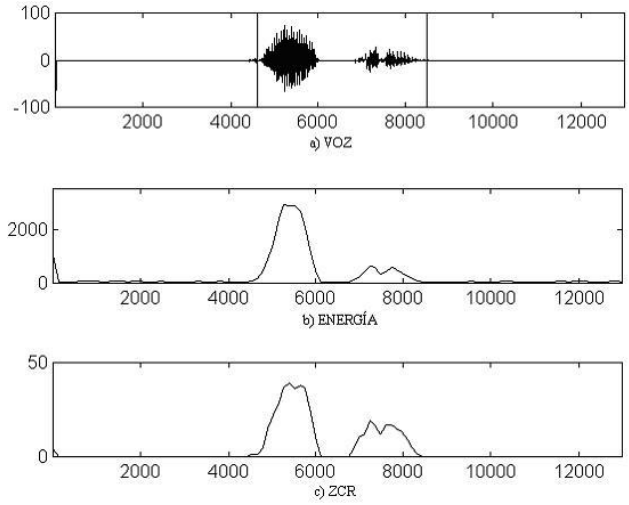


Figura 2.3: Palabra *cuatro* a) Detección activa de voz, b) Energía media de la señal, c) Densidad de cruces por cero.

cuyos valores instantáneos de energía y densidad de cruces por cero puede ser bastante bajos, incluso muy similares a los del ruido de fondo, por ejemplo la */t/* de la palabra *cuatro* mostrada en la figura 2.3. Por esta razón es necesario fijar un umbral de silencio, el cual determina el número máximo de tramas con energía y densidad de cruces por cero inferiores a los umbrales que puede ser considerados como parte de los segmentos sonoros. Por otra parte, en el medio ambiente

puede presentarse el ruido impulsivo que supere los umbrales de energía o cruces por cero durante un periodo de tiempo muy corto, lo cual puede conducir a la detección incorrecta de inicio del segmento de voz. Para evitar este problema se fija un umbral de longitud mínima que pueda tener el segmento sonoro. Las longitudes medias de segmentos sonoros e insonoros varían de un idioma a otro, por lo cual la determinación de los umbrales de segmentación se debe escoger empleando técnicas estadísticas para cada caso. En [12] se presenta el procedimiento de estimación para el caso específico de diccionarios reducidos de voz en el idioma español. La realización en línea de los algoritmos de VAD sobre intervalos cortos de análisis (típicamente, 10 ms), se lleva a cabo una vez se tienen fijados los umbrales de detección. La decisión sobre la aparición de un segmento sonoro se realiza al satisfacer conjuntamente los valores correspondientes de energía y cruce por ceros, en caso contrario se toma la hipótesis de segmento sordo.

2.2.3 Reducción de perturbaciones

Un alto número de técnicas de reducción de ruido han sido propuestas, tanto para un solo micrófono, como para arreglos de varios micrófonos. En el caso del AAV, es común el empleo de la primera disposición de conversión electro-acústica.

2.2.4 Filtración de pre-énfasis

Una vez detectados los segmentos sonoros de voz, sobre cada uno de ellos se realiza la etapa de filtración que acentúa las frecuencias altas de la señal de voz, debido a que el modelo del tracto vocal utilizado atenúa fuertemente estas componentes. El filtro de pre-énfasis obedece a la

expresión recursiva dada por la ecuación:

$$x_p[n] = x[n] + a_{pre}x[n-1] \tag{2.7}$$

cuya función de transferencia asociada está dada por:

$$H(z) = 1 + a_{pre}z^{-1} \tag{2.8}$$

La adecuada filtración pre-énfasis asegura la uniformidad en los niveles medios espectrales de cada segmento sonoro de análisis. Los objetivos de utilizar el filtro pre-énfasis son

- Reducir el efecto de la pendiente espectral de -20dB presente en los segmentos de voz.
- Amplificar la zona del espectro superior a 1kHz donde la percepción auditiva se hace sensible.

El valor óptimo del coeficiente de pre-énfasis a_{pre}, está dado en función de la señal de entrada; sin embargo, se escoge un valor constante con el fin de acentuar adecuadamente la estructura de los formantes. Un valor razonable, en este caso, oscila entre entre 0.9 y 0.95. La respuesta en frecuencia de este filtro se muestra en la figura 2.4. La apreciación del efecto del filtro de pre-énfasis, se observa mediante el uso del espectrograma, el cual provee una medida cualitativa de la acentuación de las frecuencias altas. En la parte inferior de la figura 2.5, se observa la aparición de frecuencias alrededor de los 7000Hz que no eran visibles antes del proceso de la filtración pre-énfasis.

2.2.5 Ventaneo

El proceso digital de señales de voz se hace empleando técnicas de análisis en intervalos cortos de tiempo, suponiendo la estacionariedad del proceso aleatorio de voz. En este caso, la estimación

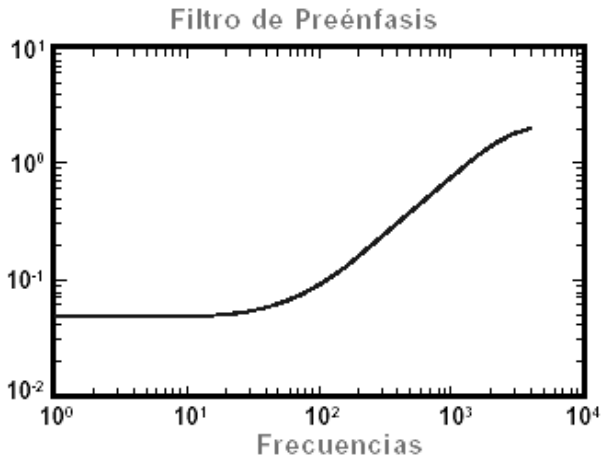


Figura 2.4: Respuesta en frecuencia del filtración de pre-énfasis

de los parámetros de voz se realiza sobre porciones de la señal, con longitud o apertura suficiente para considerar la invariabilidad estadística en el tiempo del proceso. Se considera que las señales de voz presentan una dinámica estacionaria en intervalos entre los 20 y 40 ms [12]. Sea $w(n)$ una ventana de secuencia real con apertura finita, utilizada para seleccionar el intervalo corto de análisis de la señal. Se considera que las ventanas son secuencias causales, es decir que comienzan en $n = 0$ y su apertura es representada por N. La mayoría de las ventanas utilizadas son simétricas con respecto al tiempo $(N-1)/2$ y su transformada de Fourier está dada por:

$$\hat{w}(\omega) = |\hat{w}(\omega)|\, e^{-j\omega[(N-1)/2]} \tag{2.9}$$

Donde el término de fase es de dependencia lineal, correspondiente al retardo de la ventana que la hace causal. Para obtener el intervalo corto de análisis de la señal $s(n)$ que finalice en un tiempo igual a m se utiliza la siguiente relación:

$$v(n) = x(n)w(m-n) \tag{2.10}$$

Donde se observa que las características temporales del intervalo han cambiado con respecto a la

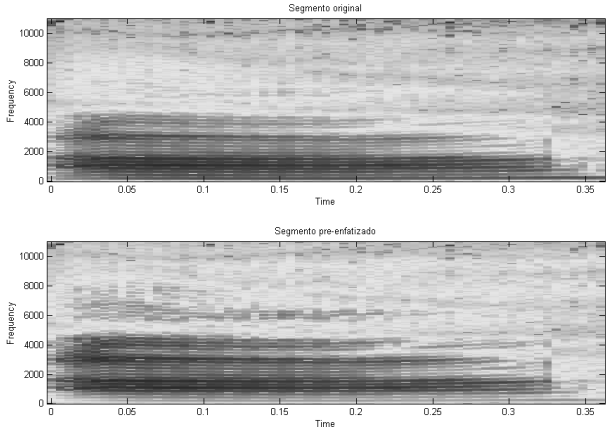


Figura 2.5: Espectrogramas para el fonema vocálico /a/. Arriba: segmento original. Abajo: segmento con pre-énfasis

señal $x(n)$ al ser multiplicado por la ventana. De la misma manera, las características espectrales también varían, la cuales para $v(n)$ quedan determinadas por la convolución entre la señal original y la ventana en el dominio de la frecuencia:

$$\hat{v}(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{x}(\omega - \theta)\hat{w}(-\theta)e^{-j\theta m} d\theta \tag{2.11}$$

Asumiendo que la ventana está centrada en $n = 0$, esto es, $\hat{w}(\omega) = |\hat{w}(\omega)|$, la ecuación anterior se puede escribir como:

$$\hat{v}(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{x}(\omega - \theta)\, |\hat{w}(\theta)| d\theta \tag{2.12}$$

$|\hat{w}(-\theta)|$ ha sido reemplazado por $|\hat{w}(\theta)|$ debido a que la magnitud del espectro es una función par de θ. De (2.12), se deduce que la magnitud ideal de la ventana corresponde a $|\hat{w}(\theta)| \approx 2\pi\delta(\theta)$, para que $\hat{v}(\omega) = \hat{x}(\omega)$. Esto implica que $w(n) = 1$ para todos los valores de n, lo cual no representa una función ventana con apertura finita. Sin embargo, la aproximación anterior se extiende a las ventanas usadas normalmente con el fin de conservar las características espectrales de la señal

$\hat{x}(w)$.

Las ventanas más comunes tienden a tener un espectro pasa bajos con el lóbulo principal en las bajas frecuencias y una serie de lóbulos laterales atenuados. Como ejemplo se tiene la ventana rectangular, que está definida por:

$$w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & Otrovalor \end{cases} \tag{2.13}$$

Otro tipo común de ventana es la de Hamming:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) & 0 \leq n \leq N-1 \\ 0 & Otrovalor \end{cases} \tag{2.14}$$

Con el fin de obtener el espectro de la ventana aproximado a la función delta impulso, esta debe tener el lóbulo principal angosto con alta atenuación en los lóbulos laterales. El lóbulo principal estrecho ayuda a conservar las características espectrales de la señal, mientras la atenuación de los lóbulos laterales evita que el ruido de otras partes del espectro deforme el espectro real en la frecuencia dada. La ventana rectangular conserva intactas las características temporales de la señal, pero realiza un corte abrupto de la misma en los bordes de la ventana. Las características espectrales de la ventana rectangular son las siguientes: lóbulo principal bastante estrecho el cual se decrementa en N. La altura de todos los lóbulos aumenta con N mientras que la atenuación de los lóbulos laterales se mantiene mas o menos constante al rededor de -20 dB con respecto al lóbulo principal, lo cual permite que gran cantidad de energía espectral indeseable se introduzca en el espectro a la frecuencia dada. Por estas razones en el procesamiento de las señales de voz no se utiliza la ventana rectangular, prefiriéndose el uso de ventanas, tales como *Hamming, Hanning, Blackman y Kaiser*. Estas ventanas tienden a cambiar las características temporales de la señal,

pero con el beneficio de producir un truncamiento menos abrupto de la señal en los bordes. Otro aspecto importante del proceso de ventaneo es que al multiplicar una señal por la ventana de Hamming o Hanning la longitud efectiva del intervalo de análisis se reduce aproximadamente en un 40% debido a la atenuación de la señal en los bordes de la ventana. Por esta razón, se recomienda que al deslizar la ventana para tomar la siguiente trama de la señal de voz se vuelvan a tomar algunas muestras de la trama anterior, es decir que debe haber traslapamiento entre dos ventanas consecutivas. En el procesamiento de voz se utilizan ventanas con duración entre 20 y 40 ms y se utiliza un traslape entre 10 y 20 ms [12].

2.3 Representación de señales de voz

2.3.1 Representación estacionaria

Cualquier señal $f(t)$ de energía o potencia finita, esto es, $f(t) \in L^2(\mathbb{R})$, puede ser representada por medio de un conjunto de valores o coeficientes $\{f_n \in \mathbb{C}\}$, expresados en dependencia de un espacio funcional de coordenadas, así:

$$f(t) = \sum_n f_n \phi_n(t) \tag{2.15}$$

donde $\{\phi_n(t) \in L^2(\mathbb{R})\}$ consiste de un conjunto de funciones elegidas a priori, denominadas *funciones base*, siendo n el orden de la función dentro del conjunto $\{\phi_n(t)\}$. La descomposición (2.15) en funciones base se denomina *representación espectral generalizada de Fourier*. La elección de conjuntos bases se realiza, preferiblemente, sobre los sistemas ortogonales. En forma general, el desarrollo de la expansión (2.15) involucra una serie con cantidad infinita de términos, cuya

convergencia está garantizada por la completitud del respectivo sistema base de expansión [25]. En el caso de considerar la ortogonalidad de la expansión (2.15), los coeficientes f_n, son determinados de acuerdo a la condición del mínimo error cuadrático medio, el cual es definido como:

$$\overline{\varepsilon^2(t)} = \frac{1}{T}\int_T \left| f(t) - \sum_{n=0}^{N} f_n \phi_n(t) \right|^2 dt, \text{ donde } \overline{\varepsilon^2} \geq 0 \tag{2.16}$$

Con lo cual los coeficientes se determinan por la expresión:

$$f_n = \frac{1}{T}\int_T f(t)\phi_n(t)dt \tag{2.17}$$

El conjunto de los coeficientes $\{f_n\}$, que genera un *subespacio* de $L^2(\mathbb{R})$, se denomina *espectro* y provee una forma de representación paramétrica de la señal $f(t) \in L^2(\mathbb{R})$, mientras el producto escalar $\langle f_n, \phi_n(t)\rangle$ se define como la *componente espectral* de la señal. Cualquiera de las dos formas: la serie generalizada de Fourier o *el espectro* $\{f_n\}$ determinan unívocamente la señal $f(t)$. Debido a que la cantidad de sistemas ortogonales completos es inconmensurable, la elección del mejor sistema base de representación tiene un amplio sentido práctico. En el proceso de voz, se difundió el empleo de las funciones exponenciales de Fourier dado en forma de funciones exponenciales complejas del tipo:

$$\phi_n(t) = e^{jn\omega_0 t}, \tag{2.18}$$

donde $n \in \{0, \pm 1, \pm 2, ...\}$ se denomina *armónico,* siendo $\omega_0 = cte \neq 0$. La forma generalizada (representación integral) de las series 2.18 se conoce como *la Transformada de Fourier* y está definida como:

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt, \tag{2.19}$$

cuya transformación inversa es

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{j\omega t} d\omega \tag{2.20}$$

siendo $\{f(t)\} \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$.

En la práctica, se consideran extensiones del análisis espectral de Fourier, entre las cuales están los siguientes:

Análisis Cepstrum

Dado por la expresión:

$$\breve{f}(\breve{t}) = \left| \int_{-\infty}^{\infty} \log |\hat{f}(\omega)|^2 e^{-i\breve{t}\omega} d\omega \right|^2 \tag{2.21}$$

considerando la función par $f(t)$, la cual se extiende a $|\hat{f}(\omega)|^2$, y por tanto a $\log |\hat{f}(\omega)|^2$. Por lo cual, (2.21), se puede escribir que:

$$\breve{f}(\breve{t}) = 4 \left[\int_{0}^{\infty} \log |\hat{f}(\omega)|^2 \cos \breve{t}\omega d\omega \right]^2 \tag{2.22}$$

El análisis *cepstrum* es empleado en casos en que se tenga un carácter oscilatorio significativo en el espectro, esto es, de hecho un caso muy frecuente.

Sin embargo, la Transformada de Fourier y sus derivaciones son herramientas que permiten trasladar un problema de un dominio al otro, conservando toda la información, pero no brindan la posibilidad de tener información cruzada en ambos dominios al mismo tiempo. Si las señales que se quieren analizar son *no estacionarias* a lo largo del tiempo ni de la frecuencia, entonces no se pueden analizar con estas transformaciones, sino que se requiere tener de forma conjunta información de la evolución temporal y frecuencial de las señales.

Transformada de Gabor

Una solución propuesta al problema de la localización tiempo-frecuencia proviene de los trabajos de Gabor [26], quién introdujo el concepto de ventana, permitiendo así la delimitación de la función a estudiar en el tiempo antes de realizar su descomposición frecuencial de (2.19), siendo $g(t)$ la función ventana, que desliza a lo largo de la función en el tiempo mediante traslaciones, determinadas por el *factor de traslación* τ. La señal a analizar, se multiplica con la ventana en la posición adecuada y seguidamente se lleva a cabo la transformación espectral.

$$S_\tau(\omega) = \hat{f}_g(\tau, \omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (f(t)g(t-\tau))e^{(-j\omega t)}dt \tag{2.23}$$

La función $g(t)$ registra la influencia de una pequeña porción de la función $f(t)$ alrededor del instante de observación $t = \tau$. De esta forma, se estudia la distribución de frecuencia y se logra una mejor localización temporal inexistente en la transformada de Fourier. La selección de la misma ventana corresponde a diferentes criterios de optimización del proceso (en los dominios temporal o frecuencial). En la aplicación de estas técnicas se han estado utilizando los espectrogramas: El *espectrograma de banda ancha* utiliza una ventana $g(t)$ muy estrecha en el tiempo, mientras el *espectrograma de banda estrecha* busca una mejor localización de la distribución de energía en el dominio frecuencial. En la figura 2.6 se aprecia la diferencia en la distribución de energía en el plano tiempo-frecuencia para estos dos tipos de espectrogramas. El primero permite una buena localización temporal, lo cual se aprecia por las franjas verticales. El espectrograma de banda estrecha localiza mucho mejor en frecuencia, reflejado por las franjas horizontales. Está técnica desplaza una ventana de apertura fija a lo largo del dominio temporal de la señal y extrae el contenido frecuencial en dicho intervalo. Las funciones base para esta transformación son generadas

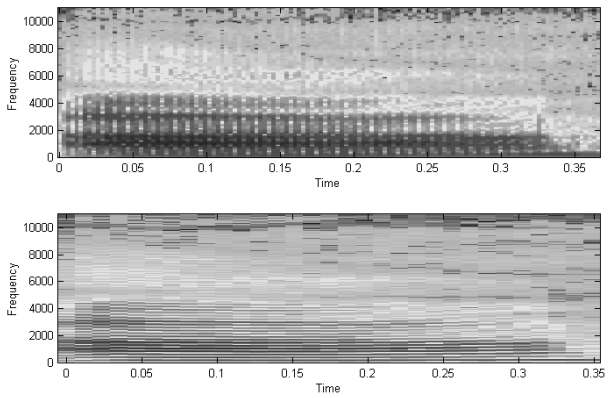


Figura 2.6: Espectrograma correspondiente al fonema vocálico /a/. Arriba: banda ancha. Abajo: banda estrecha

de la modulación y traslación de la función ventana $g(t)$. El obstáculo principal de esta técnica y por consiguiente de todas aquellas basadas en ésta, es la ventana fija tiempo-frecuencia. En la figura 2.7 se muestra el recubrimiento del plano tiempo-frecuencia de acuerdo a la transformada de Fourier y series exponenciales. Además, para poder entender el efecto de la transformada Gabor, se representa el mapeo del plano para el caso de los espectrogramas de banda ancha y banda estrecha. Suponiendo la estacionariedad de la voz, en el proceso digital de señales, es frecuente el empleo del análisis de intervalos cortos de tiempo (*Short Time Fourier Analysis*), en el cual se define el par de transformaciones discretas de Fourier

$$\hat{f}_N[k] = \sum_{n=0}^{N-1} f_N[n] e^{-j2\pi nk/N},\ 0 \leq k < N \tag{2.24}$$

$$f_N[k] = \frac{1}{N} \sum_{n=0}^{N-1} \hat{f}_N[n] e^{j2\pi nk/N},\ 0 \leq n < N \tag{2.25}$$

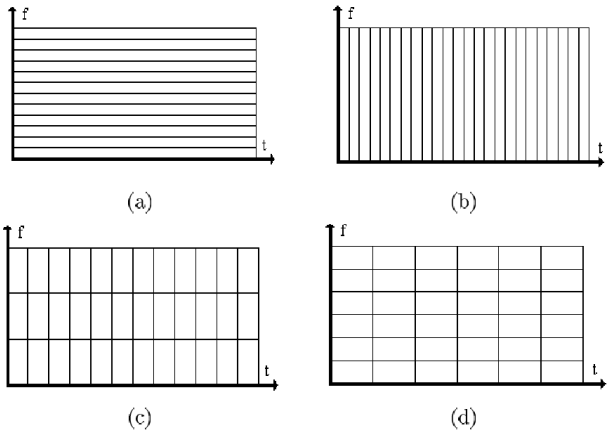


Figura 2.7: Recubrimiento del plano tiempo-frecuencia, para (a) Transformada de Fourier y series exponenciales,(b) Muestreo de Shannon, (c) Espectrograma de banda ancha y (d) Espectrograma de banda estrecha

2.3.2 Representación no estacionaria

La Transformada *Wavelet* (WT) permite la localización conjunta de eventos en tiempo-frecuencia; éste análisis incluye la técnica de ventaneo con regiones de tamaño variable (ver figura 2.8). En este caso, se usan aperturas largas de tiempo, donde se requiera información más precisa a baja frecuencia o aperturas cortas donde se requiera información de alta frecuencia. El análisis *Wavelet* es efectivo en la localización de particularidades, tales como tendencias, puntos de quiebre, discontinuidades en derivadas de alto orden, autosimilaridad, etc. [27]. En contraposición con lo que ocurre en el caso de empleo de STFT, si se desea una mejor localización de la distribución resultante en el tiempo, se escoge una ventana estrecha en tiempo, que va dividiendo el plano tiempo-frecuencia en rectángulos alargados en el sentido de la frecuencia y estrechos a lo largo del tiempo. Si por el contrario, se desea una mejor discriminación en la frecuencia, las ventanas

se rotan en su recubrimiento del plano 90º (ver figura 2.7). La uniformidad del recubrimiento, una vez elegida la ventana, lleva a difíciles compromisos de resoluciones que no siempre encuentra fácil solución. En la WT, se lleva a cabo la descomposición en diferentes componentes frecuen-

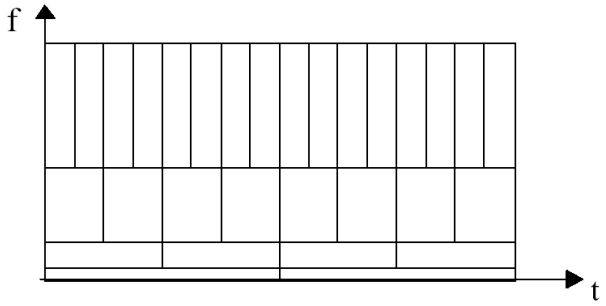


Figura 2.8: Recubrimiento del plano tiempo-frecuencia a través de la transformada Wavelet

ciales, de tal manera que cada una de las componentes tenga una resolución de acuerdo con su escala [28].

