DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN ELECTRÓNICA Y TELECOMUNICACIONES

TEMA: DISEÑO DE UN SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DE EMOCIONES A PARTIR DEL ANÁLISIS DE LA SEÑAL DE VOZ.

AUTOR: FLORES CRUZ, EVELYN MARLEY

DIRECTOR: ING. BERNAL OÑATE, CARLOS PAÚL MSc.

SANGOLQUÍ

2019

CERTIFICADO DEL DIRECTOR

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**DISEÑO DE UN SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DE EMOCIONES A PARTIR DEL ANÁLISIS DE LA SEÑAL DE VOZ**", fue realizado por la señorita **Flores Cruz, Evelyn Marley** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 24 de junio del 2019

Firma:

Ing. Carlos Paúl Bernal Oñate, Msc.

C. I: 1709775637

AUTORÍA DE RESPONSABILIDAD

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORÍA DE RESPONSABILIDAD

Yo, **Flores Cruz, Evelyn Marley**, declaro que el contenido, ideas y criterios del trabajo de titulación: "**Diseño de un sistema de reconocimiento automático de emociones a partir del análisis de la señal de voz"** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 24 de junio del 2019

Firma:

.....

Evelyn Marley Flores Cruz

C.C. 0503963357

AUTORIZACIÓN

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORIZACIÓN

Yo, **Flores Cruz, Evelyn Marley**, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: "**Diseño de un sistema de reconocimiento automático de emociones a partir del análisis de la señal de voz**", en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 24 de junio del 2019

Firma:

.....

Evelyn Marley Flores Cruz

C.I. 0503963357

DEDICATORIA

Deseo dedicar esta tesis a:

A mis padres y hermanas, Leonidas, Carmen, Lorena y Vanessa, quienes con su amor, paciencia y esfuerzo me han permitido llegar a cumplir hoy un sueño más, gracias por inculcar en mí el ejemplo de esfuerzo y valentía, de no temer las adversidades porque Dios está conmigo siempre.

A Santiago, por su cariño y apoyo incondicional, durante todo este proceso, por estar conmigo en todo momento, gracias.

A todas las personas especiales que me acompañaron en esta etapa, aportando a mi formación tanto profesional y como ser humano.

Evelyn

AGRADECIMIENTO

A mi familia, por haberme dado la oportunidad de formarme en esta prestigiosa Universidad y haber sido mi apoyo durante todo este tiempo.

De manera especial a mi tutor de tesis, por haberme guiado, no solo en la elaboración de este trabajo de titulación, sino a lo largo de mi carrera universitaria y haberme brindado el apoyo para desarrollarme profesionalmente y seguir cultivando mis valores.

De igual manera mis agradecimientos a la Universidad de las Fuerzas Armadas ESPE, por haberme brindado tantas oportunidades y enriquecerme en conocimiento.

Evelyn

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

A lo largo del tiempo las emociones han representado un elemento inherente de los seres vivos, mediante una expresión emocional los seres humanos pueden expresar cualquier acción, sentimiento o información de manera implícita y natural. El presente proyecto de investigación tiene como finalidad la identificación de emociones a partir de la señal de voz, utilizando el software Matlab®, mediante la teoría de *Machine Learning* a través de la técnica de clasificación supervisada. Para la detección de emociones se trabajó con dos bases de datos que contienen un total de 312 audios repartidos equitativamente entre hombre y mujer. Con el objetivo de diferenciar cuatro emociones fundamentales felicidad, enojo, miedo y tristeza; se evaluó las variaciones de un conjunto de características tales como, Entropía, Energía entre otras, logrando un total de 68 características obtenidas de dos maneras diferentes, señal de voz sin ningún preprocesamiento y mediante la Transformada *Wavelet*; en cada una de ellas se ejecutó un análisis descartando las características que no presentaban gran relevancia para el estudio mediante métodos de selección características propias para cada modelo de clasificación. Obteniendo así un total de 53 características para hombres y 57 para mujeres, las cuales fueron utilizadas para la detección automática de emociones. Lo resultados fueron analizados bajo cuatro parámetros que son exactitud, precisión, sensibilidad y especificidad.

PALABRAS CLAVE:

- **TRANSFORMADA WAVELET**
- **MACHINE LEARNING**
- **RECONOCIMIENTO DE EMOCIONES**

ABSTRACT

Throughout time emotions have represented an inherent element of living beings, through an emotional expression human being can express any action, feeling or information implicitly and naturally. The purpose of this research project is to identify emotions from the voice signal, using the Matlab® software through Machine Learning theory through the supervised classification technique. For the detection of emotions, I worked with two databases containing a total of 312 audios distributed equally between men and women. With the objective of differentiating four fundamental emotions: happiness, anger, fear and sadness; the variations of a set of characteristics such as Entropy, Energy, among others, were evaluated, achieving a total of 68 characteristics obtained in two different ways, voice signal without any preprocessing and through the Wavelet Transform; in each one of them an analysis was carried out discarding the characteristics that did not present great relevance for the study by means of selection methods characteristic of each classification model. Obtaining a total of 53 characteristics for men and 57 for women, which were used for the automatic detection of emotions. The results were analyzed under four parameters that are accuracy, precision, sensitivity and specificity.

KEYWORDS:

- **TRANSFORMED WAVELET**
- **MACHINE LEARNING**
- **RECOGNITION OF EMOTIONS**

CAPÍTULO I

INTRODUCCIÓN DEL PROYECTO DE INVESTIGACION

1. Introducción del proyecto de Investigación

1.1 Antecedentes y justificación del proyecto

A lo largo del tiempo las emociones han representado un elemento inherente de los seres vivos, mediante una expresión emocional los seres humanos pueden expresar cualquier acción, sentimiento o información de manera implícita y natural. Los filósofos y psicólogos se interesaron inicialmente en el estudio de las expresiones emocionales de las personas. Platón manifestó "La filosofía del alma tripartita está compuesta por emoción, cognición y motivación". De igual manera Charles Darwin en su obra "La expresión de las emociones en los animales y el hombre", explica sobre las emociones y su relación con la selección natural (Chóliz, 1995).

El reconocer el estado emocional de una persona ha servido para muchas aplicaciones alrededor del mundo tal es el caso de la creación de un Sistema de Respuesta Interactiva por Voz (IVR) que se basa principalmente en un sistema que detecte problemas de depresión en pacientes con trastornos psicológicos, mediante las características de la calidad de voz de la persona. Este sistema alerta a un doctor en caso de que el individuo presente un grado alto de depresión (G.M., 1999).

Este tipo de aplicaciones se usaron también para realizar un control de conversaciones entre servidores y usuarios en entidades que se basan principalmente en trabajos de alto estrés lo que provoca que los individuos lleguen a un estado de agotamiento, para lo cual se monitorizan los estados emocionales de las personas ya sea de un cliente enfadado o un mal trato por parte del

agente a los usuarios, y de esta manera mejorar la calidad de los servicios y del personal administrativo (Devillers, et al., 2006).

En el Instituto Tecnológico y de Estudios Superiores de Monterrey, un grupo de egresados han realizado una investigación basada en el uso de un modelo de comportamiento afectivo, el cual detectaría automáticamente el estado de ánimo y pedagógico del alumno para posteriormente darle tutorías inteligentes que motiven y capten el interés del estudiante para aprender matemáticas, historia, entre otras (Hernández, et al., 2008).

Uno de los estudios implementados actualmente en *call centers* de Estados Unidos sirve para detectar automáticamente emociones de personas que llamen pidiendo ayuda y requieran atención médica, la mayoría de los estudios anteriores a este, han utilizado bases de datos de emociones emitidas por actores profesionales, este es un caso singular en que las emociones de los individuos son naturales y espontáneas. Las personas que llamen a pedir ayuda podrían presentar diferentes emociones tales como ira, dolor, tensión, ansiedad, y dependiendo el caso y la gravedad de la llamada se dirigirá al médico indicado. Se ha obtenido una tasa de detección correcta de emociones del 80%, durante veinte horas de grabación continua (Vidrascu, et al., 2005).

En el Departamento de Salud del Instituto Tecnológico de Georgia, se realizó una investigación acerca del papel que tienen las características glóticas en la clasificación de la depresión clínica en personas que sufren un trastorno mental, dicho estudio se basa en un modelo continuo que tiene como base 3 categorías: prosódica, espectro del tracto vocal y fuente global, este tipo de características aborda la problemática que se tiene al pasar de una base de datos actuada a una espontánea (Elliot More, 2008).

El presente trabajo de investigación que trata sobre el reconocimiento de emociones a partir de la voz busca dar un apoyo a entidades que requieran conocer el estado anímico de una persona tan solo con una grabación de voz, ya que esta puede ser utilizada por psicólogos y centros de llamada y de forma remota y de esta manera mejorar por una parte la calidad de vida de las personas con trastornos emocionales y por otro lado optimizar el servicio al cliente que brinda la entidad.

1.2 Objetivos de la investigación

1.2.1 General

Implementar un sistema de reconocimiento automático de emociones a partir de la señal de voz.

1.2.2 Específicos

- Identificar mediante el estado del arte las principales características acústicas de la señal de voz usadas hasta el momento.
- Estudiar la viabilidad entre las diferentes trasformaciones Tiempo-Frecuencia para la extracción de características prosódicas, calidad de la voz y espectrales; de cada persona presente en la base de datos actuada.
- Entrenar el sistema mediante la teoría de Machine Learning a través de la técnica de clasificación supervisada.
- Analizar, comparar, evaluar y corregir el sistema con las bases de datos aplicadas, a través de Machine Learning.

CAPÍTULO II

MARCO TEÓRICO

2. Marco teórico

2.1 Anatomía del habla

La voz de los seres humanos es producto de una necesidad de comunicación, que se ejecuta en el aparato fonador. La voz es una señal formada por la comprensión de moléculas y la aplicación de energía que se comunican de forma paralela.

El lenguaje verbal aun cuando es exclusivo de los seres humanos no comienza con las primeras palabras de un bebé, es el lenguaje que se va aprendiendo con el crecimiento de la persona, en el cual hay un entrenamiento y un aprendizaje progresivo (Castañeda, 2015).

2.1.1 Elementos del habla

2.1.1.1 Cavidades infraglóticas

En primer lugar, existe una corriente de aire, que es producida por los músculos respiratorios y los pulmones. Las cavidades infraglóticas se forman por los bronquios, pulmones y tráquea. Al momento que los pulmones se llenan de aire, el diafragma se comprime haciendo que el aire sea expelido con fuerza hacia el exterior y provocando el ritmo necesario para la creación de fonemas. El preciso instante de la espiración los músculos intercostales se contraen provocando la energía necesaria para que la onda acústica atraviese los órganos fonadores superiores (Arsuaga, 2000).