La función *Wavelet madre* $\psi(t)$ de variable real t, que oscila en el tiempo, debe estar bien localizada en el dominio temporal, donde, la localización temporal se expresa en la forma habitual de rápido decaimiento hacia cero cuando la variable independiente t tiende al infinito:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \tag{2.26}$$

$$\int_{-\infty}^{\infty} t^{m-1}\psi(t)dt = 0 \tag{2.27}$$

siendo $(m-1)$ el valor del orden del momento de la función $\psi(t)$.

A partir de la función madre, se generan el resto de funciones de la familia mediante cambios de escala y traslaciones $\{\psi_{a,b}(t),\ a > 0,\ b \in \mathbb{R}\}$. La función madre, tradicionalmente se ajusta a escala unidad. El *parámetro de escala* a queda asociado a un estiramiento o encogimiento de la función madre. Así, dada una función localizada en el tiempo $s(t)$, su versión escalada $s_a(t)$ se

define como

$$s_a(t) = \frac{1}{\sqrt{a}} s\left(\frac{t}{a}\right),\ a \in \mathbb{R},\ a > 1 \tag{2.28}$$

esta función mantiene la misma forma que $s(t)$ pero sobre un intervalo de representación (soporte) más amplio. Si el parámetro de escala se hace menor que 1, pero manteniéndolo siempre positivo (para evitar una inversión de la función) se obtiene una compresión del soporte de la función. El *parámetro de traslación* b, permite la localización temporal de la distribución de energía. A partir de la función madre $\psi(t)$, se generan las funciones Wavelet $\psi_{a,b}(t)$ mediante operaciones conjuntas de cambio de escala y traslación

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right) \tag{2.29}$$

En [28], se demuestra que si la función madre $\psi(t)$ es real, entonces la familia de funciones definidas por su traslación y escalamiento conforman una base completa del espacio, y por lo tanto, se puede representar cualquier función (señal de energía finita $f(t) \in L^2(\mathbb{R})$) mediante una combinación lineal de las funciones $\psi_{a,b}(t)$, calculando los coeficientes de tal descomposición en la forma del producto escalar. La Transformada Wavelet se describe por:

$$\begin{aligned} C(a,b) &= \int_{-\infty}^{\infty} f(t)\psi^*_{a,b}(t)dt \\ C(a,b) &= \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t)\psi^*\left(\frac{t-b}{a}\right)dt = \langle f(t), \psi_{a,b}(t)\rangle \end{aligned} \tag{2.30}$$

donde el parámetro *a* es denominado de *escala*, mientras *b* se denomina de *traslación*, ambos varían de forma continua por todo el eje real, esto es, $a,\ b \in \mathbb{R},\ a > 0$. La función $f(t)$, puede ser reconstruida unívocamente utilizando la expresión [29]

$$f(t) = \frac{1}{C_\psi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \langle f(\tau), \psi_{a,b}(\tau)\rangle \psi_{a,b}(t) \frac{dadb}{a^2} \tag{2.31}$$

donde la constante C_ψ, denominada *condición de admisibilidad*, depende sólo de la función Wavelet madre $\psi(t)$, de acuerdo con

$$C_\psi = 2\pi \int_{-\infty}^{\infty} \left|\hat{\psi}(\xi)\right|^2 |\xi|^{-1} d\xi < \infty \tag{2.32}$$

La condición de admisibilidad asegura que la función Wavelet madre no tenga contenido a frecuencia nula (o que éste resulte despreciable)(ver ecuación 2.26) y con ello, que las versiones dilatadas resultantes de la función madre estén todas centradas a frecuencias diferentes. A diferencia del caso de las expresiones de Fourier, la transformada $f(t) \rightarrow C(a, b)$ representa con redundancia una función de una variable en un espacio bidimensional y, por lo tanto, estas funciones Wavelet no forman una base ortonormal real. El muestreo apropiado de los parámetros de la función Wavelet permite eliminar la redundancia, obtener una base ortonormal de Wavelets de soporte compacto y definir la metodología para el cálculo eficiente de los coeficientes Wavelet.

Capítulo 3

Estimación de características de la voz

En la identificación automatizada de señales de voz, es común el empleo de dos tipos de parámetros de información denominados características de voz (CV): las *características acústicas* (CA) que califican las cualidades vocales y poseen un sentido físico determinado; y las *características de representación*, que corresponden a valores calculados a partir de diferentes formas de representación de la voz y, a los cuales, en general, no les corresponde algún sentido físico. En el primer caso, aunque se han llevado a cabo numerosos estudios de correlación de las CA, sus resultados, además de ser ambiguos pueden ser contradictorios en el análisis de tipicidad (normalidad) o diferentes alteraciones de la voz [30]. De este modo, la selección apropiada de la medida de las CA y su interpretación no convergen a una solución totalmente confiable en la identificación de patologías de la voz. Uno de los principales problemas en el agrupamiento de las características de voz de acuerdo con su uso principal, es que muchas de ellas son sensibles a varias propiedades acústicas. Esta mutua dependencia puede ser una de las razones para que su interpretación sea aparentemente contradictoria en los diferentes resultados encontrados en la literatura [30].

Para las características de representación, especial desarrollo tienen los coeficientes obtenidos a

partir de la Transformada Wavelet (WT) o coeficientes Wavelet (CW), que han sido probados en la extracción de parámetros orientada a la discriminación de casi todo tipo de señales, mostrando ser muy efectivos [31, 32].

3.1 Características acústicas

Las CA pueden ser agrupadas de acuerdo a las respectivas propiedades acústicas que deben medir, y las cuales en la práctica se asocian en dos categorías [19]:

3.1.1 Parámetros cuasiperiódicos

Estos parámetros reflejan las variadas formas de periodicidad presentes en las perturbaciones en la señal de voz. Entre los principales parámetros de este grupo están:

Frecuencia fundamental (F_0)

La información prosódica, es decir, la velocidad de entonación, está dominada por la frecuencia fundamental F_0 de vibración de las cuerdas vocales, cuyo inverso a su vez, se conoce como *período fundamental* T_0 [33].

En forma general, la definición de periodicidad se determina para intervalos infinitos de análisis, sin embargo, para efectos prácticos (por ejemplo de cálculo computacional), su estimación se realiza sobre intervalos finitos, que permitan cubrir varios períodos del pitch, o de manera instantánea a partir de la diferencia entre dos momentos consecutivos del cierre glótico.

La estimación del pitch es importante en diversas aplicaciones entre otras, los sistemas de decodificación acústico-fonética, sistemas de codificación para voz, a baja y media velocidades (donde la evaluación precisa del pitch lleva a la notable mejora en la calidad asociada a la voz codificada), sistemas que operan en tiempo real, tales como plataformas de ayuda a discapacitados (asistencia para aprendizaje de sordos en procesos de habla, apoyos de diagnostico post-quirúrgico, etc.), obtención de sistemas de conversión de texto a voz, basados en el contorno del pitch, entre otros [34].

En la estimación del pitch, se consideran las siguientes restricciones [34]:

- Los segmentos de voz son altamente no estacionarios, o corresponden al producto de vibraciones irregulares de las cuerdas vocales.
- La excitación glótica no es rigurosamente periódica.
- Existe una importante interacción entre la excitación y el tracto vocal, en donde la periodicidad que se observa en la señal resultante, es debida a la acción conjunta de la excitación casi-periódica y los primeros formantes de banda estrecha.
- La segmentación del inicio y del final de zonas sonoras debe realizase con alta precisión, tarea que no siempre es fácil de implementar.
- Alto margen dinámico de variación del parámetro F_0. Por tanto, dicha estimación debe restringirse a un intervalo de valores permitidos, en donde las ventanas de tiempo se ajusten dependiendo del hablante considerado, pudiendo abarcar entre 2 y más de 20ms ($50 - 500Hz$), para cubrir voces desde niños o sopranos, hasta barítonos, o entre 6 y 12 ms ($80 - 170Hz$), para el caso de locutores adultos promedio.

Los diferentes métodos de cálculo del pitch, de acuerdo a las técnicas estadísticas de estimación, pueden dividirse en:

- Estimación por promedios de ensamble, cuando en su cálculo no se considera el desarrollo de la señal en el tiempo.
- Estimación por promedios de tiempo, cuando se considera el desarrollo temporal para la señal analizada.

Así mismo, de acuerdo a la implementación de las técnicas empleadas de estimación del pitch en tiempo real, estas se dividen en:

- Análisis estacionario (en intervalos cortos de tiempo).
- Análisis no estacionario (generalmente, recurriendo a las transformadas conjuntas tiempo-frecuencia).

En la mayoría de los sistemas, la metodología utilizada en el cálculo del pitch, consta de tres partes [33]:

- El *preprocesamiento* que además de adecuar las características de señal a la entrada del sistema, busca también reducir la información redundante en la misma.
- La *extracción de parámetros* asociados a la estimación del pitch.
- El *postprocesamiento* que corrige los posibles errores del sistema de detección.

En [34], se realiza un descripción mas detallada de estos métodos.

Recientemente, la utilización de la WT [2, 35, 36] para el calculo del pitch ha mejorado los niveles de confiabilidad y precisión, gracias a la estimación del *Instante de Cierre Glótico (GCI - Glottal Closure Instant),* el cual presenta discontinuidad en su forma de onda en el momento en que la glotis se encuentra cerrada, y a partir del cual, el pitch es calculado determinando la distancia entre dos instantes consecutivos.

Muchas de las alteraciones en la voz se caracterizan por la inestabilidad en la generación del pitch durante la emisión de vocales sostenidas. Con el fin de evaluar la estabilidad de la generación de F_0 y su distorsión en la forma de onda de los pulsos, se utilizan los siguientes parámetros [9]:

1. *Grado de Ausencia de Voz (DUV - degree of unvoiceness):* que se define como el número de segmentos *Nunv* que no son sonoros sobre el número total de segmentos de voz Nt durante la emisión de vocales sostenidas:

$$k1 = Nunv/Nt \tag{3.1}$$

   Con el fin de evaluar el verdadero número de cortes de voz, se eliminan los segmentos de inicio y fin de la vocal.

2. *Grado de Estabilidad:* definido como:

$$k2 = Nstab/Np \tag{3.2}$$

   donde, *Nstab* es el número de periodos en toda la zona estable y *Np* el número de todos los periodos para la emisión vocal. La detección de las zonas estables (segmentos en donde la

generación de F_0 es casi constante) puede ser llevada a cabo usando el algoritmo descrito en [9].

Los formantes

En el espectro de una señal de voz se encuentran regiones de énfasis o *resonancias* y de deénfasis o *antiresonancias* [19], ambas resonancias, denominadas *formantes*, siguen patrones comunes en la mayoría de los humanos, por cuanto estos obedecen, en primera instancia, a sus medidas antropométricas [4], [23]. En la práctica se analizan 5 formantes (notados respectivamente como F_1, F_2, F_3, F_4, F_5), junto con sus correspondientes anchos de banda ($BW_1, \cdots, BW_5$) y energías ($EF_1, \cdots, EF_5$). Si se considera el tracto vocal como un perfecto cilindro cerrado a nivel de la glotis y abierto al nivel de los labios con una longitud promedio de 17.5 cm (media aproximada de una laringe de hombre adulto), entonces, los primeros cuatro formantes estarán cerca de los $500, 1500, 2500 y 3500 Hz$, respectivamente [2].

El análisis de los formantes y su ancho de banda da una calificación sobre la eficiencia de la articulación vocal y, mediante la búsqueda de valores óptimos se encuentran las emisiones vocales correctas [3]. La estimación de los formantes emplea algoritmos basados en modelos de representación de la voz con un número reducido de indicadores, a partir de los cuales es posible su reconstrucción adecuada. Para esto, se emplean técnicas de predicción lineal, basadas en el modelado del tracto vocal mediante un filtro de solo polos, que permite predecir la próxima muestra como una combinación lineal de las muestras anteriores [37]

3.1.2 Parámetros de perturbación

Destinados a medir la componente relativa de ruido en la señal de voz. Para la estimación de las perturbaciones de los parámetros de la frecuencia fundamental y amplitud pico, es común el empleo del promedio relativo de perturbación (*RAP Relative Average Perturbation*) definido como [4], [10]:

$$\vartheta_{RAP} = \frac{\left(\sum_{i=1}^{n} \left|\frac{z_{i-1}+z_i+z_{i+1}}{3} - z_i\right|\right)}{\sum_{i=1}^{n} z_i} \tag{3.3}$$

siendo *n* el número de ciclos consecutivos analizados.

Jitter

Cuando el parámetro de perturbación z_i en 3.3 se refiere al periodo de la frecuencia fundamental. Estas variaciones fueron observadas por primera vez en [38].

Shimmer

Cuando el parámetro de perturbación z_i en 3.3 se refiere al pico máximo de la señal pico a pico.

Relación armónico - ruido (*HNR - Harmonic Noise Ratio*)

Corresponde al promedio de la componente de ruido $\eta(t)$ de la emisión vocal, y la cual no tiene una competencia glótica adecuada [2].

Relación excitación glótica - ruido (*GNE - Glottal to Noise excitation Ratio*)

Es la estimación del ruido basada en la presunción de que los pulsos glóticos resultantes de la colisión de los pliegues vocales conllevan a una excitación sincrónica en las diferentes bandas de frecuencia [30]. El ruido turbulento generado durante la constricción conlleva a una excitación no correlacionada. El sincronismo es expresado por la correlación entre envolventes de diferentes bandas de frecuencia.

3.1.3 Estimación de las características acústicas

Estimación de la Frecuencia Fundamental

Se consideran tres diferentes algoritmos de estimación del pitch, así [39]:

- Intervalos cortos de tiempo:
  - Basado en el cálculo simple de la Función de Auto Correlación (FAC).
  - Basado en el cálculo mejorado de la Función de Auto Correlación (FAC).
- Transformada conjunto tiempo - frecuencia:
  - Basado en el cálculo de la Transformada Wavelet Discreta

Intervalos cortos de tiempo: En los métodos de análisis de intervalos cortos de tiempo, los segmentos de voz con apertura T, se procesan como si cada uno de ellos tuviese propiedades estadísticas

independientes:

$$x_w[n] = \sum_{m=-\infty}^{\infty} \Gamma\left\{x[m]\right\} w[n-m] \tag{3.4}$$

Donde $x[n]$ es la señal digitalizada de voz, $w[n]$ es la función ventana centrada respecto a la apertura T, que es escogida, de tal manera, que facilite la extracción de las CA. $\Gamma(\cdot)$ es la transformación a la cual se somete la secuencia aleatoria original.

***Algoritmo simple de FAC*:** En este caso, el algoritmo es denominado *Short Time Cross Correlation Function* (STCCF), el cual toma en (3.4) la función rectangular de ventaneo, $w[n] = 1$, por lo que la FAC se determina en forma vectorial como [40]:

$$R_x(\tau) = \frac{\vec{x}_i \vec{x}_{i+\tau}}{|\vec{x}_i|\,|\vec{x}_{i+\tau}|} \tag{3.5}$$

donde la estimación T_0 del periodo fundamental se define como el argumento para el cual el STCCF toma su máximo valor, así:

$$T_0 = \arg\left\{\max(\tau)\right\}, \forall T_{0\,\min} \leq \tau \leq \forall T_{0\,\max} \tag{3.6}$$

La estimación del pitch al realizarse sobre segmentos cortos de voz, supone previamente la toma de la decisión en cada intervalo de análisis sobre las siguientes hipótesis: la presencia de sonidos (λ=1) o su ausencia (λ=0). Como umbral de detección γ puede ser empleado el primer coeficiente de reflexión derivado de (3.7) [41]:

$$r_{1x}(\tau) = \frac{\vec{x}_i \vec{x}_{i+1}}{|\vec{x}_i|\,|\vec{x}_i|} = \begin{cases} \leq \gamma, & \lambda = 0 \\ > \gamma, & \lambda = 1 \end{cases} \tag{3.7}$$

Algoritmo mejorado de FAC En general, la estimación del pitch basada en FAC está afectada por la función de autocorrelación de la forma de la ventana empleada [42]:

$$R_x(\tau) \approx \frac{R_a(\tau)}{R_w(\tau)} \tag{3.8}$$

Así, para un tren de funciones delta con periodo T_0, definidos como

$$x(k) = \sum_{n=-\infty}^{\infty} \delta(k - k_0 - nT_0), 0 \leq k_0 \leq T_0 \tag{3.9}$$

siendo k_0 la fase de los pulsos con la ventana, se tiene que (3.8) será

$$R_x(T_0) = \frac{\sum\limits_n w(k_0 + nT_0)w(k_0 + (n+1)T_0)}{R_w(T_0)\sum\limits_n w^2(k_0 + nT_0)} \tag{3.10}$$

Esto es, la exactitud del algoritmo depende directamente de la forma de la ventana empleada. Se considera que en aplicaciones de voz, la ventana de Hanning es preferible a otras formas por su menor sensibilidad a cambios rápidos de la señal [42]. Un factor adicional de distorsión en el cálculo de la función de correlación es la discretización del proceso original con periodo de muestreo $\triangle\tau$, cuya FAC también será discreta:

$$R[n] \equiv R(n\Delta\tau) \tag{3.11}$$

En el caso de encontrar un máximo local entre los intervalos de correlación $(m-1)\triangle\tau$ y $(m+1)\triangle\tau$, o sea, $r[m-1] < r[m] < r[m+1]$, entonces, una primera estimación del pitch podría ser $\tau_{max} = m\triangle\tau$, con lo cual se obtendría una resolución igual a $\triangle f = f_s\triangle\tau/T_0$. Teniendo en cuenta los valores reales del pitch ($\leq 300Hz$) y de los estándares existentes en las velocidades de discretización de señales de voz (≈10.000Hz), el valor de resolución estará alrededor de los 9 Hz,

obteniéndose una precisión límite en la estimación del orden del 3%. En [42], se proponen diferentes tipos de interpolación alrededor del punto $m\triangle\tau$, para mejorar la calidad de la estimación.

Estimación de los formantes

El ancho de banda de los formantes BW_i se define como el grupo de frecuencias que hay desde la caída posterior y anterior de 3 dB. La energía del formante será el valor del pico y el valor del formante la frecuencia donde se encuentre el pico anterior [2]. Para predecir la próxima muestra como una combinación lineal de las muestras anteriores se tiene que:

$$\hat{y}(n) = -\sum_{k=1}^{q} a_k y(n-k) \tag{3.12}$$

donde q es el orden del predictor y a_k son los coeficientes de predicción lineal (LPC) que pueden ser calculados de dos maneras:

- Calculando las raíces del filtro que representan los LPC, lo cual es computacionalmente pesado y, por tanto dificulta el análisis en tiempo real.

- Calculando la envolvente del espectro a partir de los LPC y, luego estimar sus picos. Los coeficientes a_k pueden ser determinados a partir del cálculo de coeficientes de autocorrelación.

De esta manera, se tendrán q coeficientes de predicción lineal por cada trama de la señal de voz. También es habitual aplicar la transformación homomórfica de los coeficientes LPC, obteniendo como resultado los coeficientes LPC cepstrales (la transformada de Fourier inversa del espectro de amplitud logarítmico) que representan la respuesta aproximadamente logarítmica o *psofométrica* del oído humano. Los coeficientes LPC cepstrales c_k presentan la ventaja de convertir el ruido convolucional en ruido aditivo y permiten separar la excitación glótica de los parámetros del tracto

vocal. Seguidamente, se emplea el tipo de ventaneo que acentúe los pesos de los LPC cepstrales con el objeto de obtener mayor discriminación de la envolvente espectral.

Estimación del GNE

Fundamentalmente, el algoritmo de cálculo del GNE consta de las siguientes etapas:

1. Filtración predictiva lineal inversa para obtener los pulsos glóticos (si los hay).

2. Filtración pasa banda de la señal residual empleado la ventana de Hanning (con ancho de 3000Hz) centrada en las diferentes bandas de frecuencia.

3. Cálculo de la envolvente de Hilbert para los diferentes bandas de frecuencia (con ancho de banda y frecuencias centrales fijas), así como de sus respectivos coeficientes de correlación para intervalos de análisis (retardos) en el rango de $-0.3 < t < 0.3$ ms.

4. Finalmente, el máximo coeficiente de correlación corresponde al GNE.

3.2 Características de voz usando WT

Las técnicas de estimación CV planteadas en [31, 32], proponen modelos adaptativos para la selección de un número reducido de coeficientes Wavelet que retengan la información discriminante, teniendo como criterio de optimización el error de representación. Sin embargo, estas técnicas asumen que cada una de las muestras del conjunto inicial tengan igual tamaño, es decir, para el caso de cada uno de los segmentos de voz determinados implicaría la existencia de el alineamiento temporal, que puede ser realizado a través del *Dynamic Time Warping* (DTW), no obstante

este procedimiento requiere de la supresión de información en las señales que estén muy alejadas (por defecto o por exceso) de la longitud de referencia escogida. La metodología propuesta en [43, 44, 45], no toma en consideración el alineamiento temporal de las señales discretizadas. Además, se sugiere la forma de selección de los coeficientes de descomposición de la WT que son independientes de la longitud en muestras de cada señal.

3.2.1 Estimación de características de representación usando WT

Existen variados algoritmos de selección de los coeficientes y niveles de representación de las WT en la representación de las señales de voz. En [46] se propone un algoritmo que corresponde a la modificación del presentado en [44]. Gracias a su propiedad de concentración de energía, la WT distribuye las características específicas de la señal en diferentes bandas de frecuencia. Haciendo uso de esta propiedad, el algoritmo de estimación de CV se describe de la siguiente manera:

1. Selección de la Wavelet ortogonal de soporte compacto $\psi(t)$

2. Cálculo del respectivo filtro pasabajo de descomposición $h[m]$ [47].

3. Determinación de los valores de $f_i[n]$, $i = 1, .., N_s$ (número de muestras), como una secuencia discretizada para el segmento de voz.

4. Cálculo de los coeficientes $ca_j, j = 1, ..., J$ por la expresión (A.3), sobre J escalas de aproximación, que cubren J octavas de frecuencia. Donde $ca_0 = f_i[n]$ corresponde al nivel de aproximación cero.

5. Selección de los p coeficientes de mayor amplitud en cada escala de aproximación ca_j.

6. Conformación del vector de características $\mathbf{x}$ para $f_i[n]$, con dimensionalidad $N_f = J \times p$.

3.2.2 Selección de la Wavelet Madre

Variadas aplicaciones *wavelet* explotan la capacidad de representar eficientemente una clase particular de funciones con un número pequeño de coeficientes *wavelet* diferentes de cero.

La selección de ψ debe ser tal, que produzca el máximo número de coeficientes *Wavelet*, cuyo valor sea cercano a cero. De otra manera, la señal $f(t)$ tiene pocos coeficientes *Wavelet* de representación no despreciables, si la mayoría de los coeficientes a escalas finas tienden a cero, lo cual depende, en mayor medida, de la regularidad de la función f, de los momentos de desvanecimiento de ψ y del tamaño del respectivo soporte compacto.

En [48] se prueba que si f cumple con la condición de regularidad y ψ tiene una cantidad suficiente de momentos de desvanecimiento, entonces los coeficientes *wavelet* $\mid \langle f, \psi_{j,n} \rangle \mid$ serán pequeños para las escalas finas de 2^j. Donde ψ tiene p momentos de desvanecimiento, si se cumple que

$$\int_{-\infty}^{\infty} t^k \psi(t) dt = 0 \quad para \quad 0 \leq k \leq p$$

Si f tiene una singularidad aislada en t_0 y si t_0 se encuentra dentro del soporte de $\psi_{j,n}(t) = \frac{1}{\sqrt{2}}\psi(\frac{t-2^j n}{2^j})$, entonces $\langle f, \psi_{j,n} \rangle$ podrá presentar una amplitud significativa. Si ψ tiene un soporte compacto de tamaño K, en cada escala 2^j existen K *wavelets* $\psi_{j,n}$ cuyo soporte incluye a t_0. Para minimizar el número de coeficientes de amplitud alta se debe reducir el tamaño del soporte de ψ. Se dice que la función de escalamiento ϕ tiene soporte compacto si y solo si h tiene soporte compacto y sus soportes son iguales. Si los soportes de h y ϕ son iguales a $[N_1, N_2]$ entonces el soporte de ψ es $[\frac{N_1 - N_2 + 1}{2}, \frac{N_2 - N_1 + 1}{2}]$.

Si f tiene pocas singularidades aisladas y es muy regular entre singularidades, se debe buscar la *wavelet* que presente alta número de momentos de desvanecimiento para generar una gran cantidad de coeficientes $\langle f, \psi_j, n \rangle$ de valor cercano a cero. Si la densidad de singularidades se incrementa, sería recomendable reducir el tamaño del soporte, y su costo sería la reducción de los momentos de desvanecimiento.

Capítulo 4

Selección de características de voz

La selección de las CV es, tal vez, la etapa más importante en el desarrollo de sistemas automatizados de proceso y clasificación de la voz. Todo parámetro de análisis deberá cumplir con las siguientes propiedades [49]:

- Debe ser fácilmente medible y poco dependiente de las perturbaciones ambientales (ruidos, interferencias, etc.).
- Su estimación debe ser estable en el tiempo.
- No debe ser imitable.

El análisis de datos en altas dimensiones se ha convertido en un problema común, el cual requiere de altos recursos computacionales, presenta dificultad de representación, adicionalmente, el almacenamiento, transmisión y procesamiento de estos datos demanda grandes sistemas. Por tanto, es favorable reducir la dimensionalidad de los datos, mientras se mantenga la estructura original de los mismos casi intacta. [50] Además, existen otras dos razones principales, por las cuales se debe mantener la dimensionalidad del espacio de características tan pequeño como sea posible: costo

de medida y precisión en la clasificación. Un número limitado de características simplifica la representación tanto del patrón como del clasificador, lo que resulta en un clasificador más rápido y que usa menos memoria. Por otro lado, la reducción exagerada en el número de características podría llevar a una pérdida en el poder discriminante, empobreciendo la precisión del sistema de reconocimiento. [51]

4.1 Preproceso de características

La estimación de cualquier CV, ζ_i, es muy sensible a factores tales como: las condiciones acústicas de toma de señales (ruido de fondo, hardware de registro electrónico, tiempo del día en que se le toman las muestras al paciente, contenido de las palabras, etc.), por lo que la primera tarea a resolver en su clasificación, usualmente, es el preproceso de datos de las realizaciones, que está orientado a aumentar la efectividad en el uso de los parámetros representativos de cada clase. Una parte básica en el preproceso de datos está en la eliminación de datos anómalos debidos a posibles errores en el proceso de medición, influencia de fuertes perturbaciones, etc. En este caso, cualquier valor anómalo del vector $\zeta_{ik}(l)$ se expresa mediante el valor crítico de la distribución de Student $t_{p,n-2}$ [52]:

$$\frac{|\zeta_{ik}(l) - m_{1\zeta_{ik}}|}{\sigma_{\zeta_i}} \leq \frac{t_{p,N_e-2}\,(N_e-1)^{1/2}}{\left((N_e-2)+(t_{p,N_e-2})^2\right)^{1/2}}, \tag{4.1}$$

siendo $m_{1\zeta_{ik}}$ y $\sigma_{\zeta_{ik}}$, respectivamente los valores extremos de la media, y varianza del arreglo en análisis ζ_{ik}; p es el nivel de significancia. Este método permite la agrupación de valores cada arreglo $\zeta(l)$ en tres grupos:

$$|\zeta(l) - m_{1\zeta}| \leq (5\%, N_e)\,; \tag{4.2}$$

$$(5\%, N_e) \leq |\zeta(l) - m_{1\zeta}| \leq (10\%, N_e)\, y \tag{4.3}$$

$$|\zeta(l) - m_{1\zeta}| \geq (10\%, N_e) \tag{4.4}$$

Los valores de $\zeta(l)$ relativos al primer grupo no deben ser eliminados, los del segundo necesitan un elemento de juicio adicional para ser eliminados, mientras los del tercer grupo siempre se eliminan. La detección de valores anómalos por toda la matriz $\zeta_{ik(l)}$, adicionalmente permite hacer clara la calidad de medición de cada realización de CV. Así, si en la realización del análisis acústico de un paciente dado ocurre un error sistemático de medición es de esperar que aparezcan como anómalos varios de los valores de la matriz $\zeta_{ik}(l)$ por i. En este caso, es preferible que estas muestras de pacientes sean eliminadas del ensamble y reemplazadas por otros registros. Así mismo, se puede juzgar indirectamente sobre la estabilidad de la estimación de cada CV $\zeta(l)$, en la medida en que se incremente desmesuradamente el número de valores anómalos en la matriz inicial por todos los l, para valores fijos de i, k, con mayor certeza se deducirá que hay problemas en la estimación de ζ_{ik} [2].

4.2 Reducción de dimensionalidad

4.2.1 Pruebas de independencia estadística

Con el propósito de observar y evaluar las CV que discriminen adecuadamente las dos clases de voz se realizan los siguientes procedimientos:

- *Pruebas de Hipótesis*, para comparar las clases desde el punto de vista de los promedios de cada una de las componentes del espacio de características. $\mathcal{F}$.
- *Correlación por rangos de Spearman*, para detectar algún grado de dependencia o independencia de parejas de variables.
- *Análisis de componentes principales*, con el propósito de reducir la dimensión del espacio característico, además para detectar dependencia o independencia en dicho espacio.

Prueba de hipótesis

Con el fin de asegurar que cada una de las CV sea apta para lograr el grado suficiente de discriminación entre las voces patológicas y las normales, se analizan las siguientes hipótesis

- H_0: no existe diferencia significativa en la media de cada CV $\xi_i \in \mathcal{F}$ para discriminar las dos clases.
- H_1: existe una diferencia significativa del promedio en cada $\xi_i \in \mathcal{F}$.

En este trabajo, la prueba de las anteriores hipótesis se realizó empleando el método *t-Student*. Sea $\xi_i = [x_1, x_2, ..., x_{N_i}]$ el vector correspondiente a las mediciones de la CV, con media μ y varianza σ^2, ambas desconocidas. A partir de las N_i observaciones por clase, se estiman los valores de $\overline{\mu}$ y $\overline{\sigma^2}$. Así, un intervalo de confianza bilateral al $100(1-\alpha)\%$ para la media verdadera es

$$(\bar{\mu_1} - \bar{\mu_2}) - t_{(\alpha/2, N_1+N_2-2)}\bar{\sigma}_{\bar{\mu_1}-\bar{\mu_2}} \leq \mu_1 - \mu_2 \leq (\bar{\mu_1} - \bar{\mu_2}) + \\ + t_{(\alpha/2, N_1+N_2-2)}\bar{\sigma}_{\bar{\mu_1}-\bar{\mu_2}} \tag{4.5}$$

con

$$\bar{\sigma}^2_{\bar{\mu_1}-\bar{\mu_2}} = \frac{N_1\bar{\sigma}_1^2 + N_2\bar{\sigma}_2^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right) \tag{4.6}$$

donde $t_{(\alpha/2,N_1+N_2-2)}$ representa el punto porcentual de la distribución t con $N_1 + N_2 - 2$ grados de libertad. Si el intervalo de confianza dado por la ecuación (4.5) contiene el valor 0, entonces no se rechaza la hipótesis nula (H_0), en caso contrario, se acepta la hipótesis alternativa H_1. Así, si la hipótesis nula es rechazada, se asume que existe diferencia entre la media de cada clase.

Análisis de correlación por rangos

Como medida de asociación entre las CV, el análisis no paramétrico de correlación por rangos es utilizado para observar su mutua dependencia. Una de estas medidas de asociación es el *coeficiente de rango de Spearman* [53]. Los valores de los coeficientes de correlación por rangos están entre -1 y 1, donde un valor cercano a cero indica que no existe una asociación entre las variables. El coeficiente de correlación de Spearman r_s es definido como el coeficiente de correlación lineal entre los rangos R_i de ξ_i y los rangos S_i de χ_i con $\xi_i,\ \chi_i \in \mathcal{F}$

$$r_s = \frac{\sum_{i=1}^{N_s}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{N_s}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{N_s}(S_i - \overline{S})}} \tag{4.7}$$

Aunque se pierde alguna información al reemplazar los datos originales por sus rangos, el coeficiente de Spearman es más robusto a la presencia de anomalías en los datos que la correlación lineal, debido a que pequeñas variaciones no influyen en el rango de los datos. La transformación del espacio original al espacio de rangos genera la linealización entre las CV [53].

4.2.2 Análisis de componentes principales PCA

Es tal vez uno de los métodos de reducción de dimensión más utilizado, en gran parte debido a su simplicidad conceptual y a la eficiencia computacional de sus algoritmos. También conocido

como la transformación de *Karhunen-Loève* [54], siendo una poderosa herramienta que permite extraer la estructura de un conjunto de datos de alta dimensionalidad.

PCA lineal

Inicialmente su campo de acción se centra en las transformaciones lineales de los datos, y busca maximizar la varianza direccional de una manera no correlacionada, ortogonalizando el sistema de coordenadas en el que se describen originalmente los datos; de esta manera el problema de reducción de dimensionalidad tiene un solución analítica exacta. Geométricamente, el hiperplano generado por los primeros **L** componentes Principales es el hiperplano de regresión que minimiza las distancias ortogonales a los datos. Por esta razón, PCA es un método de regresión simétrica, contrario a la regresión lineal estándar. Es común que un número pequeño de componentes principales sea suficiente para representar la mayor parte de la estructura de los datos. Estos primeros componentes principales generalmente son usados como punto de partida por otros algoritmos, tales como *Regresión Projection Pursuit, Curvas Principales* o *Mapas de Kohonen* entre otros. De igual manera, existen varias arquitecturas de *Redes Neuronales* con capacidad de extraer los componentes principales de un conjunto de datos [54].

Sin embargo, el PCA lineal solo está en capacidad de encontrar un subespacio lineal, lo que indica que no puede manejar datos con relaciones no lineales. En general, no se sabe cuantos componentes principales se deben tener en cuenta, aunque existen algunas reglas empíricas para decidir. Por ejemplo, eliminar aquellos componentes cuyos autovalores sean menores a cierta fracción del mayor de los autovalores, o tener en cuenta los necesarios para representar cierto porcentaje de la varianza total. [54].

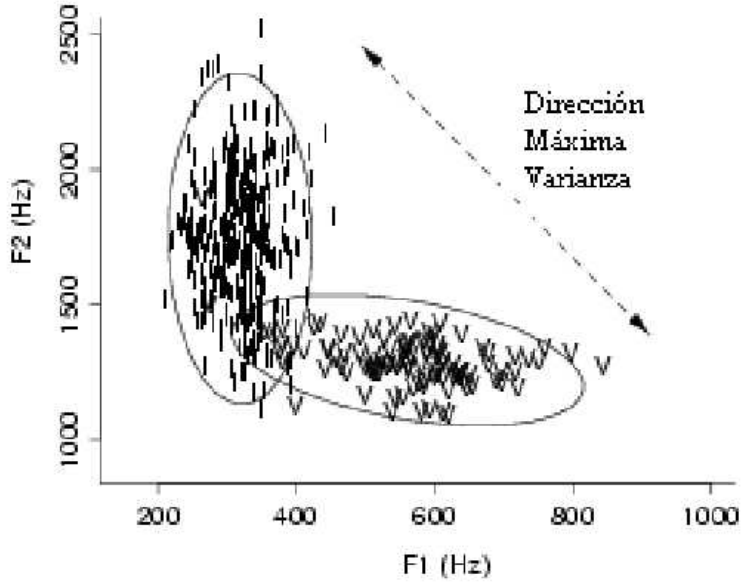


Figura 4.1: PCA representa la dirección de la máxima varianza

El algoritmo computacional de cálculo de PCA lineal empleado corresponde al *método matricial*, cuyo objetivo es encontrar los autovectores de la matriz de covarianza. Dichos autovectores describen la dirección de los componentes principales de los datos originales, y su significancia estadística está dada por el autovalor correspondiente. Este método puede estructurarse de la siguiente manera [54]:

1. Conformación de la matriz inicial de datos **X**, a partir de vectores muestra x_i, i=1,2,...,m.

2. Centralización del primer momento, calcular $\bar{x}$ y restarlo de cada x_i.

3. Cálculo de la matriz de covarianza **C**.

4. Determinación de los autovectores y autovalores para la matriz **C**.

$$C\nu = \lambda\nu \tag{4.8}$$

en donde λ es un autovalor y ν un autovector.