2.1.1.2 Cavidad glótica o laríngea

La principal función de la cavidad laríngea es la producción de los sonidos vocálicos. El efecto conjunto de la presión infraglótica y la tensión de las cuerdas permite emitir hacia el exterior una

señal audible. Las cuerdas vocales están formadas por cuatro cartílagos, cicroides que es la base en forma de anillo, tiroides que tiene una forma de escudo y las dos aritenoides que permite una gran movilidad de las cuerdas vocales.

Si las cuerdas vocales se encuentran tensas en el aire y juntas, estas chocan produciendo diferentes sonidos y cada una tiene propio espectro como la “a” y la “u” tienen su primer y segundo armónico débil y su tercer armónico más fuerte, en cambio en las vocales “e” y “o” sucede lo contrario su primer y segundo armónico es más fuerte que su tercer armónico. La mayoría de las vocales poseen armónicos transitorios. (Castañeda, 2015)

2.1.1.3 Cavidades supraglóticas

Las cavidades supraglóticas están formadas por la cavidad nasal (nariz), cavidad oral (boca) y cavidad faríngea (garganta). Todas estas cavidades que están por encima y por debajo de la laringe entran en resonancia con la vibración de las cuerdas vocales. El sistema resonante permite amplificar los sonidos de la voz que son emitidos hacia el exterior.

Dentro del tracto vocal existen lugares de articulación que permiten emitir los sonidos hacia al exterior, como los labios, dientes, paladar duro, velo del paladar y mandíbula, los cuales provocan que la voz humana sea única ya que permiten añadir intensidad, emoción y profundidad al momento de vocalizar algún fonema, sílaba o palabra.

El mecanismo de producción de voz se puede observar en la Figura 1, el cual describe claramente las cavidades infraglóticas, glóticas y supraglóticas.

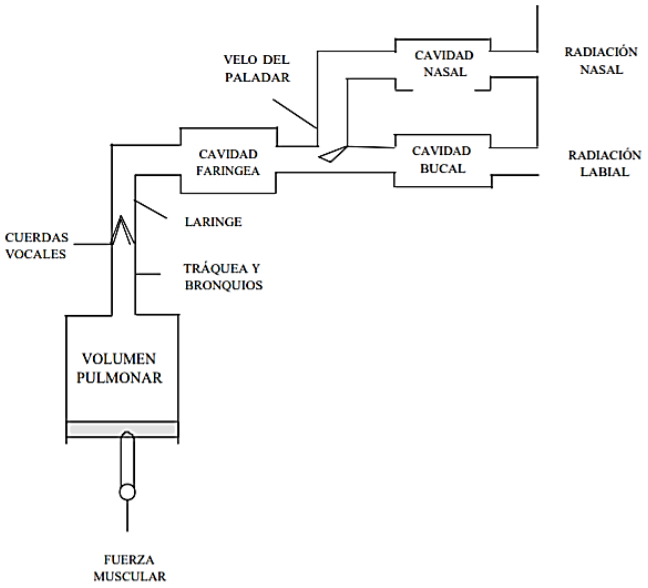

Figura 1. Esquema del mecanismo de producción de voz
Fuente: (Gonzáles, 2013)

2.2 Fonética y Fonología

2.2.1 Fonética

"La fonética es el estudio de los sonidos físicos del discurso humano. Es la rama de la lingüística que estudia la producción y percepción de los sonidos de una lengua específica, con respecto a sus manifestaciones físicas" (Concepto fundamentales de fonología, 2010).

2.2.2 Fonología

La fonología es un subcampo de la lingüística que describe el modo en que los sonidos funcionan en un nivel abstracto o mental.

2.3 Sonidos

2.3.1 Sonoros

Si las cuerdas vocales vibran se llaman sonidos sonoros, provocando un tren de pulsos cuasi-periódico. Esto sucede al pronunciar las letras /b/, /d/, /a/, /l/, /m/, /n/, /y/, /g/ del lenguaje anglosajón (Gaya, 1999).

2.3.2 Sordos

Se llaman sonidos sordos a la excitación generada mediante una oclusión en el tracto vocal que provoca una relajación para que las cuerdas vocales no vibren. Por ejemplo, al pronunciar /p/, /t/, /k/, /ch/, /z/, /s/, /j/, /f/ del lenguaje anglosajón (Duque, 2007).

2.3.3 Nasales y orales

Se llaman sonidos plosivos a aquellos que al momento de producirlos el aire pasan por la cavidad nasal. Si gran parte del aire pasa por cavidad bucal se llaman sonidos orales (Gaya, 1999).

2.4 Clasificación de las emociones

2.4.1 Emociones básicas

Las emociones primarias o básicas son muy fáciles de reconocer por la voz se inician con rapidez y duran unos pocos segundos. Las emociones primarias están grabadas en los circuitos nerviosos, estas emociones no muestran un estado emocional permanente de la persona, si no es una respuesta fisiológica causada por el cerebro (Vivas, 2001).

Existen cuatro emociones básicas las cuales poseen también los mamíferos superiores, cada una de ellas representa una variedad de representaciones:

2.4.1.1 Ira

La ira es una emoción que varía según la intensidad de la situación de la que se presente. La ira provoca un aumento en la frecuencia cardiaca y la presión arterial. La ira es la forma en la que el cuerpo nos pone en alerta frente a los problemas o situaciones que producen frustración o resultan desagradables (Arroyo, 2013). La ira se caracteriza por tener un tono medio alto de 229Hz, en promedio una persona tiene una velocidad de elocución de 190 palabras por minuto, con un 32% de pausas (Duque, 2007).

2.4.1.2 Alegría

La alegría es una emoción positiva que las personas expresan cuando han logrado una meta o por algo que le causa bienestar. Se manifiesta con un incremento en su frecuencia fundamental entre 200Hz y 300Hz para el caso de hombre y para el caso de mujer su frecuencia fundamental entre 200Hz y 350Hz (Duque, 2007).

2.4.1.3 Tristeza

Es un sentimiento de dolor y un signo de debilidad que muestran las personas mediante su voz, es una emoción necesaria que nos permite reorganizar y superar eventos traumáticos (Bericat). La emoción tristeza muestra un tono medio más bajo que la felicidad para el caso de hombres presenta una variación entre 180Hz a 220Hz y para la mujer su frecuencia fundamental varía entre 150Hz a 340Hz.

2.4.1.4 Miedo

La emoción del miedo ha presentado un tomo medio elevado con 254 Hz y una velocidad de elocución de 220 palabras por minuto. El miedo provoca en la persona un latido en su corazón

muy rápido que induce al individuo a huir. El miedo es una emoción muy importante ya que nos permite apartarnos de los peligros y actuar de manera inmediata frente a ellos (Mora, 2010).

2.4.2 Emociones secundarias

Son emociones secundarias ya que se tratan de expresiones emocionales que están conformadas por otras un poco más elementales. Las emociones secundarias se presentan en:

2.4.2.1 Pena

La pena es una emoción que incapacita a los individuos es una forma de expresar tristeza de manera extrema. Presenta un tono bajo medio, tiene una velocidad de locución baja y muestra un alto porcentaje de pausas (Duque, 2007).

2.4.2.2 Sorpresa

La sorpresa es una emoción provocada por algo imprevisto, la sorpresa puede ser negativa o positiva y prepara al individuo para que actúe de manera rápida frente al problema (Duque, 2007).

La sorpresa presenta una combinación con los sentimientos de confianza, ternura, simpatía y estima. Expresa un alto nivel de tono de elocución (Soler, 2010).

2.5 El habla y las emociones

A lo largo del tiempo el habla ha sido estudiada por diferentes investigadores con diferentes fines ya sea por efectos léxicos, prosódicos o psicológicos. Las emociones afectan a la voz y Charles Darwin fue la primera persona que manifestó esta afirmación en su obra “La expresión de las emociones en el hombre y en los animales”. Hoy en día sabemos que el tono, intensidad, vibración y demás características de la voz de las personas hablan de su estado de ánimo,

psicológico o mental y debido a que no es un método invasivo muchos doctores e investigadores alrededor del mundo están usando esta técnica para diagnosticar a sus pacientes incluso si no habla su mismo idioma (Petisco, 2010).

Las características más importantes en el habla para detectar las emociones destacadas en los diferentes documentos son:

- Frecuencia fundamental (pitch)
- Duración
- Calidad de la voz

2.5.1 Frecuencia fundamental (Pitch)

La Pitch se da gracias a la vibración de cuerdas, según los estudios realizados se ha determinado que la pitch o frecuencia fundamental es una característica de la voz que posee más información acerca de las emociones (Carmona J.).

El rango o distancia del pitch esta entre el valor máximo y mínimo de la frecuencia fundamental el cual muestra el grado de exaltación del locutor entre más extenso este se refleja una excitación emocional o psicológica. Si los cambios presentes en las velocidades del pitch son altos, suaves, bajos o abruptos son producidas psicológicamente, para emociones negativas como miedo o enfado, la curva de la frecuencia fundamental es discontinua y para emociones positivas tales como alegría, la curva del pitch es suave (Expresión de las emociones, 2010).

Según los estudios realizados por la Universidad Complutense de Madrid, los parámetros del pitch que se han obtenido para las diferentes emociones se muestran en la Tabla 1, en el cual las columnas muestran los parámetros tomados en cuenta para modelar las distintas emociones y las

filas indican los valores que tomarán los parámetros para las diferentes emociones (Gervás, 2010).

Tabla 1
Parámetros de la prosodia en las emociones

	Volumen	Velocidad (Pal/min)	Pitch	Rango Pitch
Neutral	0.9	120	100	11
Enfado	1	145	100	30
Sorpresa	1	120	125	20
Alegría	1	155	135	14
Tristeza	0.8	110	90	7
Miedo	1	135	175	24

Fuente (Gervás, 2010)

2.5.2 Duración de la voz

La duración está caracterizada por la velocidad del habla cuyos efectos son el ritmo de la voz. El ritmo en la voz de las personas radica en la combinación de las pausas y los fonemas. Una persona que tenga emociones positivas hablará rápidamente y de manera espontánea mientras que un individuo deprimido hablará más lento y con pausas mucho más largas. Un individuo en estado de exaltación acortará la duración de las sílabas y se tomará en cuenta las sílabas por segundo (Carmona J.).