5. Organización de los autovalores de tal manera que:

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n \tag{4.9}$$

6. Selección de los primeros d $\leqslant$ n autovalores y generar el nuevo set de datos en la nueva representación.

PCA no lineal

PCA es una técnica utilizada para transformar linealmente un conjunto de datos en un nuevo conjunto de variables no correlacionadas de dimensión pequeña que representa la mayor parte de la información. A través del uso de funciones Kernel, dichas variables se puede transformar de una manera no lineal. El Kernel PCA es una extensión no lineal de PCA en donde los componentes principales son calculados en un espacio de características de altas dimensiones, el cual está relacionado de una manera no lineal con el espacio de entrada. [55]

Dado un conjunto centrado de observaciones x_k, $k = 1, \ldots, M$, y debido a que todas las soluciones ν con $\lambda \neq 0$ en (4.8) están contenidas en el espacio generado por x_1,...,x_M, la ecuación (4.8) es equivalente a

$$\lambda(x_k \cdot \nu) = (x_k \cdot C\nu) \tag{4.10}$$

Para el caso de KPCA, también es necesario la solución de la ecuación (4.10), pero en un espacio producto punto F, el cual se relaciona con el espacio de entrada a través de un mapeo no lineal,

$$\Phi : R^N \rightarrow F, x \rightarrow X \tag{4.11}$$

Cabe anotar que dicho espacio F, al cual puede referirse como espacio de características, puede tener una dimensionalidad muy alta, inclusive infinita. La matriz de covarianza en F es

$$\bar{C} = \frac{1}{M} \sum_{j=1}^{M} \Phi\left(x_j\right) \Phi\left(x_j\right)^T \tag{4.12}$$

entonces, se necesita resolver

$$\lambda V = \bar{C} V \tag{4.13}$$

encontrando los autovalores $\lambda \geq 0$ y los autovectores $V \in F$. Aquí nuevamente, todas las soluciones V con $\lambda \neq 0$ están contenidas en el espacio generado por $\Phi(x_1)$,...,$\Phi(x_M)$, lo que permite concluir los siguientes momentos [55]: primero, se puede considerar que

$$\lambda(\Phi\left(x_k\right) \cdot V) = (\Phi\left(x_k\right) \cdot CV) \tag{4.14}$$

para $k = 1, \ldots, M$, y segundo, existen los coeficientes $\alpha_i(i)(i = 1, \ldots, M)$, tales que

$$V = \sum_{i=1}^{M} \alpha_i \Phi\left(x_i\right). \tag{4.15}$$

Si se combinan (4.14) y (4.15), se obtiene

$$\lambda \sum_{i=1}^{M} \alpha_i \left(\Phi\left(x_k\right) \cdot \Phi\left(x_i\right)\right) = \frac{1}{M} \sum_{i=1}^{M} \alpha_i \left(\Phi\left(x_k\right) \cdot \sum_{j=1}^{M} \Phi\left(x_j\right)\right) \left(\Phi\left(x_j\right) \cdot \Phi\left(x_i\right)\right) \tag{4.16}$$

Se define la matriz **K** de dimensiones MxM, como

$$K_{ij} = (\Phi(x_i) \cdot \Phi(x_j)) \tag{4.17}$$

lo que significa que

$$M\lambda K\alpha = K^2\alpha \tag{4.18}$$

en donde α denota el vector columna con entradas α_1,...,α_M. Para encontrar las soluciones de (4.18), se resuelve

$$M\lambda\alpha = K\alpha \tag{4.19}$$

para los autovalores diferentes de cero.

Sean $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_M$, los autovalores de K (las soluciones $M\lambda$ de (4.18)), y $\alpha^1, \ldots, \alpha^M$ el conjunto completo de los autovectores correspondientes, siendo λ_p el ultimo autovalor diferente de cero (asumiendo $\Phi \neq 0$) se normalizan $\alpha^1, \ldots, \alpha^p$ con el fin de que los correspondientes vectores en F sean normalizados, es decir

$$\left(V^k \cdot V^k\right) = 1 \tag{4.20}$$

para $k = 1, \ldots, p$, combinando (4.15) y (4.19), lo anterior se convierte en la condición de normal-

ización para $\alpha^1, \ldots, \alpha^p$:

$$\begin{aligned} 1 &= \sum_{i,j=1}^{M} \alpha_i^k \alpha_j^k \left(\Phi\left(x_i\right) \cdot \Phi\left(x_j\right)\right) \\ &= \sum_{i,j=1}^{M} \alpha_i^k \alpha_j^k K_{ij} = \left(\alpha^k \cdot K\alpha^k\right) = \lambda_k(\alpha^k \cdot \alpha^k) \end{aligned} \tag{4.21}$$

Para el cálculo de los componentes principales, es necesario realizar las proyecciones sobre los autovectores V^k en F ($k = 1, \ldots, p$). Sea x un punto de prueba, con una imagen $\Phi(x)$ en F, entonces

$$\left(V^k \cdot \Phi\left(x\right)\right) = \sum_{i=1}^{M} \alpha_i^k \left(\Phi\left(x_i\right) \cdot \Phi\left(x\right)\right) \tag{4.22}$$

puede considerarse como el Kernel PCA correspondiente a Φ [55].

Ejemplos de funciones kernel

La matriz de productos punto **K** puede ser calculada seleccionando un kernel $k(x, y)$ de tal manera que $k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) = K_{ij}$, con el fin de evitar cualquier cálculo en el espacio de características con altas dimensiones. [55]

Entre los kernel comúnmente utilizados se tienen los siguientes:

– polinomiales,

$$k\left(x, y\right) = \left(x \cdot y + 1\right)^d \tag{4.23}$$

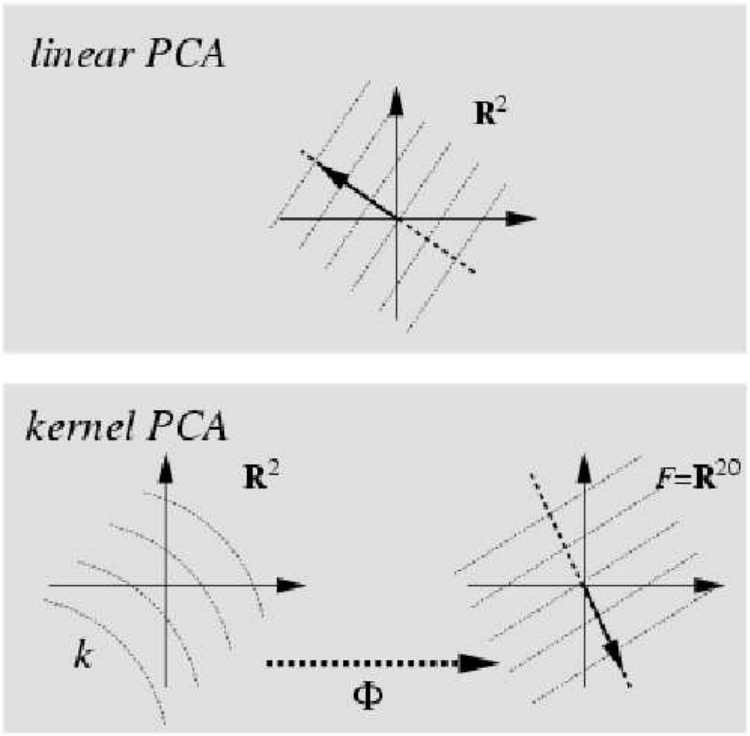


Figura 4.2: KPCA realiza PCA en un espacio de altas dimensiones

– las funciones base radiales (RBF),

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \tag{4.24}$$

– tipo Red Neuronal.

$$k(x,y) = \tanh\left((x \cdot y) + b\right) \tag{4.25}$$

4.2.3 Análisis discriminante

El objetivo del Análisis Discriminante puede resumirse como, encontrar una función que retorne valores escalares que permitan una buena discriminación entre diferentes clases de los datos de entrada. Estos discriminates son usados luego para entrenar clasificadores o para visualizar ciertos aspectos de los datos. En este sentido, el Análisis Discriminante puede entenderse como un preprocesamiento supervisado o extracción de características, supervisado en el sentido de que es necesario informarle al algoritmo que muestras de entrenamiento corresponden con que clase [56].

Análisis discriminante lineal

El Análisis Discriminante Lineal (*LDA - Linear Discriminant Analysis*), busca aquellos vectores que mejor discriminan las clases en un espacio inicial (en lugar de aquellos que mejor representan los datos). De una manera mas formal, dado un número de características independientes, el LDA crea una combinación lineal de estas, que permita la mayor diferencia entre las medias de las clase deseadas. Matemáticamente, se definen dos medidas para todas las muestras en todas las clases: (i) una llamada *Matriz de dispersión intra-clase*

$$S_w = \sum_{j=1}^{c} \sum_{i=1}^{N_j} \left(x_i^j - \mu_j\right) \left(x_i^j - \mu_j\right)^T \tag{4.26}$$

donde x_i^j es la *i*-ésima muestra de la clase *j*, μ_j es la media de la clase *j*, *c* es el número de clases , y N_j es el número de muestras en la clase *j*, y (ii) la otra es llamada *Matriz de dispersión entre-clases*

$$S_b = \sum_{j=1}^{c} \left(\mu_j - \mu\right) \left(\mu_j - \mu\right)^T \tag{4.27}$$

en donde μ representa la media de todas las clases.

EL objetivo es maximizar el valor entre-clases, mientras se minimiza el valor intra-clase. Una forma de hacerlo es maximizando la relación $\frac{\det|S_b|}{\det|S_w|}$. La ventaja de utilizar esta relación es, que se ha comprobado que si S_w es una matriz no singular, dicha relación se maximiza cuando lo vectores columna de la matriz de proyección W, son los autovectores de $S_w^{-1}S_b$.

Básicamente, se busca una función $f : \mathcal{X} \rightarrow \mathbb{R}^D$, tal que $f(x)$ y $f(z)$ son similares cuando x y z lo son, y diferentes en otro caso. En el caso especial del Análisis Discriminate Lineal se busca una función lineal

$$f(\mathbf{x}) = W^T\mathbf{x}, \qquad W \in \mathbb{R}^{NxD} \tag{4.28}$$

en donde W es la matriz de transformación.

Discriminante de fisher

Probablemente el ejemplo más conocido de un discriminante lineal corresponde al Discriminante de Fisher, que busca la dirección $\mathbf{w}$ que separa correctamente las medias de las clases (una vez se proyectan en la dirección correcta) mientras se logra una varianza pequeña alrededor de las mismas. Se espera que sea simple decidirse, con un error pequeño, por una de las clases a partir de dicha proyección. La cantidad que mide la diferencia entre las medias es llamada *Varianza entre-clases* y la cantidad que mide la varianza alrededor de la media de cada clase es llamada *Varianza intra-clase* respectivamente. Por tanto, el objetivo es encontrar una dirección que maximize la *varianza entre-clases* al mismo tiempo que minimiza la *varianza intra-clase*.

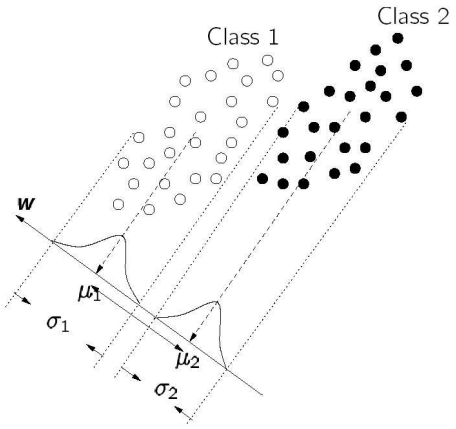


Figura 4.3: Discriminante de Fisher para dos clases. Se busca la dirección w, para la cual la diferencia entre las medias de clases (μ_1y μ_2) proyectadas en dicha dirección es grande y que la varianza alrededor de dichas medias (σ_1 y σ_2) es pequeña.

4.2.4 Análisis de información mutua

Considerese dos variable aleatorias discretas **X** y **Y** con valores en dos alfabetos [57]

$$\chi = \{x_k | k = 0, \pm 1, \ldots, \pm K\}\, y \tag{4.29}$$

$$\gamma = \{x_j | j = 0, \pm 1, \ldots, \pm J\} \tag{4.30}$$

donde x_k y y_j son valores discretos y, $(2K + 1)$ y $(2J + 1)$ son los números de niveles discretos respectivamente. Se define

$$p(x_k) = P(X = x_k), p(y_j) = P(Y = y_j) \tag{4.31}$$

como las probabilidades de que las variables **X** y **Y** tomen los valores x_k y y_j. Entonces la *Información Mutua* $I(X,Y)$ entre las variables **X** y **Y** es

$$I(X,Y) = \sum_{k=-K}^{+K} \sum_{j=-J}^{+J} p(x_k, y_j) \log \left(\frac{p(x_k, y_j)}{p(x_k)p(y_j)} \right) \tag{4.32}$$

donde $p(x_k, y_j)$ es la función de probabilidad conjunta de las variables discretas **X** y **Y**. Se considera a $I(X,Y)$ como la información relativa a la variable X que puede ser obtenida observando la variable Y.

La información mutua tiene las siguientes propiedades:

- *Simetría* de la información mutua entre **X** y **Y**, $I(X,Y) = I(Y,X)$
- *No negativa* $I(X,Y) \geqslant 0$
- La información mutua puede ser expresada en términos de la *Entropía* de las dos variables

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{4.33}$$

donde

$$H(X) = \sum_{k=-K}^{+K} p(x_k) \log \left(\frac{1}{p(x_k)} \right) \tag{4.34}$$

es la *Entropía,* que representa la cantidad promedio de información que ofrece la variable **X**,

y

$$H(X|Y) = H(X,Y) - H(Y) \tag{4.35}$$

es la *Entropía Condicional* de **X** dado **Y**, y representa la cantidad de incertidumbre existente sobre la variable **X** después de observar la variable **Y**, y donde

$$H(X,Y) = \sum_{k=-K}^{+K} \sum_{j=-J}^{+J} p(x_k, y_j) \log \left(\frac{1}{p(x_k)p(y_j)} \right) \tag{4.36}$$

es la *Entropía Conjunta* de **X** y **Y**, y representa la cantidad promedio conjunta de información dada por las variables **X** y **Y**.

Selección de CV empleando la información mutua

La información mutua puede ser determinada como el criterio de selección de las CV, así [58]:

$$f\left(X\right) = \max_{f \in F} I\left(Y, X\right) \tag{4.37}$$

el objetivo es entonces, seleccionar el conjunto de CV que tienen la mayor información mutua con la respectiva clase, al mismo tiempo que se obtiene la menor información mutua entre dichas CV.

$$f = \max_{X_i} I\left(Y, X_i\right) - \beta \sum_{i,j=1} I(X_i, X_j) \tag{4.38}$$

Capítulo 5

Marco experimental

En esta sección se muestran las pruebas y resultados obtenidos para la metodología propuesta de selección de características de voz orientada a la identificación de patologías. Se determinó que la población para la cual se llevaría a cabo el estudio sería la población urbana adulta de la Ciudad de Manizales, personas de ambos géneros y que pertenecen al periodo de *estabilidad vocal*, es decir, con edades entre los 19 y 54 años de edad. Las clases dentro de las cuales fueron ubicadas las voces de los pacientes evaluados subjetivamente fueron: clase Normal ($k = 1$) y clase con Disfonía($k = 2$).

5.1 Base de datos fuente

La muestra representativa de la población seleccionada, fue evaluada de forma subjetiva por parte del especialista en fonoaduología, y una vez realizado el diagnóstico inicial se llevaron a cabo las respectivas sesiones de grabación con aquellas personas que amablemente colaboraron de manera consciente con la investigación. El procedimiento de grabación fue realizado en condiciones controladas de ruido ambiente.

5.1.1 Recolección de señales de voz

La muestra de análisis en el presente trabajo consta de un total de 91 registros de voz, clasificados así:

Género	*Valoración*
42 Hombres	40 Normales
49 Mujeres	51 Disfonías

Tabla 5.1: Muestra de análisis y valoración del especialista

Las propiedades de los archivos de audio generados para cada una de las señales son las siguientes:

- Formato de registro tipo *.wav.
- Frecuencia de muestreo igual a $22kHz$.
- Bits por muestra igual a 16.
- Canales de grabación - 1 (monofónico).

5.1.2 Conjunto de datos

El registro de voz consta de la pronunciación en una forma natural de las cinco vocales del idioma castellano, */a/*, */e/*, */i/*, */o/*, */u/*.

Para el presente trabajo, las pruebas fueron realizadas sobre los siguientes conjuntos de datos, según la forma que se considera cada realización:

- Cada realización corresponde al segmento /a/.
- Cada realización corresponde a un vector conformado por las características de los 5 segmentos {/a/, /e/, /i/, /o/, /u/}.

- Cada realización corresponde a uno de los segmentos [/a/; /e/; /i/; /o/; /u/], es decir, cada segmento es tomado como una realización independiente.

5.2 Conformación de los espacios de características

Teniendo en cuenta lo expuesto en el capítulo 3, así como las naturaleza específica de los algoritmos de cálculo de las CV, en el presente trabajo se restringe el análisis sobre tres diferentes espacios, así:

Espacio de Características Acústicas (CA1): Consta de las siguientes CA estimadas de forma directa sobre el registro de voz:

- Frecuencia Fundamental (algoritmo AMDF)
- Jitter
- Shimmer
- Armónico Ruido
- Formante 1
- Ancho de Banda del Formante 1
- Energía del Formante 1
- Formante 2
- Ancho de Banda del Formante 2
- Energía del Formante 2
- Energía del segmento

Espacio derivado de Características Acústicas (CA2): Consta de las siguientes CA estimadas de forma derivada sobre el registro de voz:

- Frecuencia Fundamental (Algoritmo de Childers)
- Jitter
- NEP
- GNE
- HNR

Espacio de Características de Representación (CW): Consta de las CV estimadas a partir de los coeficientes de descomposición Wavelet $\mathcal{F} \subset \mathbb{R}^M$. Los parámetros escogidos en la estimación de las CV son los siguientes

- Función madre $\psi(t) \rightarrow$ familias *Daubechies, Symlets, Coiflets, Meyer, biorthogonal spline* y *reverse biorthogonal spline*
- Número de escalas de aproximación $J = 6$. Después del sexto nivel de aproximación, se encontró que no se retiene información representativa del segmento de voz (ver figura 5.1).
- Número de coeficientes seleccionados por escala, $p = 2$.

En total se obtienen 12 CV, a partir de 2 coeficientes de la Transformada Wavelet por cada uno de los 6 niveles de descomposición tenidos en cuenta para el análisis [46].

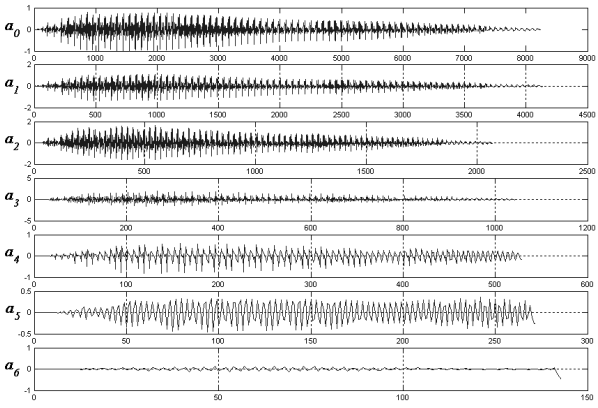


Figura 5.1: Seis escalas de aproximación del segmento de voz */a/*.

5.3 Análisis estadístico del espacio de características

Cada uno de los registros de voz fueron preprocesados acorde a los algoritmos presentados en el capítulo 2. Una vez conformada la matriz inicial de datos, con un total de 91 muestras y a partir de las cuales fueron extraídas las 11 características acústicas para **CA1**, 5 para **CA2** y 12 características de representación para **CW**; se realizó el análisis estadístico de los mismos llevando a cabo los procedimientos de Prueba de hipótesis, Análisis de correlación por rangos y el Análisis de componentes principales, descritos en el capítulo 4.

5.3.1 Prueba de hipótesis

Se realizaron las 11 pruebas de hipótesis sobre espacio de características **CA1** tabla 5.2, 5 pruebas para el **CA2** tabla 5.3 y 12 pruebas para el **CW** tabla 5.4; sobre la media de cada característica x_i siguiendo la ecuación (4.5), con un intervalo de confianza del 95%, y un nivel de significancia del 5%.

CA1	**Característica**	**Rechaza H_0**	**Intervalo de confianza**
1	$Pitch$	No	-0.01 $\leq \mu \leq$ 0.0399
2	$Jitter$	No	-0.0012 $\leq \mu \leq$ 0.0005
3	$Shimmer$	Si	0.0007 $\leq \mu \leq$ 0.003
4	HNR	No	-0.000 $\leq \mu \leq$ 0.0001
5	F_1	Si	0.0563 $\leq \mu \leq$ 0.7306
6	BWF_1	Si	0.001 $\leq \mu \leq$ 0.0094
7	EF_1	Si	0.0871 $\leq \mu \leq$ 0.3888
8	F_2	No	-0.1305 $\leq \mu \leq$ 1.1602
9	BWF_2	Si	0.0006 $\leq \mu \leq$ 0.0126
10	EF_2	Si	0.1380 $\leq \mu \leq$ 0.5609
11	E	Si	0.028 $\leq \mu \leq$ 0.0102

Tabla 5.2: Resultados prueba de hipótesis para CA1

CA2	**Característica**	**Rechaza H_0**	**Intervalo de confianza**
1	$Pitch$	No	-16.2794 $\leq \mu \leq$ 33.98159
2	$Jitter$	No	-0.0703 $\leq \mu \leq$ 0.1266
3	NEP	Si	-0.1556 $\leq \mu \leq$ -0776
4	GNE	Si	0.0294 $\leq \mu \leq$ 0.1063
5	HNR	Si	-8.9055 $\leq \mu \leq$ -5.5923

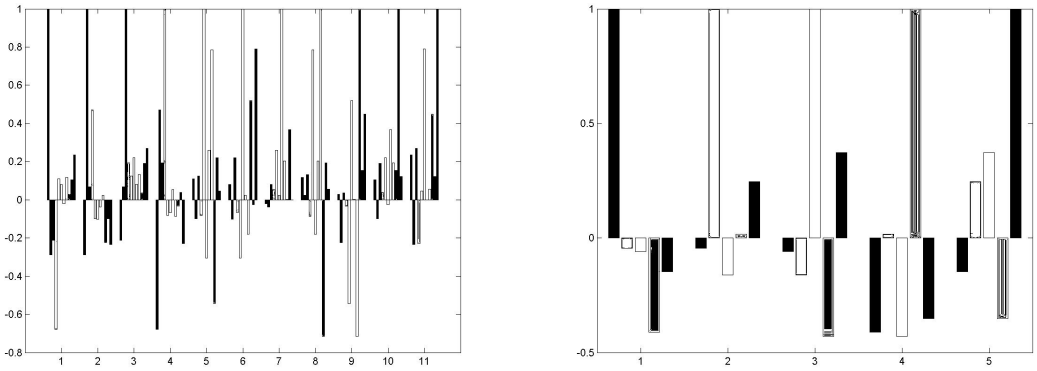
Tabla 5.3: Resultados prueba de hipótesis para CA2

5.3.2 Análisis de correlación por rangos

Se calculó el coeficiente de correlación de Spearman para cada par de características para los tres espacios conformados (**CA1, CA2, CW**). La matrices de correlación por rangos obtenidas en cada caso, se observan en las figuras 5.2(a), 5.2(b) y 5.3.

CW	**Coeficiente**	**Rechaza H_0**	**Intervalo de confianza**
1	c_{11}	No	-0.048081 $\leq \mu \leq$ 0.124842
2	c_{12}	No	-0.040943 $\leq \mu \leq$ 0.135831
3	c_{21}	No	-0.057971 $\leq \mu \leq$ 0.193823
4	c_{22}	No	-0.055608 $\leq \mu \leq$ 0.196972
5	c_{31}	Si	0.162217 $\leq \mu \leq$ 0.548874
6	c_{32}	Si	0.181570 $\leq \mu \leq$ 0.562471
7	c_{41}	No	-0.246689 $\leq \mu \leq$ 0.436232
8	c_{42}	No	-0.241890 $\leq \mu \leq$ 0.424520
9	c_{51}	Si	-1.364304 $\leq \mu \leq$ -0.547752
10	c_{52}	Si	-1.331879 $\leq \mu \leq$ -0.539226
11	c_{61}	No	-0.719508 $\leq \mu \leq$ 0.002833
12	c_{62}	Si	-0.724954 $\leq \mu \leq$ -0.008289

Tabla 5.4: Resultados prueba de hipótesis para CW



(a) (b)

Figura 5.2: Matriz de correlación por rangos para: (a) CA1; (b) CA2.

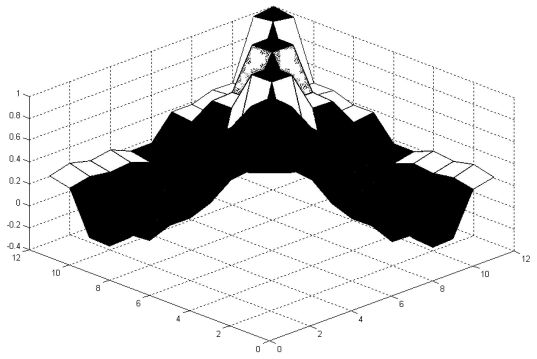


Figura 5.3: Matriz de correlación por rangos para CWT

5.3.3 Análisis de componentes principales

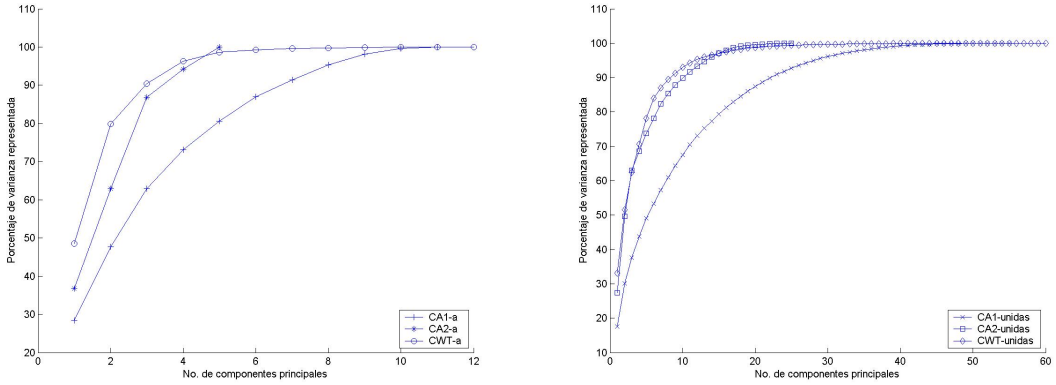
Para evaluar el desempeño de esta metodología, se conformaron 12 conjuntos de datos a partir de los conjuntos descritos en 5.1.2 y los espacios conformados en 5.2; entre estas doce variantes se encuentran conjuntos de CV de naturaleza única (acústicas ó de representación), así como aquellos en donde se tienen en cuenta CV de naturaleza diferentes (acústicas y de representación). Los conjuntos conformados y su correspondiente dimensión, son los siguientes:

Conjunto	**Dimensión**
CA1-a	[11]
CA2-a	[5]
CWT-a	[12]
CA1-unidas	[11x5]
CA2-unidas	[5x5]
CWT-unidas	[12x5]
CA1-indep	[11]
CA2-indep	[5]
CWT-indep	[12]
Combina-a	[11+5+12]
Combina-unidas	[55+25+60]
Combina-indep	[11+5+12]

Tabla 5.5: Conjuntos de datos utilizados

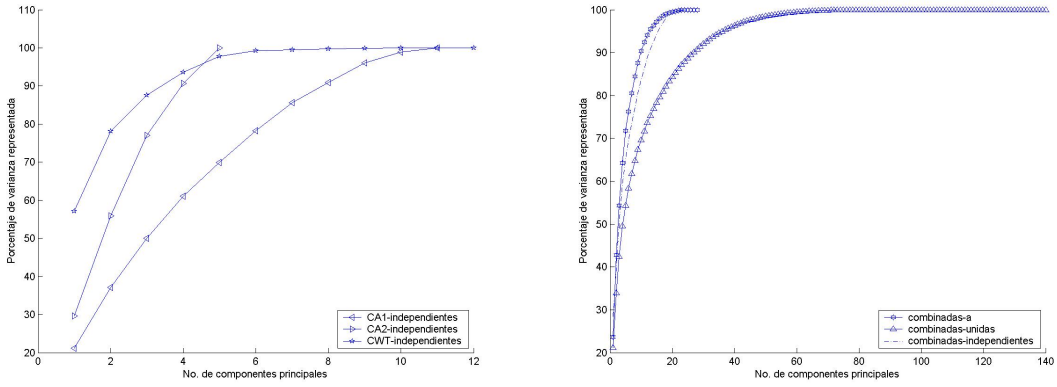
PCA Lineal

Las figuras (5.4(a)), (5.4(b)), (5.5(a)) y (5.5(b)) muestran las curvas de la varianza acumulada para cada uno de los análisis realizados, se puede observar el rápido crecimiento de algunas de estas curvas, lo que permite una mayor reducción en la dimensionalidad del conjunto de datos.



(a) (b)

Figura 5.4: Varianza acumulada para: (a) CA1-a, CA2-a y CWT-a; (b) CA1-unidas, CA2-unidas y CWT-unidas.

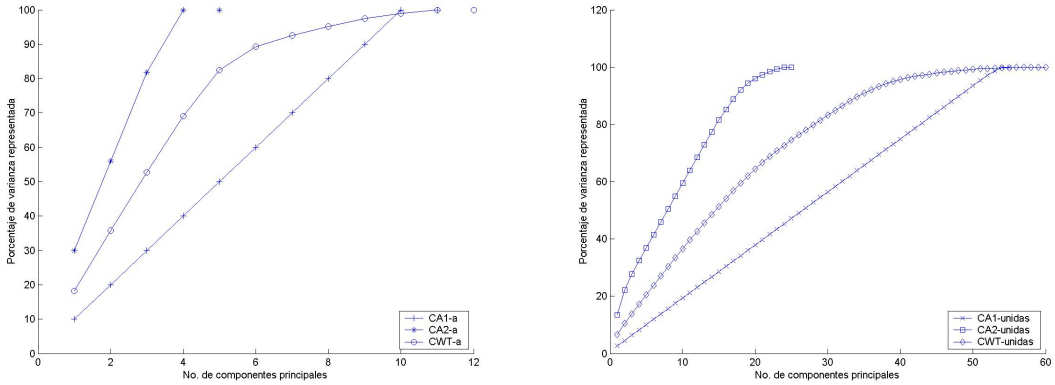


(a) (b)

Figura 5.5: Varianza acumulada para: (a) CA1-indep, CA2-indep y CWT-indep; (b) combi-a, combi-unidas y combi-indep.

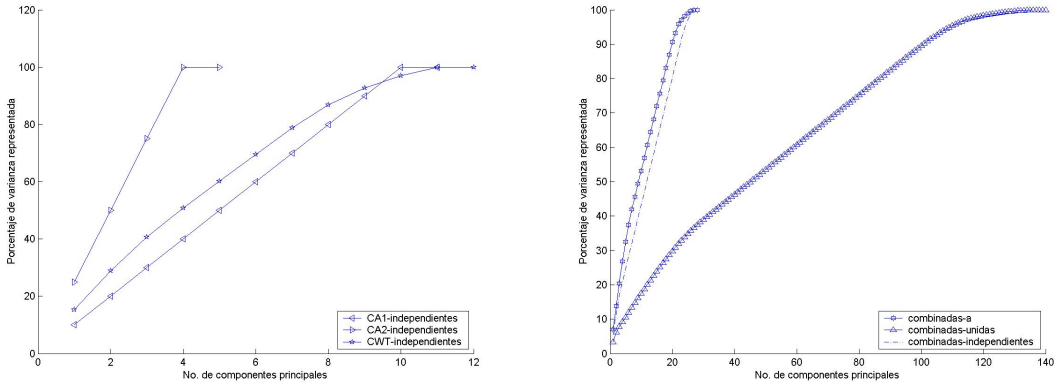
PCA No Lineal

Las figuras 5.6(a), 5.6(b) , 5.7(a) y 5.5(b) muestran las curvas de la varianza acumulada obtenidas en cada uno de los análisis realizados para el caso no lineal.



(a) (b)

Figura 5.6: Varianza acumulada para: (a) CA1-a, CA2-a y CWT-a; (b) CA1-unidas, CA2-unidas y CWT-unidas.



(a) (b)

Figura 5.7: Varianza acumulada para: (a) CA1-indep, CA2-indep y CWT-indep; (b) combi-a, combi-unidas y combi-indep.

5.4 Pruebas de clasificación

Una vez se tiene el clasificador, es necesario evaluar su utilidad midiendo el porcentaje de observaciones que fueron clasificadas correctamente. Esto genera una estimación de la probabilidad de casos correctamente clasificados [46].

5.4.1 Validación del clasificador

Se presentan dos métodos para estimar la probabilidad de éstos: *prueba de muestras independientes*(ITS) y *validación cruzada*(VC).

- *Prueba de muestras independientes:* si el conjunto de muestras es grande, se puede dividir en un conjunto de entrenamiento y un conjunto de validación. Se usa el conjunto de entrenamiento para construir el clasificador y se clasifican las observaciones del conjunto de validación usando la regla de clasificación. La proporción de observaciones correctamente clasificadas es el porcentaje de clasificación estimado. Como el clasificador no ha visto los patrones en el conjunto de validación, el porcentaje de clasificación estimado no está sesgado. Los pasos para evaluar el clasificador usando este método son

  1. Separar aleatoriamente la muestra en dos conjuntos de tamaño n_{TEST} y n_{TRAIN}, donde $n_{TEST} + n_{TRAIN} = N_s$.
  2. Construir el clasificador (e.g., Bayesiano, Red Neuronal, SVM) usando el conjunto de entrenamiento.
  3. Presentar cada patrón del conjunto de validación al clasificador y obtener una etiqueta de clase para él. Dado que se conoce la clase correcta para estas observaciones, se

puede contar el número de correctamente clasificados N_{CC}.

4. El porcentaje en que las observaciones son correctamente clasificadas es

$$P(CC) = 100 * \frac{N_{CC}}{n_{TEST}} \tag{5.1}$$

La ventaja de este método reside en que no emplea mucho tiempo de proceso; sin embargo, su evaluación puede tener una alta varianza, que puede depender en gran medida de los datos que finalmente quedan tanto para el conjunto de entrenamiento como para el de validación. De este modo la evaluación puede ser significativamente diferente dependiendo de cómo es hecha la partición inicialmente.

– *Validación cruzada con k-particiones:* El concepto básico consiste en dividir el conjunto de muestras en *k* particiones de tamaño $k_p = N_s/k$, con N_s siendo el número total de patrones en $\mathcal{F}X$. Una partición es reservada como conjunto de validación mientras los restantes $k-1$ son usadas como conjunto de entrenamiento. El modelo es entrenado k veces a través del conjunto de muestras completo. Cuando éste muy pequeño para dividirlo en un sólo conjunto de entrenamiento y validación, es recomendable usar validación cruzada [59]. El siguiente procedimiento permite estimar el costo de clasificar de forma incorrecta una observación usando validación cruzada.

1. Dividir aleatoriamente el conjunto de muestras en k particiones.
2. Retener una de las k particiones con objeto de validación. Denominar k_T
3. Usar las restantes $k-1$ particiones para entrenar el clasificador.
4. Presentar el conjunto de validación k_T al clasificador y obtener las etiquetas de clase.

5. Determinar el error de clasificación. $e_{ci} = \frac{N_{CI}}{k_p}$.

6. Repetir desde el paso 2 al 5 hasta que las k particiones hayan sido usadas como conjunto de validación.

7. Determinar el promedio de los k errores.

La ventaja de este método es su poca sensibilidad a la partición de los datos. Cada dato logra estar en el conjunto de validación una vez, y $k - 1$ veces en el conjunto de entrenamiento. La estimación de la varianza se reduce a medida que k incrementa. La desventaja de este método es que el algoritmo de entrenamiento debe ser evaluado k veces, elevando el tiempo de cálculo.

– *LOO:* Por último el método *LOO Leave-one-out* es una validación cruzada de k particiones tomada en el límite donde $k = N_s$. Esto significa que k_s veces, de forma independiente, el modelo es entrenado sobre todos los datos exceptuando 1 punto. Luego, la validación es hecha para tal punto. Nuevamente el error promedio es calculado y usado para evaluar el modelo.

Debido al reducido número de muestras por clase, la técnica de *Validación cruzada de k particiones* fue usada como medida de confiabilidad [60] en la evaluación del desempeño de cada uno de las técnicas de reconocimiento de patrones. Esta medida, a su vez permitirá una selección de la técnica más adecuada en términos de generalización dados por el error promedio y su desviación [61].

5.4.2 Resultados del clasificador usando Prueba de Hipótesis

Conjunto	**Completo**	**Reducido**	σ_C	σ_R
CA1-a	78.75	78.75	12.96	3.42
CA2-a	87.50	92.50	7.65	5.23
CWT-a	92.50	86.25	2.80	5.23
CA1-unidas	83.75	82.50	16.89	13.55
CA2-unidas	86.25	87.50	6.85	0.00
CWT-unidas	98.75	99.00	2.80	0.00
CA1-indep	81.25	77.50	13.26	5.59
CA2-indep	82.50	91.25	9.27	3.42
CWT-indep	92.50	75.00	8.15	8.84

Tabla 5.6: Clasificación utilizando CV seleccionadas con PH

En 5.6, se muestran los resultados obtenidos en la clasificación utilizando el conjunto completo de datos, al igual que los resultados que se obtienen al utilizar solamente aquellas características señaladas por la prueba de hipótesis. En un primer grupo $\{CA1 - a, CA2 - a, CWT - a\}$, el porcentaje de acierto se mantiene estable o se mejora para las dos primeras, al mismo tiempo que se nota un descenso en el valor de la varianza de clasificación para el procedimiento VC; pero en el caso de las CWT, el porcentaje desciende considerablemente y la varianza aumenta. Para el segundo grupo $\{CA1 - unidas, CA2 - unidas, CWT - unidas\}$, el porcentaje de acierto se mantiene estable para todas, y para el caso de la varianza se presenta reducción en los tres casos, siendo destacable el valor 0 alcanzado para CA2-unidas y CWT-unidas. Para el tercer grupo $\{CA1 - indep, CA2 - indep, CWT - indep\}$, el porcentaje solo se mantiene estable para CA2-indep, en los otros dos casos este desciende, y en cuanto a la varianza, su valor disminuye para CA1-indep y CA2-indep, pero aumenta para CWT-indep.

Conjunto	**Completo**	**Reducido**	σ_C	σ_R	**CP retenidos**
CA1-a	76.25	78.75	11.18	7.13	8/11
CA2-a	85.00	86.25	5.59	9.27	5/5
CWT-a	87.50	82.50	11.69	6.85	4/12
CA1-unidas	87.50	78.75	6.25	12.18	29/55
CA2-unidas	90.00	73.75	8.39	8.15	14/25
CWT-unidas	95.00	96.25	5.23	3.42	12/60
CA1-indep	79.25	78.50	4.73	2.05	9/11
CA2-indep	81.25	81.75	4.76	3.49	5/5
CWT-indep	83.50	78.25	5.11	5.42	5/12
combina-a	97.50	96.25	3.42	5.59	13/28
combina-unidas	97.50	92.50	3.42	2.80	37/140
combina-indep	92.75	88.50	1.85	4.28	16/28
	Error prom.	3.42	σ prom.	5.97	

Tabla 5.7: Clasificación SVM de CV seleccionadas con PCA Lineal

5.4.3 Resultados del clasificador usando PCA lineal

En 5.7 se resumen los resultados obtenidos al aplicar la metodología de reducción de dimensionalidad a través de PCA. Se muestran resultados para doce conjuntos de características, en donde se tienen en cuenta CV de naturaleza única, así como conjuntos que incluyen la combinación de las mismas. Se observan los porcentajes de acierto alcanzados por el clasificador para el conjunto *completo* y para el conjunto *reducido*, al igual que el numero de componentes principales retenidos para cada prueba, esto para un criterio del 90% de la varianza acumulada. Para los conjuntos CA1-a, CA2-unidas, CWT-unidas, combi-a y combi-unidas, se manifiesta un aumento en el porcentaje de acierto en la clasificación. Para CWT-a, CA1-unidas y CA2-indep el porcentaje se mantiene estable. Caso contrario ocurre para CA2-a, CA1-indep, CWT-indep y combi-indep, en donde se ve una disminución, aunque leve, en el porcentaje de acierto.