2.5.3 Calidad de voz

Las principales características son:

2.5.3.1 Respiración o *Breathiness*

Esta característica del tracto vocal es propia de la voz del individuo. La respiración describe como la generación de ruido en el aparato respiratorio tiende a ser más fuerte, mientras que las frecuencias altas son reemplazadas con ruidos aspiratorios. Una voz estrepitosa presenta una vibración aperiódica de las cuerdas vocales y un pitch bajo (Expresión de las emociones, 2010).

2.5.3.2 Intensidad

Está característica refleja la amplitud de la onda de voz y la percepción del volumen de la señal (Petisco, 2010).

2.5.3.3 Irregularidades vocales

Entre las características vocales más importantes son *jitter* la cual refleja la fluctuación de un pulso glotal frente a diferentes emociones como el enfado o la desaparición de la voz en emociones como tristeza, pena (Expresión de las emociones, 2010).

2.5.3.4 Cociente de alta y baja frecuencia

Las emociones como enfado o ira presentan una alta frecuencia mientras una baja frecuencia presentan las emociones como tristeza o pena (Expresión de las emociones, 2010).

2.6 Procesamiento de la señal de voz

2.6.1 Transformadas tiempo-frecuencia (TFR)

Las transformadas tiempo-frecuencia son utilizadas cada vez más debido a que el uso de métodos espectrales clásicos no permite reflejar cambios en frecuencia con respecto al tiempo como lo hacen las transformadas tiempo-frecuencia, el uso combinado de estos dos dominios es muy efectivo para encontrar características útiles y fiables, en la actualidad se utiliza en todos los campos de tratamientos de señales digitales (Duque, 2007).

Una característica importante de las señales es la estacionariedad, ya que no varían en el tiempo y su análisis es más fácil, pero en la naturaleza la mayoría de las señales varían con respecto el tiempo.

2.6.2 TFR Cuadráticas

La transformada cuadrática trata acerca de la distribución de energía y esta transformada de igual manera recibe el nombre de "representaciones energéticas", las fórmulas que se presentan a continuación son:

Potencia instantánea:

$$p_s(t) = |s(t)^2| \quad (1)$$

Densidad espectral de energía:

$$P_s(f) = |S(f)^2| \quad (2)$$

Muchas veces la transformada cuadrática no significa realmente una distribución energética, en cualquier caso, que se presente esta transformada puede llegar a una aproximación de la señal analizada (Duque, 2007).

2.6.3 Transformada Gabor

En la transformada de Gabor, se usa una ventana temporal para el análisis en todas las bandas, esto significa que las variables de tiempo y frecuencia son constantes. Una de las principales desventajas que presenta es que esta transformada solo necesita señales estacionarias ya que si aparecen otras componentes de la frecuencia y no son múltiplos del tamaño de la ventana este se puede distorsionar, al utilizarse el mismo tamaño de las ventanas para frecuencias altas y bajas, esto puede ocasionar que se pierdan componentes de bajas frecuencias (Vuletich, 2005)

El uso de la transformada de Gabor para el procesamiento de señales acústicas resulta eficiente cuando se trata de caracterizar eventos con patrones de frecuencia bien definida y de largo tiempo

ya que los de corta de duración no tienen oscilaciones largas. En la Figura 2 se puede observar la transformada de Gabor a una señal de voz.

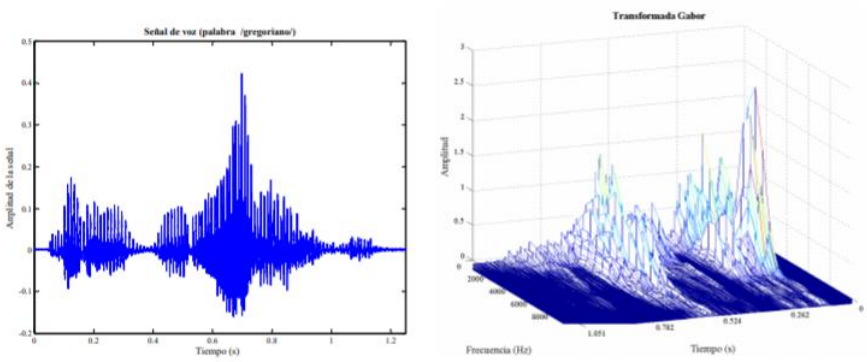

Figura 2. Proyección 3D de la transformada Gabor
Fuente: (Duque, 2007)

2.6.4 Transformada Wavelet

La transformada Wavelet contiene la mayor cantidad de energía localizada en un intervalo frecuencial y temporal. Cada dilatación que presenta la transformada wavelet está asociada a una frecuencia central. “La transformada mide para cada instante y cada frecuencia la correlación entre la señal original y la wavelet” (Vuletich, 2005).

El análisis clásico para Wavelet en 1980 se descompone en sumas de senos y cosenos y toma como base un conjunto de dilataciones de una función madre exponencial y toma como base dilaciones y traslaciones de una misma función (Vizzarri, 2016)

A continuación, se describe la fórmula de la transformada Wavelet

$$\psi(t) = \frac{1}{\sqrt{|a|}}\psi\left(\frac{t-b}{a}\right) \qquad (3)$$

Se preserva la energía de las funciones mediante un factor de normalización como lo indica la Figura 3:

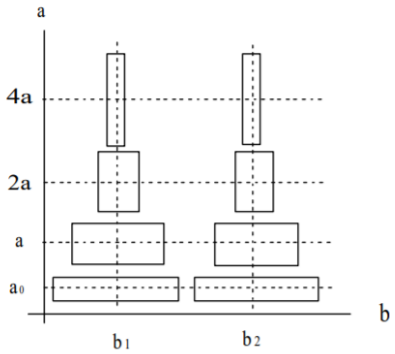

Figura 3. Plano tiempo-frecuencia para la Transformada Wavelet
Fuente: (Duque, 2007)

2.7 Extracción de características

Las características lingüísticas están muy relacionadas con el contexto en el que se van a emplear, la información lingüística para la clasificación de las emociones se obtiene principalmente a partir de la voz o de datos provenientes de una base de datos, según su género, edad u otro tipo de información acerca de la persona (Forbes, 2004). Las características contextuales en algunos casos se extraen semiautomáticamente y son dependientes de la aplicación que se va a utilizar (Liscombe, 2005). De acuerdo con el tipo de procesamiento de características se conoce dos enfoques modelo dinámico y modelo estático, este último es del que se va a hablar a continuación:

2.7.1 Modelo estático

En este modelo se usan métodos de clasificación como *Support Vector Machines (SVM)* o Redes Neuronales. La clasificación se hace en base a todo el audio que se obtuvo en diferentes tamaños. Las características son obtenidas a través de *Low Level Descriptors*, las cuales son el conjunto de características que se pueden extraer en cualquier punto de la señal. En el caso de una representación en el dominio del tiempo, la mayoría de los valores instantáneos útiles que se pueden calcular están relacionados con la amplitud o la energía de la señal. Los descriptores de

bajo nivel son por ejemplo entonación, energía o coeficientes espectrales, y de igual manera se va a aplicar medidas estadísticas como media, desviación estándar, densidad espectral (Planet, 2009).

2.7.2 Descriptores de bajo nivel

2.7.2.1 Prosódicas

Las características prosódicas describen los fenómenos de la voz como entonación, volumen, velocidad, pausas, duración y ritmo (Nuñez, 2010). En la Tabla 2 se puede observar las características prosódicas.

Tabla 2

Características Prosódicas de la voz

Características prosódicas	
speech rate	speech duration std
speech duration mean	syllable duration mean
syllable duration standard deviation	pause duration std
pause to speech ratio	pause duration mean
Contorno Melódico en la Elocución	
Pitch Average	Pitch Standard Deviation
Pitch Range	Pitch 25 % quantile
Mínimum Pitch Point	Pitch 75 % quantile
Máximum Pitch Point	Pitch Median
Pitch QuartRange	
Contorno de Energía	
Intensity Average	intensity_quartlow 25 % quantile
intensity_quartup 75 % quantile	intensity_max
intensity_range rango entre minimo y maximo	intensity_min
intensity_range_q rango entre quantiles	intensity_std

2.7.2.2 Calidad de voz

Estas características definen al hablar como neutral, jadeante, estrepitoso, sonoro, ruidoso y resonante. La mayoría de estas características han sido usadas para el reconocimiento de emociones por (Dubuisson, 2009) en la cual clasifica voces patológicas y normales. En la Tabla 3 se puede observar las características de calidad de la voz.

Tabla 3

Características de la Calidad de la voz

Características de calidad de la voz	
Jitter	Shimmer
Fraction of locally unvoiced frames	Number of voice breaks
Degree of voice break	Harmony autocorrelation mean
Noise-to-harmonics ratio mean	Harmonics-t-noise ratio mean
Harmonicity mean	Power rising mean
Power rising std	Power falling mean
Harmonicity standard deviation	Harmonicity min
Harmonicity max	Energy difference between frequency bands 60-400 Hz, 400-2000Hz, 2000-5000Hz, 5000-8000H
frequency bands ratio 60-400 Hz, 400-2000Hz, 2000-5000Hz, 5000- 8000Hz	

2.7.2 Espectrales

Describen las propiedades de una señal en el dominio de la frecuencia mediante formantes y harmónicos (Nuñez, 2010). En la Tabla 4 se puede observar las diferentes características espectrales.

Tabla 4

Características Espectrales de la voz

Características Espectrales	
Skewness	kurtosis
std	centroid
Long Term Average	
slope	mean
std	min
max	
Wavelets	
variance	min
std	max
mean	median
Cocleagrama	
variance	min
std	median
mean	max
LPC	
variance	min
std	median

2.8 *Machine Learning*

El aprendizaje, como la inteligencia, cubre una gama tan amplia de procesos que es difícil de definir con precisión. Ciertamente, muchas técnicas en aprendizaje automático se derivan de los esfuerzos de los psicólogos para hacer más precisas sus teorías del aprendizaje humano a través de modelos computacionales.

Con respecto a las máquinas, podríamos decir, muy ampliamente, que una máquina aprende cada vez que cambia su estructura, programa o datos tal manera que mejore su rendimiento futuro esperado.

Un ejemplo claro, es cuando el rendimiento de una máquina de reconocimiento de voz tiene una gran mejora después de escuchar varias muestras del habla de una persona, en ese caso se dice que la maquina ha aprendido (Nilsson, 1998).