Conjunto	**Completo**	**Reducido**	**Reducido 2**	**CP retenidos**	**CP retenidos 2**
CA1-a	76.25	78.75	76.25	8/11	4/11
CA2-a	85.00	86.25	77.50	5/5	3/5
CWT-a	87.50	82.50	85.00	4/12	3/12
CA1-unidas	87.50	78.75	67.50	29/55	16/55
CA2-unidas	90.00	73.75	60.00	14/25	7/25
CWT-unidas	95.00	96.25	96.25	12/60	10/60
CA1-indep	79.25	78.50	61.50	9/11	4/11
CA2-indep	81.25	81.75	66.00	5/5	3/5
CWT-indep	83.50	78.25	68.50	5/12	3/12
combina-a	97.50	96.25	82.50	13/28	8/28
combina-unidas	97.50	92.50	90.00	37/140	26/140
combina-indep	92.75	88.50	78.00	16/28	8/28

Tabla 5.8: Clasificación SVM de CV seleccionadas con PCA Lineal con dos criterios de retención

La tabla 5.8, es una versión ampliada de 5.7, en donde se analiza el desempeño con otra estrategia de selección del número de componentes principales retenidos. Para este segundo caso, se retienen los componentes principales que tienen un autovalor mayor o igual al valor de la media de los autovalores.

Conjunto	**Completo**	**Reducido**	σ_C	σ_R	**CP retenidos**
CA1-a	67.50	70.00	11.18	5.23	8/11
CA2-a	78.75	78.75	5.59	5.59	5/5
CWT-a	85.00	81.25	12.18	9.88	4/12
CA1-unidas	50.00	51.25	0.00	11.18	29/55
CA2-unidas	72.50	66.25	10.46	11.35	14/25
CWT-unidas	50.00	82.50	0.00	20.44	12/60
CA1-indep	62.00	59.75	9.38	11.26	9/11
CA2-indep	69.50	69.50	4.89	4.89	5/5
CWT-indep	79.75	74.00	7.26	2.40	5/12
combina-a	73.75	86.25	13.55	8.15	13/28
combina-unidas	50.00	45.00	0.00	12.02	37/140
combina-indep	81.50	80.50	7.83	2.27	16/28

Tabla 5.9: Clasificación Bayesiano de CV seleccionadas con PCA Lineal

Conjunto	**Completo**	**Reducido**	**Reducido 2**	**CP retenidos**	**CP retenidos 2**
CA1-a	67.50	70.00	68.75	8/11	4/11
CA2-a	78.75	78.75	68.75	5/5	3/5
CWT-a	85.00	81.25	85.00	4/12	3/12
CA1-unidas	50.00	51.25	53.75	29/55	16/55
CA2-unidas	72.50	66.25	53.75	14/25	7/25
CWT-unidas	50.00	82.50	85.00	12/60	10/60
CA1-indep	62.00	59.75	44.75	9/11	4/11
CA2-indep	69.50	69.50	56.25	5/5	3/5
CWT-indep	79.75	74.00	69.00	5/12	3/12
combina-a	73.75	86.25	80.00	13/28	8/28
combina-unidas	50.00	45.00	66.25	37/140	26/140
combina-indep	81.50	80.50	72.50	16/28	8/28

Tabla 5.10: Clasificación Bayesiano de CV seleccionadas con PCA Lineal con dos criterios de retención

5.4.4 Resultados del clasificador usando PCA No lineal

Conjunto	**Completo**	**Reducido**	σ_C	σ_R	**CP retenidos**
CA1-a	83.75	80.00	5.59	12.02	9/11
CA2-a	88.75	78.75	8.15	5.59	4/5
CWT-a	93.75	95.00	4.42	5.23	8/12
CA1-unidas	92.50	91.25	5.23	7.13	49/55
CA2-unidas	87.50	85.00	4.42	5.59	21/25
CWT-unidas	97.50	96.25	5.59	5.59	44/60
CA1-indep	78.50	70.25	5.33	7.15	10/11
CA2-indep	83.00	76.75	3.38	6.99	4/5
CWT-indep	81.25	81.75	2.50	5.63	10/12
combina-a	97.50	98.75	5.59	2.80	22/28
combina-unidas	98.75	95.00	2.80	5.23	113/140
combina-indep	92.25	92.75	3.11	5.26	24/28

Tabla 5.11: Clasificación SVM de CV seleccionadas con PCA No Lineal

En 5.11 se resumen los resultados obtenidos al aplicar la metodología no lineal de reducción de dimensionalidad a través de KPCA. El kernel utilizado fué del tipo RBF. Se muestran resultados para los mismos conjuntos de datos tenidos en cuenta para PCA. Se observan los porcentajes de acierto alcanzados por el clasificador para el conjunto *completo* y para el conjunto *reducido*, al igual que el numero de componentes principales retenidos para cada prueba, esto para un criterio del 90% de la varianza acumulada. Para los conjuntos CA2-a, CWT-a, CET-unidas, CA2-indep y combi-unidas se manifiesta un aumento en el porcentaje de acierto en la clasificación. Para CA1-a, combi-a y combi-indep el porcentaje se mantiene estable. Caso contrario ocurre para CA1-unidas, CA2-unidas, CA1-indep y CWT-indep, en donde se ve una disminución, aunque leve, en el porcentaje de acierto.

La tabla 5.12, es una versión ampliada de 5.7, en donde de analiza el desempeño con otra estrategia

Conjunto	**Completo**	**Reducido**	**Reducido 2**	**CP retenidos**	**CP retenidos 2**
CA1-a	83.75	80.00	81.25	9/11	10/11
CA2-a	88.75	78.75	81.25	4/5	4/5
CWT-a	93.75	95.00	88.75	8/12	7/12
CA1-unidas	92.50	91.25	93.75	49/55	52/55
CA2-unidas	87.50	85.00	87.50	21/25	21/25
CWT-unidas	97.50	96.25	97.50	44/60	31/60
CA1-indep	78.50	70.25	70.25	10/11	10/11
CA2-indep	83.00	76.75	76.75	4/5	4/5
CWT-indep	81.25	81.75	82.00	10/12	10/12
combina-a	97.50	98.75	95.00	22/28	21/28
combina-unidas	98.75	95.00	97.50	113/140	104/140
combina-indep	92.25	92.75	91.50	24/28	24/28

Tabla 5.12: Clasificación SVM de CV seleccionadas con PCA No Lineal con dos criterios de retención

de selección del número de componentes principales retenidos. Para este segundo caso, se retienen los componentes principales que tienen un autovalor mayor o igual al valor de la media de los autovalores.

Conjunto	**Completo**	**Reducido**	σ_C	σ_R	**CP retenidos**
CA1-a	67.50	63.75	11.18	11.18	7/11
CA2-a	78.75	81.25	5.59	9.88	4/5
CWT-a	85.00	87.50	12.18	7.65	3/12
CA1-unidas	53.75	46.25	14.39	7.13	23/55
CA2-unidas	72.50	67.50	10.46	18.96	11/25
CWT-unidas	50.00	37.50	0.00	9.88	9/60
CA1-indep	62.00	51.50	9.38	10.80	8/11
CA2-indep	69.50	57.50	4.89	6.79	4/5
CWT-indep	79.75	78.50	7.26	4.09	4/12
combina-a	73.75	83.75	13.55	5.59	10/28
combina-unidas	50.00	50.00	0.00	0.00	27/140
combina-indep	81.50	81.50	7.83	5.48	13/28

Tabla 5.13: Clasificación Bayesiano de CV seleccionadas con PCA Lineal

Conjunto	**Completo**	**Reducido**	**Reducido 2**	**CP retenidos**	**CP retenidos 2**
CA1-a	67.50	63.75	63.75	7/11	4/11
CA2-a	78.75	81.25	81.25	4/5	3/5
CWT-a	85.00	87.50	87.50	3/12	3/12
CA1-unidas	53.75	46.25	42.50	23/55	16/55
CA2-unidas	72.50	67.50	70.00	11/25	7/25
CWT-unidas	50.00	37.50	83.75	9/60	10/60
CA1-indep	62.00	51.50	51.50	8/11	4/11
CA2-indep	69.50	57.50	57.50	4/5	3/5
CWT-indep	79.75	78.50	78.00	4/12	3/12
combina-a	73.75	83.75	83.75	10/28	8/28
combina-unidas	50.00	50.00	50.00	27/140	26/140
combina-indep	81.50	81.50	81.50	13/28	8/28

Tabla 5.14: Clasificación Bayesiano de CV seleccionadas con PCA No Lineal con dos criterios de retención

Conjunto	Completo	Reducido	Reducido 2	Reducido K	Reducido K2
CA1-a	75.00	78.75	80.00	80.00	81.25
CA2-a	82.50	88.75	70.00	78.75	81.25
CWT-a	93.75	82.50	80.00	95.00	88.75
CA1-unidas	91.25	86.25	72.50	91.25	93.75
CA2-unidas	86.25	75.00	56.25	85.00	87.50
CWT-unidas	96.25	96.25	96.25	96.25	97.50
CA1-indep	81.00	78.75	57.25	70.25	70.25
CA2-indep	81.00	81.75	65.25	76.75	76.75
CWT-indep	82.75	76.25	68.25	81.75	82.00
combina-a	96.25	96.25	81.25	98.75	95.00
combina-unidas	98.75	86.25	88.75	95.00	97.50
combina-indep	92.25	88.50	75.75	92.75	91.50

Tabla 5.15: Comparación de resultados con PCA y KPCA

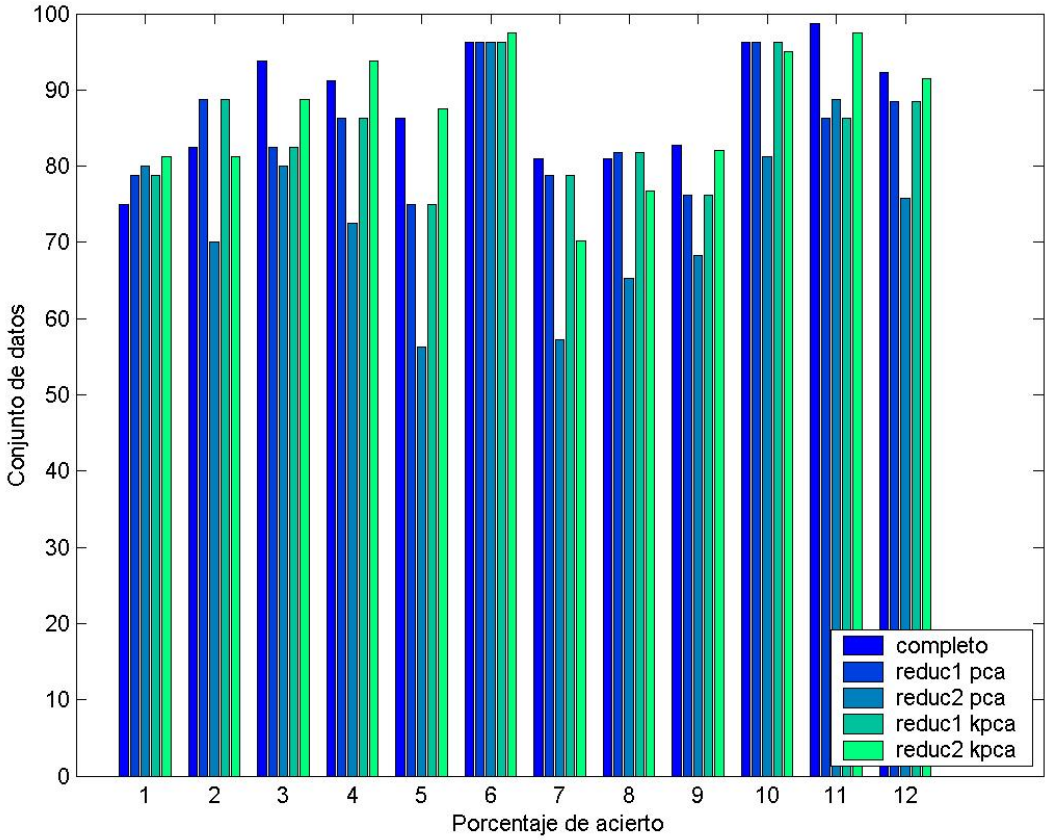


Figura 5.8: Comparación gráfica de resultados con PCA y KPCA, con dos criterios de retención

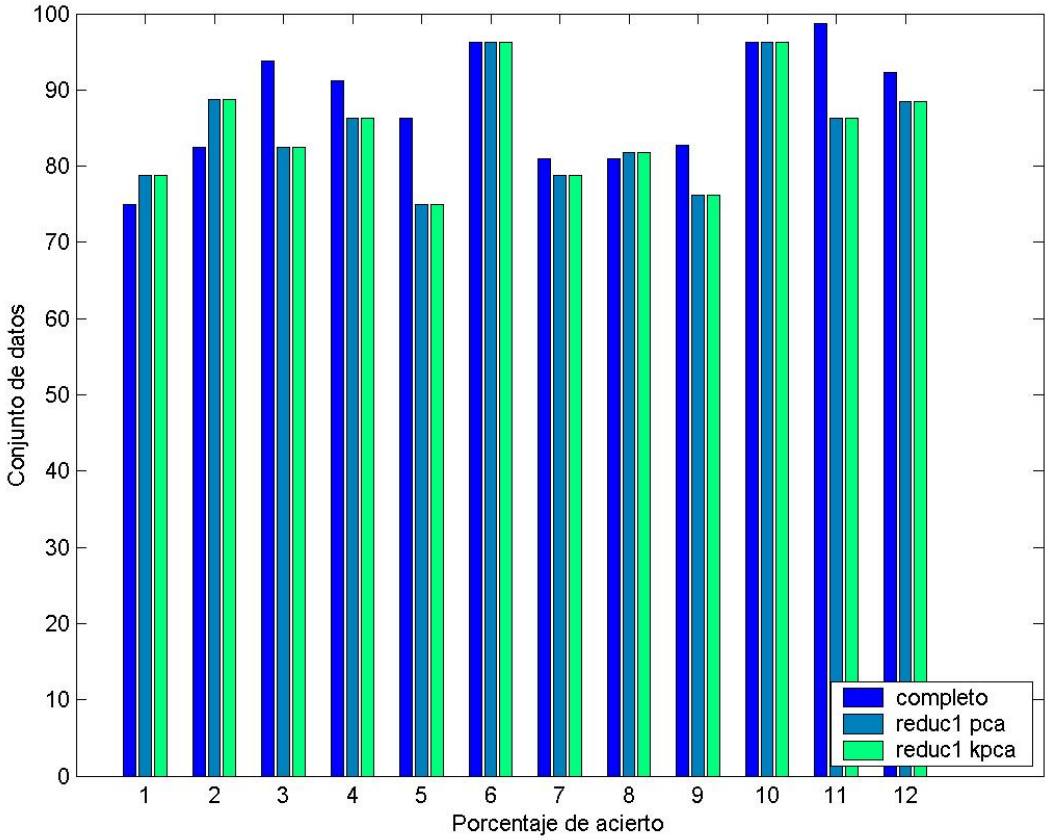


Figura 5.9: Comparación gráfica de resultados con PCA y KPCA

Analizando los resultados obtenidos para cada grupo analizado, se puede concluir que: para los conjuntos que solo tienen en cuenta el fonema */a/*, en el caso de las características acústicas, se presenta un aumento en el porcentaje de clasificación, mientras que para las características de representación este porcentaje disminuye; para los conjuntos que unen las cinco vocales, se presente lo contrario al caso anterior, es decir, se ve una disminución en el porcentaje de clasificación para las características acústicas, y se presenta estabilidad para las características de representación; cuando se tomaron las vocales como muestras independientes, el comportamiento para el primer grupo de características acústicas y para el de características de representación es el mismo, en cuanto a una disminución en el porcentaje, mientras que para el segundo grupo de características acústicas se presenta un leve aumento; para el caso en que se combinaron las características, se presenta estabilidad en el porcentaje para el primer grupo, es decir, aquel que tenía en cuenta solo la vocal */a/*, para los otros dos casos, se presenta una disminución en el porcentaje de acierto.

Capítulo 6
Conclusiones

Teniendo en cuenta los resultados presentados, se pueden destacar como conclusiones de este trabajo, las siguientes:

- Se presenta la clasificación detallada de los algoritmos de estimación automatizada de las características acústicas de voz, sobre las cuales, el interés de la comunidad científica se encuentra en aumento, esto, evidenciado en el creciente número de publicaciones en el tema. En este sentido, se analiza el empleo de las características acústicas en la identificación de las alteraciones vocales en la población colombiana, teniendo en cuenta para este caso, una de las alteraciones mas comunes en la función vocal humana, *la disfonía*.

- El análisis de las condiciones de registro sobre los resultados del clasificador, mostró la necesidad de realizar el preprocesamiento del conjunto inicial de características con el fin de evidenciar la presencia de datos anómalos, permitiendo su eliminación, al mismo tiempo que se normaliza estadísticamente la muestra como etapa anterior al análisis y procesamiento de la misma.

- Se estudian las características de la voz, con naturaleza distinta a las propiedades acústicas. En particular, se evalúan algoritmos de estimación de características de voz empleando los coeficientes de representación de la transformada Wavelet.

- Se propone la metodología de selección de características de voz, orientada a la identificación de voces con algún grado de disfonía. La metodología está basada en el estudio de dependencia estadística de las características iniciales de voz y comprende las siguientes etapas:

  1. La prueba de hipótesis, compara las clases desde el punto de vista de los promedios de cada una de las componentes del espacio de características. $\mathcal{F}$.
  2. El análisis de correlación, que detecta algún grado de dependencia o independencia de parejas de variables.
  3. La reducción de dimensionalidad, que reduce el número de características necesario para llevar a cabo la identificación, utilizando para ello el criterio de independencia estadística entre las mismas.

- El desarrollo de la metodología en el caso concreto de identificación de alteración de voz propuesto, mostró como resultado que para 8 de los 12 casos analizados, para PCA Lineal, y 9 de 12 para PCA No lineal; la metodología presentada permite realizar una reducción en la dimensión del conjunto inicial de características manteniendo estable el porcentaje de acierto en la clasificación, e incluso mejorándolo. A pesar de tener un mejor desempeño en cuanto al porcentaje de clasificación, el análisis PCA No lineal ocasiona, de manera general, un aumento en la varianza de clasificación, lo que se convierte en un efecto no deseado. Adicionalmente, la reducción alcanzada no es tan buena como la alcanzada para el caso del análisis PCA lineal.

- Los mejores resultados en cuanto a la reducción de dimensionalidad, se obtienen para el

conjunto *CWT-unidas*, en donde se obtiene una reducción de casi el 80%. Adicionalmente el porcentaje de clasificación se mantiene estable, y aparece como uno de los mas altos dentro de todo el análisis realizado. De lo anterior se puede decir que, las caracterísitcas de representación a partir de todas las vocales unidas, presentan un mejor desempeño respecto a las características acústicas en la tarea de clasificación entre voces normales y con algún grado de disfonía. Al revisar las características seleccionadas para los conjuntos ***-unidas, es posible notar la presencia de por lo menos una característica calculada a partir de cada una de las cinco vocales, lo que permite pensar que cada uno de los fonemas vocálicos provee cierta información propia e independiente, respecto a los demás.

– En la actualidad el Grupo de Control y Procesamiento Digital de Señales (GC&PDS) de la Universidad Nacional de Colombia, Sede Manizales, cuenta con una base reducida de voces. Dicha base está conformada por 91 voces de hombres y mujeres con edades entre los 19 y 54 años, discriminadas en voces normales (40) y con algún grado de disfonía (51). Es necesario continuar con las labores de recolección y etiquetamiento de muestras de señales de voz con estas características, especialmente con algún grado de disfonía, lo anterior con el fin de ofrecer a mediano plazo, una muestra representativa que contenga un número considerable de elementos para cada una de los posibles tipos de disfonía: leve, moderada y severa; hiperfuncional e hipofuncional.

Apéndice A
Transformada Wavelet Discreta

En este caso, los parámetros de dilatación a y traslación b toman solamente valores discretos. La dilatación de la Wavelet madre, se relaciona como potencias enteras de una escala de referencia a_0, normalmente mayor que 1 así $a = a_0^j$. Para la discretización del parámetro b, se debe tener en cuenta que el recubrimiento discreto del plano tiempo-frecuencia es localizado en cada escala, así el parámetro de traslación depende del parámetro de escala. Para escalas mayores, la traslación debe ser mayor. Dado que el ancho de las funciones a cada escala es directamente proporcional con la misma, se toma una discretización del parámetro b directamente relacionada con la escala que se está trabajando $b = kb_0a_0^j$.

$$\begin{aligned} a &= a_0^j \\ b &= kb_0a_0^j \end{aligned} \tag{A.1}$$

con $j,\ k \in \mathbb{Z}$ y $a_0 > 1,\ b_0 > 0$.

Una base ortonormal de Wavelets de soporte compacto pueden ser obtenidas al extender $L^2(\mathbb{R})$, el espacio de todas las señales de energía finita, por medio de traslaciones y dilataciones de la función Wavelet (ver ecuación (2.29)), así

$$\psi_{j,k}(t) = a_0^{-j/2}\psi\left(a_0^{-j}t - kb_0\right) \tag{A.2}$$

donde m representa la escala y n la traslación temporal. Si se seleccionan escalas y posiciones

basadas en potencias de 2($a_0 = 2$), llamadas escalas y posiciones *diádicas*, el análisis será mucho más eficiente e igual de preciso que el análisis continuo. Una vía para implementar este esquema usando filtros fue desarrollada por Mallat [62, 63], cuyo algoritmo es en efecto un esquema clásico conocido como *codificador sub-banda* de dos canales.

En este caso, la señal $f(t)$ se representa como una serie de *aproximaciones* (baja frecuencia) y *detalles* (alta frecuencia) en diferentes resoluciones. En cada etapa, un par de filtros $h,\ g$ son aplicados a la señal de entrada para producir una señal de aproximación y una de detalle respectivamente. La señal de detalle, representa la información perdida desde una resolución alta, hasta una más baja. La representación Wavelet es entonces, el conjunto de coeficientes de detalle en todas las resoluciones y los coeficientes de aproximación en la resolución más baja.

El algoritmo rápido para calcular los coeficientes Wavelet, está dado por la siguiente expresión

$$ca_{j,k} = \sum_m h[2k-m]ca_{j-1}[m] \tag{A.3}$$

$$cd_{j,k} = \sum_m g[2k-m]ca_{j-1}[m] \tag{A.4}$$

Los filtros h y g son llamados *filtros espejo en cuadratura* y satisfacen la siguiente propiedad

$$g[n] = (-1)^{1-n}h[1-n] \tag{A.5}$$

La etapa de filtrado es seguida por una decimación diádica o submuestreo por un factor de 2. El esquema para una etapa de filtrado a una escala j se muestra en la figura A.1. La descomposición Wavelet de una señal s analizada en una escala o nivel j, tiene la siguiente estructura: $[ca_j, cd_j, ..., cd_1]$ (ver figura A.2). Los filtros h y g son derivados de bases de Wavelets ortonormales y, por lo tanto la reconstrucción de la señal a partir de la descomposición Wavelet es exacta

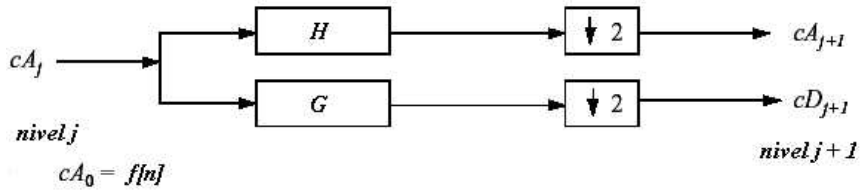


Figura A.1: Etapa de descomposición

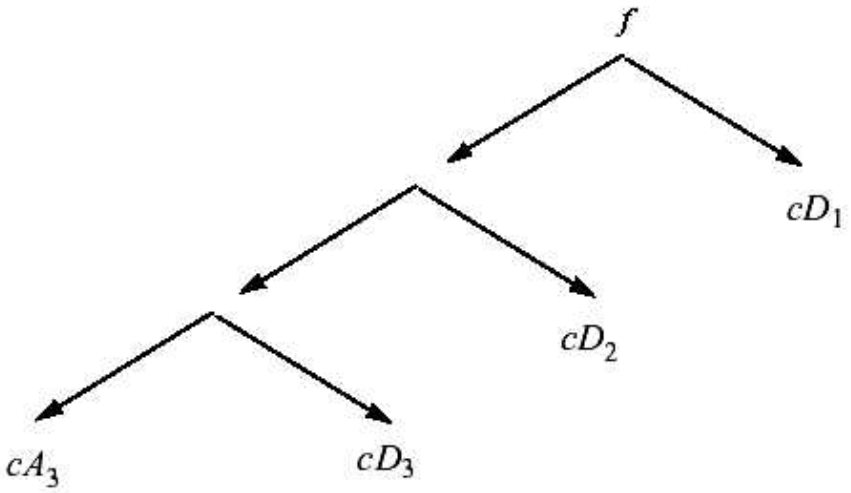


Figura A.2: Estructura de la descomposición Wavelet: *árbol Wavelet*

y dada por la ecuación (A.6) y se representa en la figura A.3.

$$ca_{j-1,k} = 2\sum_{m}(ca_{j,k}[m]h[k-2m] + cd_{m,k}[m]g[k-2m]) \tag{A.6}$$

hace alusión al cálculo real de los coeficientes.

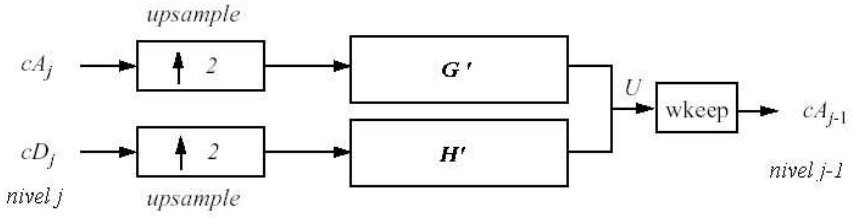


Figura A.3: Etapa de reconstrucción

Apéndice B
Reconocimiento Automático de Patologías de Voz

B.1 Clasificador bayesiano

En este caso de clasificación, el criterio de trabajo consiste en minimizar la probabilidad de error en un problema de clasificación [64]. El algoritmo de decisión bayesiana evalúa el punto a clasificar en cada una de las funciones discriminantes construidas para cada clase. En este trabajo se ha supuesto que las clases tienen igual probabilidad *a priori* de aparición.

Sea $\mathbf{X}_i$ la matriz que contiene los hiperpuntos de cada clase, de tamaño N_c(muestras por clase) $\times$ D(No. características) $\times$ C(No. de clases), se procede del siguiente modo:

1. Se calcula el vector de medias μ_i de $\mathbf{X}_i$.
2. Se calcula la matriz de covarianza $\mathbf{\Sigma}_i$ de $\mathbf{X}_i$.
3. Se calculan los coeficientes de la función discriminante para cada clase:

$$\begin{aligned} \mathbf{W}_i &= -\frac{1}{2}\mathbf{\Sigma_i^{-1}} \\ \mathbf{w}_i &= \mu_i \Sigma_i^{-1} \\ w_{i0} &= -\frac{1}{2}\mu_i \Sigma_i^{-1} \mu_i^T - \frac{1}{2}\ln(\mid \Sigma_i \mid) + \ln(P) \end{aligned} \tag{B.1}$$

4. Se construye la función discriminante para cada clase con los coeficientes calculados en B.1:

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i + \mathbf{w}_i \mathbf{x} + w_{i0} \tag{B.2}$$

El punto pertenece a aquella clase que da un mayor valor al evaluarlo en su función discriminante.

B.2 Redes Neuronales Artificiales

Las redes neuronales están compuestas de elementos simples operando en paralelo. Estos elementos están inspirados en los sistemas nerviosos biológicos. El modelo de una neurona está compuesto de una entrada escalar p que es multiplicada por un escalar de peso w, un escalar de polarización b, una función de transferencia f y una salida a. Este modelo puede ser descrito por medio de la ecuación,

$$a = f(wp + b) \tag{B.3}$$

La función de transferencia f es típicamente una función de paso o una función sigmoide.

Cabe notar que w y b son parámetros ajustables de la neurona. La idea central de las redes neuronales es que dichos parámetros pueden ser ajustados tal que la red exhiba algún comportamiento deseado. De este modo, se pueden entrenar redes para realizar un trabajo particular ajustando estos parámetros, o tal vez la propia red pueda ajustarse para alcanzar alguna salida deseada [65].

Una red puede tener varias capas, la primera capa se denomina capa de entrada, la última capa es la capa de salida y las capas restantes se denominan capas ocultas. Cada capa tiene una matriz de pesos $\mathbf{W}$, un vector de polarización b, y un vector de salida a. La arquitectura de una red neuronal de tres capas se muestra a continuación: Si una capa de la red tiene únicamente conex-

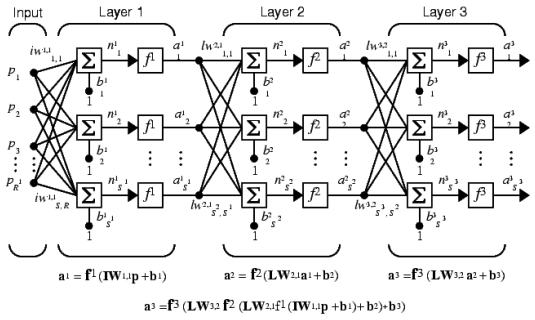


Figura B.1: Arquitectura de una red neuronal de 3 capas

iones hacia las capas que se encuentran a su derecha, la red se denomina de alimentación hacia adelante. Típicamente, las redes de alimentación hacia adelante son entrenadas con una función de desempeño de gradiente descendiente para determinar cómo ajustar los pesos para minimizar el desempeño. El gradiente es determinado usando una técnica llamada *backpropagation*, que involucra cálculos hacia atrás a través de la red. En su implementación más simple del aprendizaje *backpropagation*, la red actualiza los pesos y las polarizaciones en la dirección en la cual la función de desempeño decrece más rápidamente – el negativo del gradiente. Una iteración del algoritmo puede ser escrita como

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \mathbf{g}_k \tag{B.4}$$

donde $\mathbf{w}_k$ es un vector de pesos actuales, $\mathbf{g}_k$ es el gradiente actual, y α_k es la tasa de aprendizaje. Sin embargo la convergencia de este algoritmo es lenta y no siempre se logra alcanzar un mínimo global. El método de Newton o gradiente conjugado, es una alternativa basada en técnicas de optimización para acelerar la convergencia del algoritmo *backpropagation*. El paso básico del método de Newton es

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mathbf{H}_k^{-}1\mathbf{g}_k \tag{B.5}$$

donde $\mathbf{H}_k$ es la matriz *Hessiana* (derivadas de segundo orden) del índice de desempeño para los valores actuales de pesos y polarizaciones. Este método es más eficiente, pero sus requerimientos computacionales y de almacenamiento son hasta el cuadrado del tamaño de la red. Existe una clase de algoritmos llamados *cuasi-Newton* que actualizan una aproximación de $\mathbf{H}$ en cada iteración, éstos convergen más rápido mientras limitan los requerimientos de memoria [66].

Del mismo modo que los métodos cuasi-Newton, el algoritmo *Levenberg-Marquardt* fue diseñado para acelerar la convergencia de los métodos de segundo orden sin necesidad de calcular $\mathbf{H}$. Cuando la función de desempeño tiene la forma de una suma de cuadrados, la matriz Hessiana puede ser aproximada como

$$\mathbf{H} = \mathbf{J^T J} \tag{B.6}$$

y el gradiente calculado como

$$\mathbf{g} = \mathbf{J^T e} \tag{B.7}$$

donde $\mathbf{J}$ es la matriz *Jacobiana* que contiene las primeras derivadas de los errores de la red con respecto a los pesos y las polarizaciones, y $\mathbf{e}$ es el vector de errores de la red. La matriz Jacobiana puede ser calculada a través de la técnica *backpropagation* [67] que es un cálculo menos complejo que el de la matriz Hessiana. El algoritmo *Levenberg-Marquardt* usa esta aproximación a la matriz Hessiana en la siguiente actualización

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \left[\mathbf{J^T J} + \mu \mathbf{I}\right]^{-1} \mathbf{J^T e} \tag{B.8}$$

Cuando el escalar μ es cero, la ecuación (B.8) se traduce al método de Newton (ecuación (B.5)), usando la aproximación de la matriz Hessiana (ecuación (B.6). Cuando μ es grande, se convierte

en el método de gradiente descendiente (ecuación (B.4)) con un pequeño paso. De este modo, la función de desempeño es siempre reducida en cada iteración del algoritmo.

En este trabajo se usó una red neuronal artificial *backpropagation*. Se usaron 10 nodos en la capa oculta y 1 nodo de salida. Las funciones de activación en cada capa son del tipo *tansig* [65]. Cuando la red produce un +1 corresponde a la voz normal y para la voz patológica se asigna un -1. Se usó un algoritmo de entrenamiento Levenberg-Marquardt [67].

B.3 Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (SVM) están sustentadas en el principio de *minimización de riesgo estructural* (SRM) propuesto en [68]. Un subconjunto de funciones encontradas en el proceso de optimización minimizan el riesgo actual del problema, de manera que entrenando una serie de máquinas para el objetivo dado, se minimizan el riesgo y la confiabilidad de la dimensión Vapnik-Chervonenkis (VC). Esta dimensión implica los requerimientos de almacenamiento de la técnica de aprendizaje y la calidad de sus respuestas para responder a un problema de clasificación. En forma general, la función de riesgo actual $R(\alpha)$ es expresada como una cota, para la definición de la cual se determina el riesgo empírico $R_{emp}(\alpha)$ como el promedio de los errores de entrenamiento para un número finito y fijo de observaciones $\{\mathbf{x}, \mathbf{y}\}$ ($\mathbf{x}_i$: patrón, y_i: etiqueta del patrón i):

$$R_{emp}(\alpha) = \frac{1}{2l} \sum |y_i - f(\mathbf{x}_i, \alpha)| \tag{B.9}$$

La cantidad $\frac{1}{2} |y_i - f(\mathbf{x}_i, \alpha)| \in [0, 1]$ es llamada pérdida. Para un número η tal que $0 < \eta < 1$, que

representa las pérdidas se tiene que [68]

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h)+1) - \log(\eta/4)}{1}\right)} \tag{B.10}$$

donde h es la dimensión Vapnik-Chervonenkis (VC).

Sea un grupo de datos de entrenamiento $\{\mathbf{x}_i, y_i\}$ con $i = 1, ..., l$, $y_i \in \{-1, 1\}$ y $\mathbf{x}_i \in \mathcal{F} \subset \mathbb{R}^M$. Existe un hiperplano que separa los datos de etiquetas positivas y negativas [69].

$$\begin{aligned} &\mathbf{x}_i.\mathbf{w} + b \geq 1 - \zeta_i \text{ para } y_i = 1 \\ &\mathbf{x}_i.\mathbf{w} + b \leq -1 + \zeta_i \text{ para } y_i = -1 \\ &\zeta_i \geq 0, \forall i \end{aligned} \tag{B.11}$$

donde $\mathbf{w}$ es la normal al hiperplano y ζ_i son las variables introducidas por errores de clasificación como violaciones del hiperplano, de manera que $\sum \zeta_i$ es la cota del error de clasificación. Una manera natural de añadir un costo a la función objetivo es minimizar $\|\mathbf{w}\|^2/2 + C\sum \zeta_i$ [69], donde C es una constante elegida por el usuario correspondiente al inverso de la penalización de los errores. Así, la anterior función objetivo (ecuación (B.9)) corresponde a un problema de optimización convexa entendido como un problema de programación cuadrática (QP), cuya forma dual Wolfe es [70]:

Maximizar:

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j \tag{B.12}$$

Sujeto a:

$$0 < \alpha_i < C \tag{B.13}$$

$$\sum \alpha_i y_i = 0 \tag{B.14}$$

con solución en forma de:

$$\mathbf{w} = \sum^{n_s} \alpha_i y_i \mathbf{x}_i \tag{B.15}$$

donde n_s es el número de vectores de soporte. Por cuanto, en la mayoría de los casos el espacio

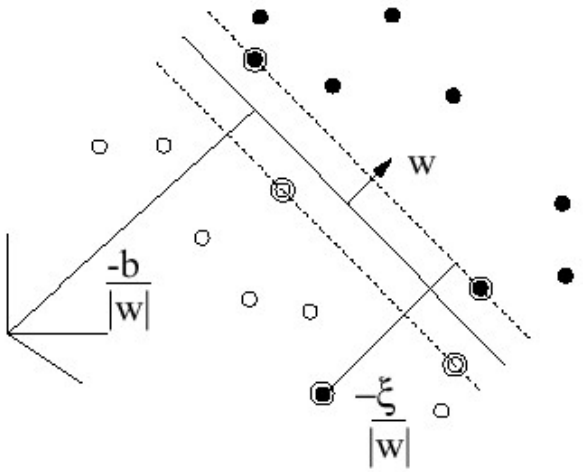


Figura B.2: Hiperplano separando los datos

de entrada no es lineal, es necesario hacer la transformación de los datos basándose en el producto interno para mapearlos en el espacio euclidiano $\mathcal{H}$, de manera que [71]:

$$\Phi : \mathbb{R}^n \rightarrow \mathcal{H} \tag{B.16}$$

Luego, el algoritmo de entrenamiento, solo depende de los datos a través de los productos punto de la forma $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. En este caso, se tiene una función K llamada kernel definida como

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

De manera que solo es necesario reemplazar el anterior kernel en el algoritmo de entrenamiento (ecuación (B.12)).

El *kernel* empleado en el presente trabajo corresponde al más utilizado (RBF - *Radial Basis Function*) definido como [72]:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2) \tag{B.17}$$

Las máquinas de soporte vectorial (SVM), son un tipo de algoritmo relativamente nuevo con un alto nivel de desempeño introducido por V. Vapnik [68]. Usualmente, la gran capacidad de generalización de las SVM es explicada por la existencia de un *gran margen*: fronteras sobre la tasa del error de un hiperplano que separa los datos con algún margen [73]. De hecho, para problemas con un número de muestras reducido, esta técnica de clasificación ha mostrado mejores resultados que otras, debido a que su esquema de optimización depende de un margen y no de una superficie de error o espacio, esto es, las máquinas de soporte vectorial siempre encuentran un mínimo global [71].