Machine Learning desarrolla algoritmos que hacen que las máquinas puedan aprender por su cuenta, para lograr este propósito se distinguen dos modalidades:

2.8.1 Aprendizaje no supervisado

Este tipo de aprendizaje está más relacionado con la inteligencia artificial debido a que los datos que se proporciona son totalmente desconocidos para la máquina, y se tiene una idea que el computador pueda aprender por si solo como identificar emociones en este caso. Es decir, se entra de manera ciega y mediante operaciones lógicas se logra resolver problemas complejos utilizando solo los datos de entrada y los algoritmos lógicos, ya que en ningún momento se tiene datos de referencia. Algunos ejemplos de algoritmos complejos son *Clustering, k-means* y reglas de asociación (Chávez, 2012)

2.8.2 Aprendizaje supervisado

El aprendizaje supervisado es el más utilizado entre los dos, los algoritmos de clasificación que utiliza son SVM, regresión lineal y logístico. Se llama supervisado ya que el desarrollador actúa en el aprendizaje para enseñar a la máquina a las conclusiones que se debe llegar, por ejemplos actúa de manera similar en la que un niño aprende de su profesor de escuela. Es de vital importancia que los datos que se proporcione para entrenar el algoritmo estén previamente etiquetados (Chávez, 2012).

2.8.2.1 Regresión

Tiene como resultado un número específico ya que las etiquetas suelen ser un valor numérico, mediante las variables de las características se logra obtener dígitos como dato resultante. En la Figura 4 se observa una regresión mediante la técnica de aprendizaje supervisado.

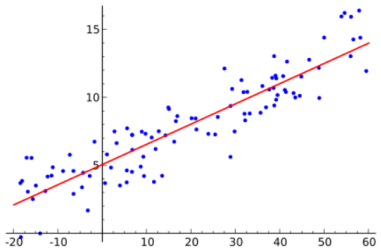

Figura 4. Regresión mediante la técnica de aprendizaje supervisado
Fuente: (Chávez, 2012)

2.8.2.2 Support Vector Machine

Los SVMs están entre los mejores algoritmos de aprendizaje supervisados, los SVM construyen un modelo que asigna nuevos ejemplos a una categoría u otra, lo que lo convierte en un clasificador lineal binario no probabilístico. Además de realizar una clasificación lineal , las SVM pueden realizar una clasificación no lineal de manera eficiente, mapeando implícitamente sus entradas en espacios de características de alta dimensión. Un SVM es un clasificador discriminativo definido formalmente por un hiperplano separador. En otras palabras, dados los datos de entrenamiento etiquetados (aprendizaje supervisado), el algoritmo genera un hiperplano óptimo que categoriza nuevos ejemplos. En dos espacios dimensionales, este hiperplano es una línea que divide un plano en dos partes, donde en cada clase se encuentra en cada lado.

SVM se basa en que cada nuevo dato puede ser utilizado dentro de la categoría que corresponderá basado en el análisis de datos. En el ejemplo de la Figura 5 los objetos pertenecen a solo una clase ya sea verde o rojo. La línea que tienen de separación define un límite en el lado derecho en el cual todos los objetos son de color verde y a la izquierda donde todos los objetos son rojos (Shwartz, 2014).

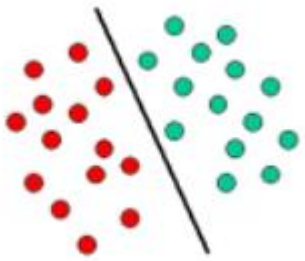

Figura 5. Ejemplo Support Vector Machine
Fuente: (Shwartz, 2014)

2.8.2.2 *K-Nearest Neighbors (KNN)*

El algoritmo KNN es un algoritmo de aprendizaje automático supervisado simple y fácil de implementar que se puede utilizar para resolver problemas de clasificación y regresión.

Un algoritmo de aprendizaje automático supervisado (a diferencia de un algoritmo de aprendizaje automático no supervisado) es uno que se basa en los datos de entrada etiquetados para aprender una función que produce un resultado apropiado cuando se le dan nuevos datos sin etiquetar.

El algoritmo KNN asume que existen cosas similares en la proximidad. En otras palabras, cosas similares están cerca unas de otras.

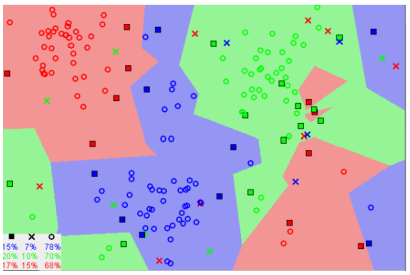

Figura 6. Imagen que muestra cómo los puntos de datos similares suelen existir cerca uno del otro
Fuente: (Bericat)

En la Figura 6 se logra observar que la mayoría de las veces, los puntos de datos similares están cerca uno del otro. El algoritmo de KNN se basa en esta suposición de ser lo suficientemente cierto para que el algoritmo sea útil. KNN captura la idea de similitud (a veces llamada distancia, proximidad o cercanía) con algunas matemáticas que podríamos haber aprendido en nuestra infancia, al calcular la distancia entre puntos en una gráfica.

CAPÍTULO III

METODOLOGÍA DEL PROYECTO DE INVESTIGACIÓN

3 Metodología del proyecto de investigación

3.1 Descripción general del proyecto de investigación

El presente trabajo de investigación tiene la intención de reconocer emociones a partir de la voz de hombres y mujeres, el Software Matlab®, se utilizó para el procesamiento de la señal, para la extracción de características mediante diferentes métodos y para entrenar el sistema mediante la teoría de *Machine Learning* a través de la técnica de clasificación supervisada.

Las bases de datos que se van a utilizar son la primera libre actuada en inglés, la cual contiene 2880 archivos de audio, que consta con 60 intentos por actor, con intensidad fuerte y normal actuando las emociones básicas, felicidad, enojo, depresión, temor, entre otras (RAVDESS, 2018). La segunda base de datos que se trabajó es de la Universidad de Toronto que consta de cuatro actrices y cuatro actores (entre 26 y 40 años) las cuales pronunciaron un conjunto de 200 palabras objetivo en la frase del portador "Diga la palabra _" y se realizaron grabaciones del conjunto que retrataba diferentes emociones. De dichas bases de datos se seleccionaron específicamente las emociones básicas de diferentes individuos. Se trabajó con dos bases de datos a la par para aseverar que el sistema no memorice y las ventajas que presentaron estas dos bases de datos, es que todos los audios presentes son de diferentes frases para cada emoción. De la base de datos RAVDESS se seleccionó los audios que simulaban solamente las cuatro emociones básicas, de lo cual se obtuvo 216 audios repartidos equitativamente para cada emoción; de la base de datos de Toronto, se realizó de igual manera se trabajó solo con los audios que simulaban las cuatro emociones, de lo cual se obtuvo 96 audios. Finalmente, en el presente proyecto de

investigación se trabajó con 312 audios, con un total de 62 actores repartidos equitativamente para cada género.

La extracción de características se realizó de dos maneras:

- Señal de voz sin ningún preprocesamiento.
- Transformada Wavelet

En cada una de ellas se ejecutó un análisis descartando las características que no presentaban gran relevancia para el estudio, con el fin de reducir el costo computacional. Finalmente se entrenó el sistema mediante la teoría de *Machine Learning* a través de la técnica de clasificación supervisada para detectar los diferentes estados emocionales de las personas.

En la Figura 7 se observa el diagrama de bloques del proyecto de investigación.

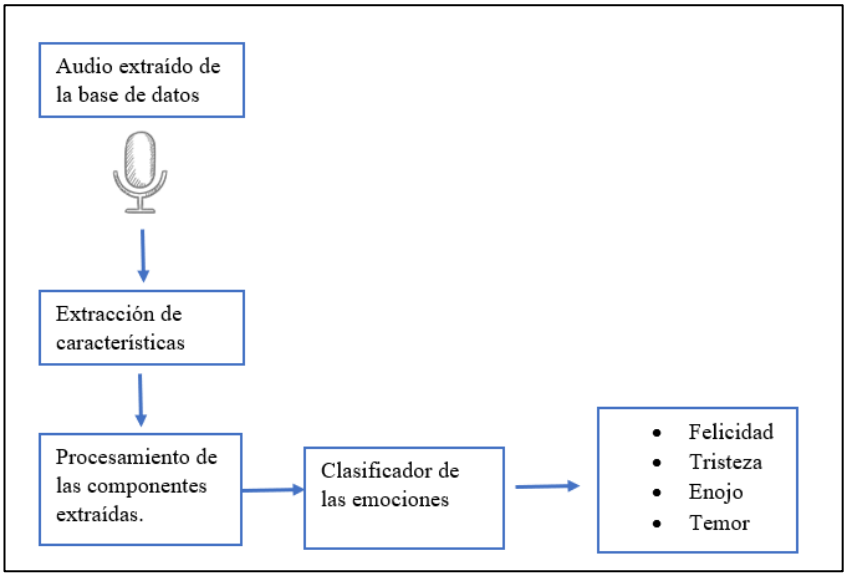

Figura 7. Diagrama de bloques

3.2 Extracción de características

A continuación, se presentan los métodos utilizados para la extracción de las características de la señal de voz más importantes. El audio se extraerá de las bases de datos mencionadas anteriormente. Estas muestras se utilizaron para el procesamiento, entrenamiento y clasificación del sistema.

Se utilizó la plataforma Matlab® ya que es una herramienta que permite obtener el procesamiento de las señales de voz por medio de un lenguaje de programación más amigable al usuario.

3.2.1 Frecuencia fundamental (*Pitch*)

La frecuencia fundamental se relaciona con el número de veces que vibran por segundo las cuerdas vocales de las personas, esta percepción en algunos casos puede corresponder a la intensidad o a las propiedades espectrales del sonido de la voz. En los hombres que no presentan ningún tipo de patología esta frecuencia varía entre los 120Hz y 125Hz, y en las mujeres se encuentra en 225Hz (Diaz, 2015).

La frecuencia fundamental se obtuvo mediante el análisis cepstral de una señal de habla la que permite trabajar con la señal de la glotis (excitación) y la del tracto vocal (resonancia) por separado. Dentro de las propiedades matemáticas involucradas en el proceso, se destaca principalmente las transformadas de *Fourier* y funciones logarítmicas que resultaron en una función llamada *Cepstral* o *Cepstro*. El método del *Cepstro* es entonces una operación matemática que consiste en extraer la Transformada de Fourier del espectro de la señal en forma de logaritmos. Una vez obtenido la señal espectral se extrae la frecuencia fundamental, anchos de

banda y amplitud de los formantes. ya que la frecuencia fundamental tiene ciertas diferencias según el sexo (Teixeira, Barbosa, & Moreira, 2011).