La máquina de soporte vectorial usada en este trabajo es *C-SVM* entrenada con el método de descomposición *SMO* (Sequential Minimal Optimization) [74] y utilizando un *kernel* gaussiano (Radial Basis Function). Se utilizó la librería de funciones para máquinas de soporte vectorial **LIBSVM** [75] para el desarrollo de este esquema de clasificación.

Apéndice C
Tablas de Resultados

1.000	-0.287	-0.211	-0.679	0.109	0.081	-0.021	0.118	0.029	0.104	0.235
-0.287	1.000	0.068	0.472	-0.100	-0.100	-0.037	0.024	-0.224	-0.098	-0.234
-0.211	0.068	1.000	0.193	0.126	0.221	0.081	0.133	0.036	0.190	0.269
-0.679	0.472	0.193	1.000	-0.082	-0.066	0.054	-0.088	-0.034	0.039	-0.228
0.109	-0.100	0.126	-0.082	1.000	-0.304	0.261	0.785	-0.543	0.221	0.047
0.081	-0.100	0.221	-0.066	-0.304	1.000	0.024	-0.180	0.520	-0.026	0.791
-0.021	-0.037	0.081	0.054	0.261	0.024	1.000	0.204	0.001	0.367	-0.002
0.118	0.024	0.133	-0.088	0.785	-0.180	0.204	1.000	-0.715	0.193	0.056
0.029	-0.224	0.036	-0.034	-0.543	0.520	0.001	-0.715	1.000	0.154	0.449
0.104	-0.098	0.190	0.039	0.221	-0.026	0.367	0.193	0.154	1.000	0.122
0.235	-0.234	0.269	-0.228	0.047	0.791	-0.002	0.056	0.449	0.122	1.000

Tabla C.1: Matriz de correlación para CA1 tomando $/a/$

1.000	-0.239	-0.217	-0.598	0.065	0.127	0.004	0.089	0.062	-0.037	0.224
-0.239	1.000	0.206	0.386	0.000	-0.035	0.015	-0.015	-0.055	0.025	-0.035
-0.217	0.206	1.000	0.329	0.017	0.268	0.024	0.050	0.035	0.050	0.244
-0.598	0.386	0.329	1.000	-0.006	-0.180	0.077	-0.040	-0.068	0.077	-0.242
0.065	0.000	0.017	-0.006	1.000	-0.221	0.437	0.271	0.042	0.013	0.151
0.127	-0.035	0.268	-0.180	-0.221	1.000	-0.156	-0.046	0.239	0.154	0.782
0.004	0.015	0.024	0.077	0.437	-0.156	1.000	0.121	0.005	0.009	0.008
0.089	-0.015	0.050	-0.040	0.271	-0.046	0.121	1.000	-0.692	0.111	0.047
0.062	-0.055	0.035	-0.068	0.042	0.239	0.005	-0.692	1.000	-0.022	0.322
-0.037	0.025	0.050	0.077	0.013	0.154	0.009	0.111	-0.022	1.000	0.104
0.224	-0.035	0.244	-0.242	0.151	0.782	0.008	0.047	0.322	0.104	1.000

Tabla C.2: Matriz de correlación para CA1 tomando todas las vocales como muestras independientes.

1.000	-0.045	-0.058	-0.453	-0.146
-0.045	1.000	-0.161	0.041	0.247
-0.058	-0.161	1.000	-0.386	0.373
-0.453	0.041	-0.386	1.000	-0.298
-0.146	0.247	0.373	-0.298	1.000

Tabla C.3: Matriz de correlaciones para CA2 tomando $/a/$

1.000	0.197	-0.027	-0.272	-0.024
0.197	1.000	-0.186	0.021	0.160
-0.027	-0.186	1.000	-0.135	0.019
-0.272	0.021	-0.135	1.000	-0.440
-0.024	0.160	0.019	-0.440	1.000

Tabla C.4: Matriz de correlaciones para CA2 tomando todas las vocales como muestras independientes

1.000	0.917	0.882	0.855	0.667	0.660	-0.016	-0.028	-0.181	-0.171	0.223	0.246
0.917	1.000	0.855	0.870	0.705	0.700	0.069	0.060	-0.167	-0.160	0.257	0.287
0.882	0.855	1.000	0.965	0.821	0.814	0.061	0.055	-0.086	-0.077	0.231	0.252
0.855	0.870	0.965	1.000	0.846	0.844	0.131	0.127	-0.048	-0.039	0.249	0.283
0.667	0.705	0.821	0.846	1.000	0.986	0.319	0.315	-0.089	-0.089	0.172	0.192
0.660	0.700	0.814	0.844	0.986	1.000	0.337	0.335	-0.064	-0.065	0.164	0.184
-0.016	0.069	0.061	0.131	0.319	0.337	1.000	0.998	0.599	0.595	0.321	0.337
-0.028	0.060	0.055	0.127	0.315	0.335	0.998	1.000	0.610	0.607	0.323	0.338
-0.181	-0.167	-0.086	-0.048	-0.089	-0.064	0.599	0.610	1.000	0.998	0.420	0.403
-0.171	-0.160	-0.077	-0.039	-0.089	-0.065	0.595	0.607	0.998	1.000	0.415	0.400
0.223	0.257	0.231	0.249	0.172	0.164	0.321	0.323	0.420	0.415	1.000	0.977
0.246	0.287	0.252	0.283	0.192	0.184	0.337	0.338	0.403	0.400	0.977	1.000

Tabla C.5: Matriz de correlaciones para CWT tomando $/a/$

1.000	0.865	0.723	0.719	0.544	0.546	0.385	0.383	0.062	0.058	0.201	0.213
0.865	1.000	0.787	0.813	0.680	0.693	0.504	0.506	0.124	0.115	0.223	0.235
0.723	0.787	1.000	0.969	0.821	0.819	0.591	0.588	0.185	0.174	0.279	0.301
0.719	0.813	0.969	1.000	0.843	0.853	0.619	0.622	0.207	0.196	0.271	0.294
0.544	0.680	0.821	0.843	1.000	0.980	0.679	0.677	0.214	0.201	0.257	0.271
0.546	0.693	0.819	0.853	0.980	1.000	0.694	0.698	0.241	0.232	0.248	0.264
0.385	0.504	0.591	0.619	0.679	0.694	1.000	0.992	0.583	0.567	0.454	0.465
0.383	0.506	0.588	0.622	0.677	0.698	0.992	1.000	0.581	0.567	0.444	0.454
0.062	0.124	0.185	0.207	0.214	0.241	0.583	0.581	1.000	0.994	0.489	0.510
0.058	0.115	0.174	0.196	0.201	0.232	0.567	0.567	0.994	1.000	0.468	0.499
0.201	0.223	0.279	0.271	0.257	0.248	0.454	0.444	0.489	0.468	1.000	0.970
0.213	0.235	0.301	0.294	0.271	0.264	0.465	0.454	0.510	0.499	0.970	1.000

Tabla C.6: Matriz de correlaciones para CWT tomando todas las vocales como muestras independientes

Bibliografía

[1] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, Dekker, New York, 1989. 1, 2

[2] Castellanos G. Vargas, F., "Clasificación automatizada de las características acústicas de la voz normal en la ciudad de manizales.," Trabajo de grado, Universidad Nacional de Colombia - Sede Manizales, Sept 2001. 3, 8, 12, 26, 46, 47, 48, 52, 59

[3] J. Gurlekian, *El hombre dialoga con la máquina.*, Buenos Aires, 1986. 4, 47

[4] J. Menaldi, *La Voz Normal.*, Panamericana - Argentina, 1992. 4, 5, 47, 48

[5] James L. Hieronymus, "Ascii phonetic symbols for the world's languages: Worldbet," Tech. Rep., AT&T Bell Laboratories, Murray Hill, USA, 1994. 6

[6] Ramishvili G.S., *Identificación automatizada del hablante por voz (Rus).*, Radio y Svjaz. Moscu, 1981. 8, 24, 25

[7] F. Le Huche and A. Allali, *Patólogía Vocal: Semiología y disfonías disfuncionales*, vol. 2 of *LA VOZ*, MASSON, 1994. 8, 13

[8] Jhon Hansen, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment, *IEEE transactions on Biomedical Engineering*, vol. 45, Marzo 1998. 9

[9] B. Boyanov, Hadjitodorov S., and Baudoin G., "Acoustic analysis of patological voices," Tech. Rep., Center for Biomedical Engineering, Julio 1997. 10, 46, 47

[10] Bielamovicz S. Parsa, V., "Comparison of voice analysis systems for perturbation measurement., *Journal of Speech and Hearing*, vol. 39, pp. 126–134, Feb. 1996. 11, 48

[11] J. Sundberg, "Perceptual aspects of singing., *Journal of Voice: Official journal of the voice foundation*, vol. 8, no. 2, pp. 106–122, Jun 1994. 12

[12] Betancourth C. Vitola, F., "Reconocimiento de palabras aisladas en español mediante técnicas de predicción lineal y alineamiento temporal," Trabajo de grado, Universidad Nacional de Colombia - Sede Manizales, Jul 1999. 15, 21, 25, 28, 30, 33

[13] Andrew. Tanenbaum, *Redes de computadoras.*, Pearson., 1997, Tercera Edición. 17

[14] Villegas. E. Castellanos. G., "Comparación de métodos de compresión en el análisis acústico de voz sobre la web.," in *II Congreso Internacional de Telemática*. Instituto Superior Politécnico José Antonio Echeverría. La Habana, Cuba., Nov 2002, vol. 1, Instituto Superior Politécnico José Antonio Echeverría. 17

[15] A. Vargas and A. Duque, "Análisis psicoacústico para la codificación de audio," Tech. Rep. 18

[16] Juan B. MANSOUR D., "A family of distortion measures based upon projection operation for robust speech recognition, *IEEE TRANS on Speech and Signal Processing*, vol. 37, 1989. 19

[17] Yoshiaki. OHSHIMA, "Environmental robustness in speech recognition using physiologically motivated signal processing. phd thesis.," Tech. Rep., University. Pittsburgh, Pennsylvania., 1993. 20

[18] Hiroshi SARUWATARI, "Speech enhancement using nonlinear microphone array based on complementary beamforming., *IEICE Trans, Fundamentals.*, vol. E00-A, no. 1, Jan 1999. 20

[19] Hansen J. Deller J., Proakis J., *Discrete-Time Processing of Signals.*, Prentice Hall, 1993. 20, 25, 43, 47

[20] Simon DOCLO, "Novel iterative signal enhancement algorithm for noise reduction in speech.," Tech. Rep., Katholieke Universiteit Leuven, Kardinaal. Belgium, 1994. 20

[21] Maurizio. OMOLOGO, "On the future trends of hands-free asr: Variabilities in the enviromental condition and in the acoustic transduction.," Tech. Rep., 1994. 21

[22] Maurizio. OMOLOGO, "Enviromental conditions and acoustic transduction in hands-free speech recognition.," Tech. Rep., 1996. 21

[23] N.Ñavarro, A. Lopez, and G. Castellanos, "Tesis diseño y desarrollo del analizador acústico computarizado de voz.," Trabajo de grado, Universidad Nacional de Colombia-Sede Manizales, 2000. 21, 47

[24] Jane CHANG and Victor. ZUE, "A study of speech recognition system robustness to microphone variations.," Tech. Rep., Massachusetts Institute Of Tecnology. Cambride, Massachusetts., 1996. 21

[25] A. Papoulis, *The Fourier Integral and Its Applications*, McGraw-Hill, 2ª edition, 1987. 34

[26] D. Gabor, "Theory of communication, *Journal IEE*, pp. 429–457, 1946. 36

[27] MathWorks, *Wavelet Toolbox: User's Guide - Version 2. For Use with MATLAB*, The MathWorks, Inc., Natick, MA, 2000. 38

[28] Yves. Meyer, "Wavelets: Algorithms and applications, *Society for Industrial and Applied Mathematics*, 1993. 39, 40

[29] Ingrid Daubechies, "Ten lectures on wavelets. conference board of the mathematical society, national science foundation (CBMS-NSF), *Society for Industrial and Applied Mathematics*, 1992. 40

[30] D. Michaelis, M. Frohlich, and H. Strube, "Selection and combination of acoustic features for the description of pathologic voices," Tech. Rep., Drittes Physikalisches Institut, 1998. 42, 49

[31] Y. Mallet, D. Coomans, J. Kaustsky, and O. De Vel, "Classification using adaptive wavelets for feature extraction, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 10, pp. 1058–1066, 1997. 43, 53

[32] S. Pittner and S.V. Kamarthi, "Feature extraction from wavelete coefficients for pattern recognition tasks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 1, pp. 83–88, 1999. 43, 53

[33] Janer García, "Transformada wavelet aplicada a la extracción de información en señales de voz," Tech. Rep., Universidad Politécnica de Cataluña, 1998. 43, 45

[34] Ricardo Alzate, "Estimación del pitch en tiempo real orientado al aav," Trabajo de grado, Universidad Nacional de Colombia - Sede Manizales, Sept 2003. 44, 45

[35] Blalock P. Koufman, J, "Classification and approach to patients with functional voice disorder," Publications related to the larynx, voice and voice disorders, Center For Voice Disorders of Wake Forest University, 1982. 46

[36] A. Wendt, C.and Petropulu, "Pitch determination and speech segmentation using the discrete wavelet transform, *IEEE International Symposium on Circuits and Systems*, vol. 2, pp. 45–48, 1996. 46

[37] Th. Parson, *Voice and speech processing.*, McGraw Hill - New York, 1987. 47

[38] P. Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges, *J. Acoust. Soc. Am.*, vol. 35, pp. 344–353, 1963. 48

[39] Omar D. Castrillón, F. Vargas, J. F. Suárez, G. Castellanos , "Comparación de algoritmos de estimación del pitch en el análisis acústico de la voz normal y patológica," in *VII Simposio de Tratamiento de Señales, Imagenes y Vision Artificial*, Universidad Industrial de Santander, Ed. Sociedad Colombiana de Tratamiento de Señales, Noviembre 2002. 49

[40] Michaelis, D.; Frohlich, M., "Empirical study to test the independence of different acoustic voice parameters on a large voice database, *University of Gottingen*, p. 15, 2000. 50

[41] Childers D., *Speech Processing & Synthesis Toolboxes,* John Wiley and Sons, 1ra edition, 1999. 50

[42] Paul Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," 2000. 51, 52

[43] M. Krishnan, C. Neophytou, and G. Prescott, "Wavelet transform speech recognition using vector quantization, dynamic time warping and artificial neural networks," 1994. 54

[44] B. T. Tan, M. Fu, A. Spray, and P. Dermody, "The use of wavelet transforms in phoneme recognition," in *Proc. ICSLP '96*, Philadelphia, PA, 1996, vol. 4, pp. 2431–2434. 54

[45] C. Long and S. Datta, "Wavelet based feature extraction for phoneme recognition," in *Proc. ICSLP '96*, Philadelphia, PA, 1996, vol. 1, pp. 264–267. 54

[46] F. Ojeda, "Extracción de características usando transformada wavelet en la identificación de voces patológicas," Trabajo de grado, Universidad Nacional de Colombia - Sede Manizales, Sept 2003. 54, 77, 86

[47] Ingrid Daubechies, "Orthogonal bases of compactly supported wavelets, *Comm. Pure Appl. Math.*, vol. 41, no. 7, pp. 909–996, 1988. 54

[48] Mallat. S., *A wavelet tour of signal processing,* Academic press, 1997. 55

[49] Botero L.M. Vargas. F., Castellanos G, "Análisis acústico en la clasificación de señales de voz empleando rna.," in *Congreso Internacional de Inteligencia Computacional*. Universidad Nacional de Colombia - Sede Medellin, Sept 2001, vol. 1, Universidad Nacional de Colombia - Sede Medellin. 57

[50] M. Partridge and R. Calvo, "Fast dimensionality reduction and simple pca," 1998. 57

[51] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao, "Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000. 58

[52] E. L'VOVSKI, *Métodos estadísticos de construcción de fórmulas empíricas*, VSh. Moscú, 1988. 58

[53] I. Doltsinis, F. Rau, and M. Werner, *Stochastic Analysis of Multivariate Systems in Computational Mechanics and Engineering*, chapter Analysis of Random Systems, pp. 9–159, International Center for Numerical Methods in Engineering, first edition, September 1999. 61

[54] Miguel A. Carreira-Perpinan, "A Review of Dimension Reduction Techniques," Tech. Rep. CS–96–09, Dept. of Computer Science, University of Sheffield, January 1997. 62, 63

[55] B. Sch, o Smola, and K. uller, "Nonlinear component analysis as a kernel eigenvalue problem," 1998. 64, 65, 67

[56] S. Mika, "Kernel fisher discriminants," Trabajo de grado, Universidad Tecnológica de Berlin, Dec. 2002. 69

[57] F. Masulli and G. Valentini, "Mutual information methods for evaluating dependence among outputs in learning machines," 2001. 71

[58] K. Bollacker and J. Ghosh, "Linear feature extractors based on mutual information," 1996. 73

[59] Wendy L. Martinez and Angel R. Martinez, *Computational Statistics Handbook with* MATLAB, Chapman & Hall/CRC, 2002. 87

[60] Y. Lin G. Wahba and H. Zhang, *Generalized Approximated cross-validation for Support Vector Machines: Another way to look at margin like quantities. In A. Smola, P. Bartlett, B. Schölkopf and D.Schuurmans, Advances in Large Margin Classifiers*, MIT Press, Cambridge, MA, 2000. 88

[61] Cullen Schaffer, "Selecting a classification method by cross-validation," Tech. Rep., Department of Computer Science CUNY/Hunter College, March 1993. 88

[62] Stéphane G. Mallat, "Multiresolution aproximations and wavelet orthonormal bases of $\mathcal{L}^2(\mathcal{R})$, *Trans. Amer. Math. Soc*, vol. 315, no. 1, pp. 69–87, 1989. 103

[63] Stéphane G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989. 103

[64] R. O. Duda, P. E Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, New York, second edition, 2001. 105

[65] MathWorks, *Neural Network Toolbox: User's Guide - Version 4. For Use with MATLAB*, The MathWorks, Inc., Natick, MA, 2000. 106, 109

[66] R. Battiti, "First adn second order methods for learning: Between steepest descent and newton's method, *Neural Computation*, vol. 4, no. 2, pp. 141–166, 1992. 108

[67] Martin T. Hagan adn M. B. Menhaj, "Training feedforward networks with the marquardt algorithm, *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, November 1994. 108, 109

[68] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, NY, 1995. 109, 110, 112

[69] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition, *Knowledge Discovery and Data Mining*, vol. 2, pp. 22, 1998. 110

[70] P. Wolfe, "The simplex method for quadratic programming, *Econometrica*, vol. 27, pp. 382–398, 1959. 110

[71] Bernhard Schölkopf and Alex Smola, *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*, MIT Press, Cambridge, 2002. 111, 112

[72] Bernhard Schölkopf, K. Sung, C. Burges, F. Girosi, P.Ñiyogi, T. Poggio, and V. Vapnik, "Comparing support vector machines with gaussian kernels to radial basis function classifiers, *URL: http://citeseer.nj.nec.com/63569.html*, 1996. 112

[73] P. Bartlett and J. Shawe-Taylor, "Generalization performance of support vector machines and other pattern classifiers," 1998. 112

[74] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998. 112

[75] Chih-Chung Chang and Chih-Jen Lin, "**LIBSVM: a library for support vector machines," Tech. Rep., National Taiwan University, Taipei, March 2003. 112**

[76] E. Wesfried, "Adapted local trigonometric transforms and speech processing, *IEEE on Signal Processing*, vol. 41, 1993.

[77] J. B. Allen, "Cochelar modeling, *IEEE ASSP Magazine*, pp. 3–29, 1985.

[78] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing*, Prentice Hall, Upper Saddle River, New Jersey, 2001.

[79] T. Grewin, C. Ryden, "Subjective assessments on low bit-rate audio codecs," in *10th AES convention. London*, 1991, vol. 1.