3.2.2 Jitter

Es una medida que sirve para observar las irregularidades que presentan en la frecuencia, ya que unas ondas pueden ser más anchas que otras, por lo que algunos pliegues vocales no vibran a la misma velocidad. El Jitter es considerado un parámetro microscópico que mide una alteración mínima en una ventana de 20 milisegundos, la cual es la que el programa considere más estable. (Casado, 2002)

Los valores pequeños del jitter se consideran normales, mientras que los valores que son relativamente grandes se refieren a una patología (Garcia, 2015).

***Jitter* (local):** Representa la diferencia media absoluta entre dos periodos consecutivos, dividido por el periodo medio, ecuación 4.

$$jitt = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i}, \quad (4)$$

donde T_i representa la longitud de cada periodo de la frecuencia fundamental y N el número de periodos.

***Jitter* (absoluto):** Mide la variación de la frecuencia en unidades de tiempo ecuación 5.

$$jitta = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}| \quad (5)$$

Uno de los objetivos principales del cálculo del parámetro *Jitter* es la detección de patologías que presenta una persona con algún trastorno emocional, en el estado del arte existe mucha

variación al respecto al umbral que se requiere para detectar una patología, la cual oscila para *Jitter* local >1.04% y > 3% y para *Jitter* absoluto >0.83% (Sintas, 2013).

3.2.3 Shimmer

Shimmer es la variación de la amplitud de la señal. Son medidas de la perturbación de la amplitud en general. La medición del *Shimmer* se usó para cuantificar pequeños lapsos de inestabilidad acústica, puede ser medido en decibelios, a través de la ecuación 6 (Garcia, 2015).

$$Shimmer(dB) = \frac{20 \times \sum_{i=0}^{i=N-1} |log_{10} \frac{A_i}{A_{i+1}}|}{N-1} \quad (6)$$

El *Shimmer* puede variar por dos causas:

- Por cambios en el tono muscular, debido a desordenes neurológicos.
- Por alteraciones aerodinámicas causadas por problemas broncopulmonares.

***Shimmer* Local:** Representa una diferencia absoluta entre las amplitudes de dos períodos consecutivos, dividido para la amplitud media, como lo indica la ecuación 7.

$$Shimm = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (7)$$

El cálculo del parámetro *Shimmer* es la detección de patologías que presenta una persona con algún trastorno emocional, con respecto al umbral que se requiere para detectar una patología, el *Shimmer local* debe ser >3.810%, *Shimmer local* (dB) >0.350 (Sintas, 2013).

3.2.4 Entropía de la energía

La entropía de energía de una señal de voz es una medida de la cantidad de información que posee la señal, esta característica lleva cambios más bruscos en el nivel de energía de una señal de audio, los cuales son datos relevantes que pueden usarse para mejorar la capacidad del clasificador.

La entropía de banda completa en el dominio del tiempo captura la distorsión que existe en personas que presentan patologías, para mejorar la resolución la entropía se calcula a partir de regiones basadas en energía (Fatiha, 2016).

La entropía se denota con la siguiente fórmula:

$$E(x) = -\sum_{i}^{\infty} p(x_i).\log_2(\mathrm{p}(\mathrm{x_i})) \quad (8)$$

Según la investigación realizada en el presente proyecto, el tamaño de la ventana varía dependiendo el número de muestras de la señal entre 20ms a 50ms, los audios presentados en las bases de datos mencionadas anteriormente manifestaron un número de muestras que oscilaban entre 200000 a 300000, razón por la cual para disminuir el costo computacional se trabajó con una ventana de 50ms y con una frecuencia de muestreo de 44100Hz dando como resultado un tamaño de 2205 muestras por ventana, lo cual sirvió para calcular por partes la entropía de la señal y como resultado final se obtuvo un vector del cual se extrajo los valores estadísticos como la desviación estándar, media, mediana y varianza, respectivamente.

3.2.5 Energía

La energía es una medición de la intensidad sonora, para eso se usa el análisis temporal en el dominio de la frecuencia. La energía es un parámetro útil para el reconocimiento automático de emociones ya que proporciona características de diferenciación entre segmentos sordos y sonoros de la señal de voz, ya que la amplitud en los segmentos sordos es mucho más baja que en los segmentos sonoros. Cuando se pretende distinguir entre sonidos sordos y fricativos, puede existir cierta imprecisión si se utiliza únicamente la métrica de la energía (Rabiner, 1978).

La energía se define como:

$$E_{seg} = N_A . E\{[s(n) - \mu_s(n)]^2\}, \quad (9)$$

donde, $s(n)$ es la señal de voz, $\mu_s(n)$ es la media de $s(n)$ y N_A es el número de muestras del segmento en análisis.

La Figura 8 representa el comportamiento de la energía, para las voces normales y para las que presentan patología.

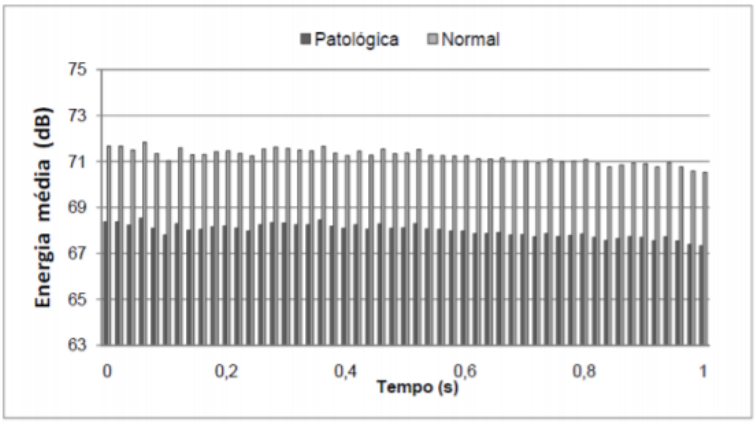

Figura 8. Energía para voces patológicas y normales
Fuente: (Cavalcanti, 2010)

3.2.6 Tasa de cruce por cero

La tasa de cruce por cero de una señal de audio indica el número de veces en que la señal continua cambia del valor positivo a negativo o viceversa. Las señales con altas frecuencias presentan una alta tasa de cruce por cero, la señal debe estar totalmente sin ruido, ya que este genera un gran número de cruces por cero (Eldho, 2017).

La tasa de cruce por cero está definida por la siguiente ecuación:

$$Z(i) = \frac{1}{W_L}\sum_{n=1}^{W_L}|sgn[x_i(n)] - sgn[x_i(n-1)]|, \quad (10)$$

donde sgn es la señal de audio y W_L es el tamaño de la ventana, que en este caso se usó de 50 milisegundos la cual sirvió para calcular por partes la tasa de cruce por cero de la señal y finalmente se obtuvo un vector del cual se extrajo los valores estadísticos como la desviación estándar, media, mediana y varianza, respectivamente.

3.2.7 *Rolloff* espectral

El *Rolloff* espectral se define como la frecuencia por debajo de la cual un determinado porcentaje de la distribución de la magnitud del espectro es concentrado (Teixeira, Barbosa, & Moreira, 2011). Para lo cual se utiliza la siguiente ecuación:

$$\sum_{k=1}^{m} X_i(k) = C\sum_{k=1}^{W_L} X_i(k), \quad (11)$$

donde C es el porcentaje adoptado, la frecuencia espectral rolloff es usualmente normalizada al dividirla para el tamaño de la ventana W_L, por lo que toma valores entre 0 y 1.

3.2.8 Centroide espectral

Centroide espectral se asocia comúnmente con la medida de claridad de una señal de audio. Esta medida se obtiene al evaluar el centro de gravedad utilizando información de frecuencia de la transformada de Fourier. El centroide espectral se define como la frecuencia promedio ponderada por las amplitudes, dividida por la suma de las amplitudes (Nam, 2001), como le muestra la siguiente ecuación:

$$C_i = \frac{\sum_{k=1}^{W_L} kX_i(k)}{\sum_{k=1}^{W_L} X_i(k)}, \quad (12)$$

donde, W_L es el tamaño de la ventana que en este caso se usó de 50 milisegundos y $X_i(k)$ es la transformada de Fourier de la señal original, posteriormente se obtuvo un vector del cual se adquirió los valores estadísticos como la desviación estándar, media, mediana y varianza.

3.2.9 Flujo espectral

El flujo espectral calcula el cambio en cuanto al espectro que existe entre dos ventanas sucesivas y se calcula como la diferencia entra las magnitudes normalizadas de los espectros de los cuadros sucesivos (Teixeira, Barbosa, & Moreira, 2011). La ecuación que define el flujo espectral se define de la siguiente forma:

$$Fl_{i,i-1} = \sum_{k=1}^{W_L} \left(EN_i(k) - EN_{i-1}(k)\right)^2, \quad (13)$$

donde, W_L es el tamaño de la ventana que en este caso se usó de 50 milisegundos finalmente se obtuvo un vector del cual se adquirió los valores estadísticos como la desviación estándar, media, mediana y varianza, respectivamente.

3.2.10 Relación señal a ruido

Se define como relación señal a ruido, S/N es el cociente de la potencia de la señal entre la potencia de ruido en un punto dado de un sistema, es decir:

$$\frac{S}{N} = \frac{Potencia\ de\ señal}{Potencia\ de\ ruido} \quad (14)$$

La relación señal a ruido representa una medida de la calidad de la señal de audio y depende, del nivel de señal recibida como del ruido total que existe en la señal.

3.2.11 Características estadísticas

La señal se normalizó de 0 a 1, para prevenir el *Clipping*, ya que esto sucede cuando una señal de audio presenta altos componentes, causando distorsión en su frecuencia. El tamaño de *winLength* que se usó es de 50 milisegundos y el *numofBlocks* varió según el tamaño de cada audio (Mora, 2015).

3.2.11.1 Media

La media es el valor central que existe en un conjunto de realizaciones. Es el punto más cercano a todos los posibles valores de la variable aleatoria (Morales, 2007). En Matlab® se empleó el comando *mean*. La media se define de la siguiente forma:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} A_i \quad (15)$$

3.2.11.2 Varianza

La varianza es el promedio que mide la distancia de los valores de la variable a la media de ésta. Si el valor es muy grande indica que existe una alta dispersión de los valores con respecto a su media, y si el valor de la varianza es pequeño indica una alta concentración de los valores de la

variable en torno a su media. El problema que presenta la varianza es que se manifiesta en unidades cuadráticas, para obviar este problema se presenta la desviación estándar, que es la raíz cuadrada de la varianza (Gómez, 2017). En Matlab® se empleó el comando *var*. La varianza se expresa como el momento de segundo orden:

$$W_k = E[(g(x) - U_1)^k] = E[(g(x) - U_1)^2] \quad (16)$$

3.2.11.3 Sesgo *(Skewness)*

Es una medida de que observa que tan asimétrica es la distribución de la variable aleatoria, es decir estudia las variaciones extremas en algún sentido, sean frecuencias altas o bajas (Sintas, 2013) En Matlab® se empleó el comando *skewness*. El sesgo se expresa de la siguiente forma:

$$W_3 = E[(g(x) - U_1)^3] \quad (17)$$

Si se presenta una asimetría negativa, la variable toma valores muy bajos. Si por el contrario es positiva la variable toma valores muy altos con mayor frecuencia que los bajos.

3.2.11.4 Curtosis

La curtosis es una medida que observa la forma de distribución de la variable aleatoria, que determina el grado de concentración que presentan los valores de una variable alrededor de la zona media de la distribución de sus frecuencias (Diaz, 2015). En Matlab® se empleó el comando *kurtosis*. La curtosis se define con la siguiente ecuación:

$$W_4 = E[(g(x) - U_1)^4] \quad (18)$$

3.2.11.5 Mediana

La mediana es el valor útil cuando la distribución de la variable es poco simétrica, ya que representa el valor de la posición central en un conjunto de datos (Bolaños, 2018). En Matlab® se empleó el comando *median.*

3.2.11.6 Desviación estándar

La desviación estándar es la medida que muestra que tan dispersos están los datos con respecto a la media. Si la desviación estándar es alta, la dispersión de datos es mayor. En Matlab® se empleó el comando *std*. La desviación estándar es la raíz cuadrada de la varianza que se muestra de la siguiente forma:

$$s = \sqrt{E[(g(x) - U_1)^2]} \quad (19)$$

3.2.12. Transformada Wavelet

Esta transformada es capaz de concentrarse en fenómenos transitorios y de alta frecuencia mejor que la Transformada de Fourier con Ventana. Las wavelets permiten representaciones de funciones en las cuales se retiene la información espacial como la escala. La transformada Wavelet se usó para aproximar con gran exactitud las diferentes características de la señal de voz que posteriormente se usaron para la clasificación de las emociones, con un pequeño número de coeficientes wavelet (Azor, 2017). Para obtener los coeficientes de wavelet, se utilizó el comando de Matlab® "*[c,l] = wavedec(x,n,wname*)", el cual devuelve la descomposición wavelet de la señal *x*, en *n* nivel y *wname* permite seleccionar el tipo de wavelet que se quiere trabajar (Matlab, 2019). El tipo de transformada wavelet que se utilizó es *Daubechies5* de nivel 3, ya que según

(Oliveira, 2019) el *db5* permite obtener fluctuaciones pequeñas debido a la anulación de momentos. En la Figura 9, se puede observar los niveles de descomposición sobre la frase /*Kids are talking by the door*/ en el estado emocional de felicidad.

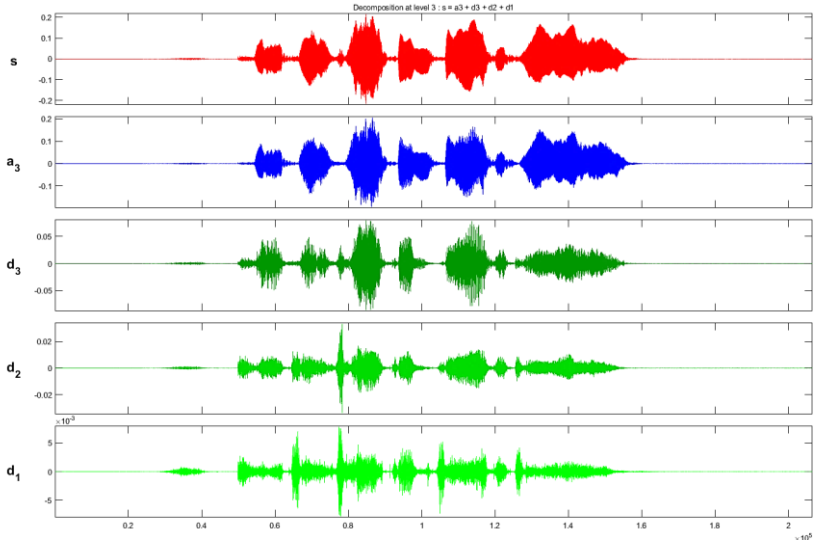

Figura 9. Descomposición en 3 niveles para la frase /Kids are talking by the door/ Felicidad. Matlab®

3.3 Base de datos

Se utilizó dos bases de datos diferentes, la primera base de datos que se empleó es libre actuada en inglés, clasificada y la cual contiene 2880 archivos de audio, que consta con 60 intentos por actor, con intensidad fuerte y normal (RAVDESS, 2018) entre diferentes emociones. La segunda base de datos que se trabajó es de la Universidad de Toronto que consta de cuatro actrices y cuatro actores (entre 26 y 40 años) los cuales pronunciaron un conjunto de 200 palabras objetivo en la frase del portador "Diga la palabra _____" y se realizaron grabaciones del conjunto que retrataba cada una de las cuatro emociones (felicidad, miedo, tristeza, enojo, entre otras). Los actores hablan inglés como primer idioma, tienen estudios universitarios y tienen

formación musical. Las pruebas audio métricas indicaron que todos los actores de las dos bases de datos tienen umbrales dentro del rango normal (Dupuis, 2010).

Fueron analizados un total de 312 audios que contienen las emociones tales como felicidad, enojo, miedo y tristeza. Comúnmente se recomienda para una clasificación binaria que la distribución sea 70% y 30%, para el conjunto de entrenamiento y prueba respectivamente, pero para este caso que es una clasificación multiclase según el estado del arte se investigó que la distribución puede ser 60% y 40%, ya que el conjunto de pruebas debe ser un poco grande para que haya más intervalos de confianza y asegurar que el modelo no haya memorizado, sino que haya aprendido las diferentes características. Dentro de la cual se utilizó un total de 192 audios para el entrenamiento, repartidos equitativamente entre las cuatro emociones, para hombre y mujer respectivamente y un total de 120 audios divididos equitativamente, para realizar las pruebas obtenidas con el clasificador para cada género.

3.4 Procesamiento de características extraídas

En el Aprendizaje supervisado, los algoritmos aprenden de datos etiquetados. Después de comprender los datos, el algoritmo determina qué etiqueta debe asignarse a los nuevos datos en función del patrón y asociando los patrones a los nuevos datos sin etiquetar. La clasificación es una técnica para determinar la clase a la que pertenece el dependiente en función de una o más variables independientes (Teixeira, Barbosa, & Moreira, 2011).

3.5 Aprendizaje automático supervisado

En el presente trabajo el aprendizaje automático se utilizó para la clasificación de emociones a partir de señales de voz, el cual consta con dos fases, la de entrenamiento y la de evaluación.

En primera instancia se realiza un entrenamiento en base al aprendizaje automático supervisado, ya que se proporcionó al sistema un conjunto de datos etiquetados. Para esta etapa se utilizó la aplicación "*Classification Learner*" de la herramienta de Matlab® (Figura 9), con esta aplicación se seleccionó el archivo que contenía los predictores. En el cual se realizó un entrenamiento automatizado para buscar el mejor tipo de modelo de clasificación, incluidos los árboles de decisión, el análisis discriminante, SVM, KNN y la clasificación por conjuntos. En sus diferentes variaciones respectivamente. Este proceso se realizó con una base de datos de 192 audios repartidos equitativamente entre las 4 emociones.

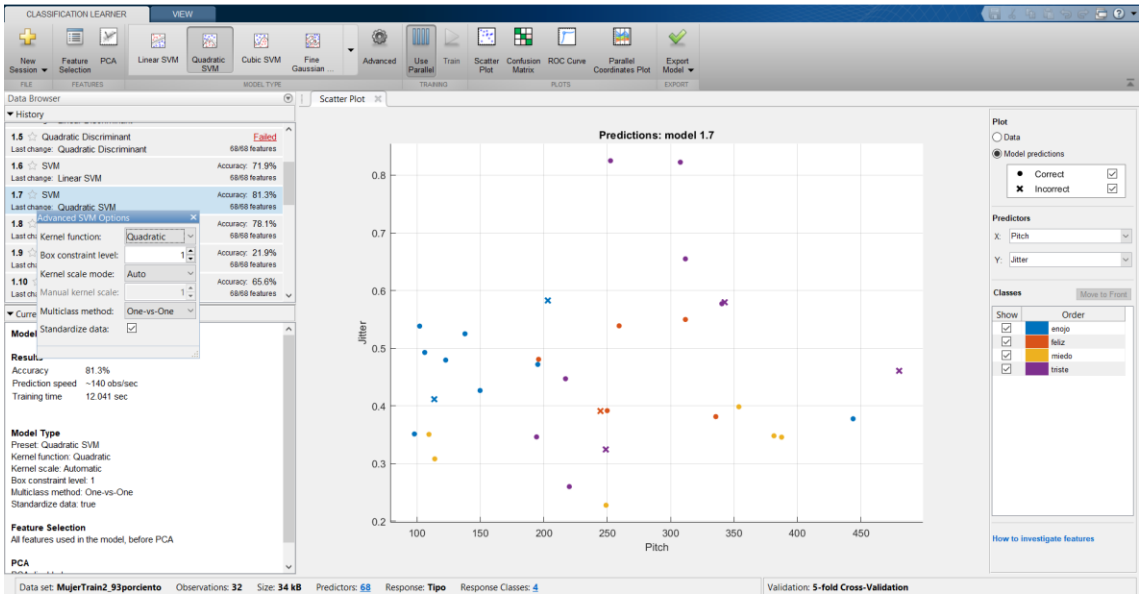

Figura 10. App Classification Learner Matlab®

Mediante esta aplicación (Figura 10) se extrajeron dos modelos de entrenamiento tanto para hombre, como para mujer.

En el caso de hombre, el mejor modelo de entrenamiento que se obtuvo fue el *Weighted k-Nearest-Neighbors (wkNN),* en el cual se varió el número de vecinos, la distancia métrica optima,

después de descartar la de coseno, *chebyshev*, *cubic* entre otras, resultó la distancia Euclideana ya que es la distancia en línea recta o la trayectoria más corta posible entre dos puntos. Ya que con esta resultó el porcentaje más alto de exactitud durante el entrenamiento (Hechenbichler, 2015). El modelo de entrenamiento *wkNN* de igual manera presenta la posibilidad de variar *"Distance Weight"* entre "*equal, inverse, squared inverse*" en este caso el peso de la distancia más eficaz resultó ser el "*squared inverse*". Los pesos son proporcionales a la distancia inversa elevada al valor de óptimo de la potencia *p*. Como resultado, a medida que aumenta la distancia, los pesos disminuyen rápidamente. Según (Larrañaga, 2015) se deben usar funciones de potencias mayores que 1, en este caso Matlab® usa *p=2*, la cual se la conoce como la interpolación ponderada al cuadrado de la distancia inversa.

En el caso de la mujer, el mejor modelo de entrenamiento que se presentó fue el *SVM Quadratic*, en el cual se puede cambiar el orden polinomial de la función núcleo (kernel), ya sea lineal, cuadrática, cúbica o gaussiana. Además, en Matlab® se permite variar la Función de Costo o parámetro de regularización el cual le dice a la optimización de SVM cuánto quiere evitar clasificar erróneamente cada ejemplo de entrenamiento. Un valor más alto conduce a resultados correctos (Carmona E. , 2013) (Figura 11). Según Matlab® el valor de la función de costo más optima es 1 para *Quadratic SVM*.

Figura 11. Izquierda: bajo valor de regularización, derecha: alto valor de regularización
Fuente: (Carmona E. , 2013)

El parámetro *gamma* define hasta dónde llega la influencia de un solo ejemplo de entrenamiento, con valores bajos que significan 'lejos' y valores altos que significan 'cerca'. En otras palabras, con gamma baja, los puntos alejados de la línea de separación plausible se consideran en el cálculo de la línea de separación. Donde alto gamma significa que los puntos cercanos a la línea plausible se consideran en el cálculo (Figura 12). El valor de gamma para *SVM Quadratic* que resultó en este caso es 1.

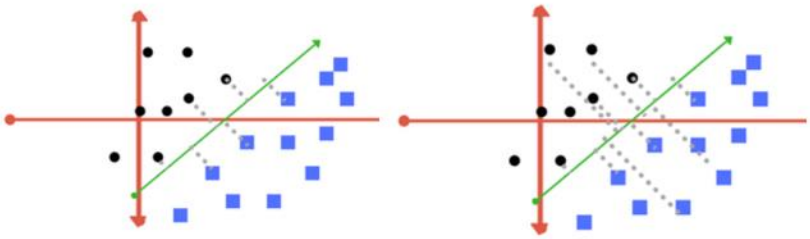

Figura 12. Izquierda: valor alto de gamma, derecha: valor bajo de gamma
Fuente: (Carmona E. , 2013)

3.6 Selección de características

La selección adecuada de características puede mejorar la precisión y la eficiencia de los métodos de clasificación. La clasificación tiene como objetivo reducir la dimensionalidad y el ruido en los conjuntos de datos. La selección de características realiza filtrado de información, ya que elimina información redundante o no deseada de un flujo de información y así evitar el *Overfitting*, que es el ajuste excesivo cuando un modelo de entrenamiento aprende el ruido de los datos y de esta manera entorpecer la clasificación. A veces, en muchos dominios de aprendizaje, un operador humano define las características potencialmente útiles. Sin embargo, no todas estas características pueden ser relevantes y algunas de ellos puede ser redundantes (Roffo, 2018).

Para lo cual se utilizó la biblioteca de selección de características (FSLib 2018) es una biblioteca Matlab®, desarrollada por Giorgio Roffo, de amplia aplicación para la selección de características (selección de atributos o variables), la cual se empleó para eliminar información irrelevante, redundante y ruidosa de los datos.

Las técnicas de selección de características se dividen en tres clases: *wrapper* se utiliza para los modelos *SVM* los cuales usan clasificadores para puntuar un conjunto dado de características, *embedded* se utiliza para modelos de árbol de decisión y *filter* se utiliza para modelos KNN, los cuales identifican las propiedades intrínsecas de los datos, ignorando el clasificador (Roffo, 2018).

En el presente estudio se usó el método *wrapper* para la selección de características de las mujeres, establecida para SVM (Figura 13), los modelos *wrapper* implican la optimización de un predictor como parte del proceso de selección. Ellos tienden obtener mejores resultados, pero los métodos *filter* suelen ser computacionalmente menos costosos que los *wrapper*. De igual manera, se usó el método *filter* para la selección de características de los hombres, establecida para KNN (Figura 14), estos modelos confían sobre las características generales de los datos de entrenamiento para seleccionar características con independencia sobre cualquier predictor.

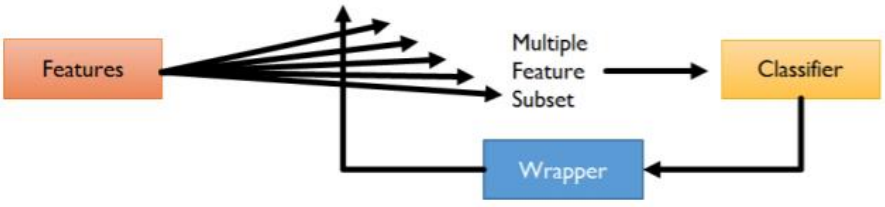

Figura 13. Modelo wrapper
Fuente: (Roffo, 2018)

Figura 14. Modelo Filter
Fuente: (Roffo, 2018)

De acuerdo con la librería *Feature Selection,* para el modelo KNN el método que resultó óptimo es *Features Selection via Eigenvector Centrality* (ECFS) que clasifica las características mediante la identificación de las más importantes en un conjunto arbitrario de señales (Roffo, 2018). La selección de características a través de ECFS es un método de filtro que mapea el problema de FS en un gráfico de afinidad, donde las características son los nodos, la solución se da evaluando la importancia de los nodos a través de algunos indicadores de centralidad. La esencia de ECFS es estimar la importancia de una característica en función de la importancia de sus vecinos. Individualiza las características de los candidatos, que resultan ser efectivas a partir de una clasificación (Roffo, 2018).

Para el modelo SVM el método que resultó mejor es: *Feature selection via concave minimization and support vector machines* (FSV), este método según (Bradley, 2015) presenta mejores resultados de clasificación, donde el proceso de selección de características se inyecta en el entrenamiento de un modelo SVM por una técnica de programación lineal (Roffo, 2018).

Una vez realizado la selección de características, se realizó pruebas descartando características consideradas menos importantes, que no aportaban al aprendizaje automático del entrenador. Se realizó pruebas para hombres y para mujeres por separado.

CAPÍTULO IV

ANÁLISIS DE RESULTADOS

4 Análisis de resultados

Los resultados obtenidos a través de este estudio fueron analizados para un total de 312 audios divididos en dos partes, el primero corresponde al conjunto de entrenamiento que incluye a 96 audios para hombre y 96 para mujer, repartidos equitativamente entre las cuatro emociones y el segundo es el conjunto de prueba que corresponde a 60 audios para hombre y 60 para mujer. Dicho de otra forma, el conjunto de entrenamiento equivale al 61.54%, mientras que el de prueba a 38.46%. Esto permite los resultados sean factibles y se evite el ajuste excesivo.

Con el fin de evaluar los resultados alcanzados, estos fueron analizados bajo cuatro parámetros:

Exactitud (A)

$$A(\%) = \frac{N_E}{N_T} \times 100 \quad (20)$$

Precisión (P)

$$P(\%) = \frac{N_{VP}}{N_{VP} + N_{FP}} \times 100 \quad (21)$$

Sensibilidad (R)

$$R(\%) = \frac{N_{VP}}{N_{VP} + N_{FN}} \times 100 \quad (22)$$

Especificidad (S)

$$S(\%) = \frac{N_{VN}}{N_{VN} + N_{FP}} \times 100 \quad (23)$$

donde:

N_E, número de eventos clasificados correctamente

N_T, número de eventos empleados en la clasificación

N_{VP}, número de verdaderos positivos

N_{FP}, número falsos positivos

N_{VN}, número de verdaderos negativos

N_{FN}, número de falsos negativos

4.1 Análisis del modelo de clasificación de hombres

4.1.1 Primer experimento

El primer experimento se realizó tras la extracción de las 68 características (Anexo A), estas fueron ingresadas a dos modelos diferentes de clasificadores a la *App Clasification Learner* de la Plataforma Matlab®, *wKNN* y *SVM*; con sus respectivos tipos independientemente. Se modificó inicialmente en el caso de *wKNN* el número de vecinos para *k=1, k=10, k=15, k=20, k=35, k=40, k=50, k=60, k=70, k=80 y k=96* y para SVM se modificó el parámetro gamma para γ *=1, =2.1, =8.2, y* la función de costo *C=1* permaneció constante, ya que se según Matlab® estos valores son óptimos para cada tipo de clasificador dentro de SVM, posteriormente se extrajo cada modelo de entrenamiento y se lo evalúo con el conjunto de prueba de 60 audios. Como se

puede observar en la Tabla 5, se indican los porcentajes de exactitud para cada tipo de clasificador obtenidos de Matlab®, los porcentajes de los tres clasificadores más altos son *Weighted KNN* con k=10 con el 68.8%de exactitud, *Weighted KNN* con k=20 con el 65.8% y *Weighted KNN* con k=1 con el 63.5%.

Tabla 5

Porcentaje de exactitud de diferentes tipos de clasificadores

Entrenamiento			**Evaluación**
Clasificador	**C**	γ	**Exactitud (%)**
Quadratic SVM	1	1	50
Fine Gaussian SVM	1	2.1	55.6
Medium Gaussian SVM	1	8.2	56.3
Weighted KNN	**k**		
	1		63.5
	10		68.8
	15		60,9
	20		65.8
	35		42,5
	40		40
	50		41.5
	60		40
	70		42
	80		40
	96		40

Según los resultados obtenidos en la primera etapa, el modelo de clasificador óptimo para la detección de emociones en el caso de los hombres es el *Weighted k-Nearest-Neighbors (wKNN),* con *k=10,* la distancia métrica más optima, después de descartar la de coseno, *chebyshev*, *cubic* entre otras, resultó la distancia Euclidiana ya que es la distancia en línea recta o la trayectoria más corta posible entre dos puntos, de igual manera se varió el parámetro *"Distance Weight"* entre "*equal, inverse, squared inverse*" en este caso el peso de la distancia más eficaz resultó ser el "*squared inverse*".

4.1.2 Segundo experimento - Selección de características mediante el método ECFS

Tras la selección del modelo de clasificador óptimo y con el objetivo de mejorar el rendimiento del sistema se realizó una selección de características con el método *Feature Selection via Eigenvector Centrality* (ECFS), que es tipo *Filter* y es el apropiado para el modelo wKNN, por ende, el método ECFS evitó el ajuste excesivo de datos y redujo el costo computacional al menorar el número de características irrelevantes para el estudio. En la Tabla 6 se puede observar que en primer lugar se extrajo el modelo wKNN dependiendo del número de características y se lo evalúo con el conjunto de prueba de 60 audios. En la Figura 15 se observa que se realizó una evaluación y reducción paulatina del número de características irrelevantes en el cual se determinó que emplear 53 características (Anexo B) en lugar de las 68, ofrece un rendimiento del 88.5% y menora el costo computacional del sistema, caso contrario de lo que sucede al utilizar 52 características, que ofrece un rendimiento del 85%. Para realizar estas pruebas se trabajó con *k=10,* dato que posteriormente se evaluará.

Tabla 6

Selección de características mediante el método ECFS

Entrenamiento		**Evaluación**
	N. Características	**Exactitud (%)**
Weighted KNN	68	68
	63	68
	58	78
	54	86
	53	88.5
	52	85
	48	80
	43	80

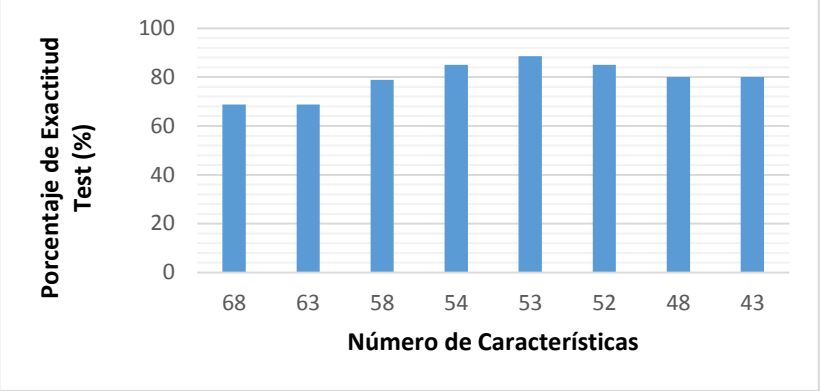

Figura 15. Selección de características utilizando el método ECFS

La Tabla 7 indica las 10 características más importantes obtenidas a través del método ECFS. En la cual *Shimmer* tiene mayor importancia para la clasificación de emociones en el caso de hombre.

Tabla 7
Orden de características obtenidas a través del método ECFS

Orden	Característica
1	Shimmer
2	Media de entropía de la energía
3	Media del centroide
4	Media spectral Roll-Off
5	Desviación estándar del centroide
6	Media con coeficientes Wavelet
7	Desviación estándar Centroide Wavelet
8	Media Centroide Wavelet
9	Desviación estándar energía
10	Jitter

4.1.3 Tercer experimento – Características sin procesamiento

Tras la selección del modelo de clasificador y el número de características óptimos, se realizó pruebas con el conjunto de datos de entrenamiento con un rango de vecinos desde 1 hasta 96 para verificar el número de vecinos correctos y se determinó mediante el conjunto de evaluación con qué número se maximizan todos los parámetros de evaluación del clasificador. De esta forma, los datos obtenidos se indican en la Figura 16.

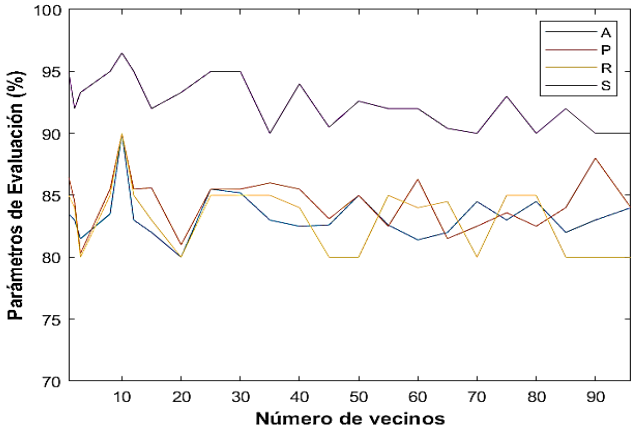

Figura 16. Parámetros de evaluación en función del número de vecinos después de evaluar 53 características sin procesamiento

Tabla 8

Porcentaje de los parámetros de evaluación en función al número de vecinos

Entrenamiento		**Evaluación**				
		N. Vecinos	**Exactitud**	**Precisión**	**Sensibilidad**	**Especificidad**
Weighted KNN	**53 características**	1	83.5%	86.5%	85%	95%
		2	83%	84.5%	84%	92%
		3	81.5%	80.3%	80%	93.3%
		8	83.5%	85.5%	85%	95%
		10	88.5%	90%	90%	90%
		12	83%	85.5%	85%	95%
		15	82%	85.6%	83%	92%
		20	80%	81%	80%	93.3%
		25	85.5%	85.5%	85%	95%
		30	82.2%	85.5%	85%	95%
		35	82%	85%	80%	92.5%
		40	83%	86%	85%	90%
		45	82.5%	85.2%	84%	94%

	50	82.6%	83.1%	80%	90.5%
	55	85%	85%	80%	92.6%
	60	82.6%	82.5%	85%	92%
	65	81.4%	86.3%	84%	92%
	70	82%	81.5%	84.5%	90.4%
	75	85%	82.5%	80%	90%
	80	81.5%	83.6%	85%	93%
	85	82%	82.5%	85%	90%
	90	83%	84%	80%	92%
	96	84%	88%	90%	90%

En la Figura 16, se observa que el número de vecinos que maximiza todos los parámetros de evaluación es *k*=10, este valor se puede corroborar en la Tabla 8, que indica de forma más específica los porcentajes obtenidos. En el cual se obtuvo una exactitud del 88.5%, una precisión de 90%, una sensibilidad del 90% y una especificidad del 90%.

4.1.4 Cuarto experimento - Características normalizadas

En este caso se normalizaron las 53 características seleccionadas anteriormente, con lo cual se varió el rango de vecinos de 1 hasta 96, con el fin de determinar con qué número de vecinos se maximizan todos los parámetros de evaluación, por ende, los datos obtenidos se muestran en la Figura 17.

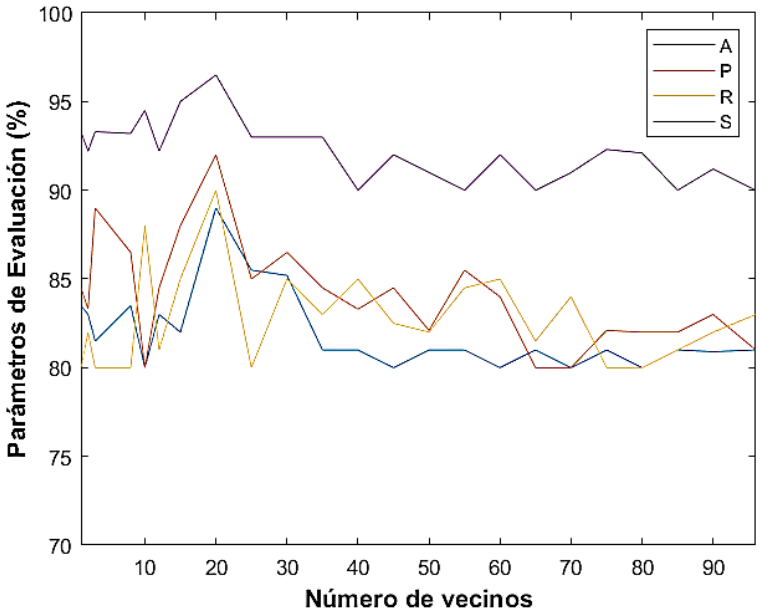

Figura 17. Parámetros de evaluación en función del número de vecinos después de evaluar 53 características normalizadas

Tabla 9

Porcentaje de los parámetros de evaluación en función al número de vecinos

Entrenamiento		**Evaluación**				
		N. Vecinos	**Exactitud**	**Precisión**	**Sensibilidad**	**Especificidad**
Weighted KNN	**53 características**	1	81%	84.5%	80%	93.3%
		2	81%	83.3%	82%	92.2%
		3	83%	89%	80%	93.3%
		8	81%	86.5%	80%	93.2%
		10	87%	90%	90%	96.5%
		12	80%	84.5%	81%	92.2%
		15	85%	88%	85%	95%
		20	88.5%	92%	90%	96%
		25	80%	90%	90%	93%
		35	80%	90%	90%	93%
		40	81%	83.3%	80%	90%
		45	80%	84.5%	82%	92%
		50	81%	82.1%	82%	91%
		55	81%	85.5%	90%	90%
		60	80%	84%	85%	92%
		65	81%	80%	90%	90%
		70	80%	80%	84%	91%

	75	81%	82.1%	80%	92.3%
	80	80%	82%	80%	92.1%
	85	81%	82%	81%	90%
	90	87%	83%	82%	91.2%
	96	81%	81%	83%	90%

En la Figura 17, se observa que el número de vecinos que maximiza todos los parámetros de evaluación es *k*=20, este valor se puede corroborar en la Tabla 9, que indica de forma más específica los porcentajes obtenidos. En el cual se obtuvo una exactitud del 88.5%, una precisión de 92%, una sensibilidad del 90% y una especificidad del 96%.

4.1.5 Análisis - Matriz de confusión

Tras la selección del clasificador, número de vecinos y características correctos, se evalúo con el conjunto de prueba de 60 audios. En el caso de características normalizadas y sin procesamiento, se obtuvo la misma matriz de confusión, como se indica en la Tabla 10 el porcentaje de acierto más alto se obtuvo para enojo con un 93% de acierto, respectivamente, mientras que los más bajos se obtuvo para feliz, triste y miedo con un 87% de acierto.

Tabla 10
Matriz de confusión para características sin procesamiento y normalizadas

			Salida		
Entrada	Feliz	Triste	Miedo	Enojo	**Acierto (%)**
Feliz	13	1	0	1	87
Triste	0	13	1	1	87
Miedo	0	1	13	1	87
Enojo	1	0	0	14	93
Acierto Global (%)					**88.5**

4.2 Análisis del modelo de clasificación de mujeres

4.2.1 Primer experimento

El primer experimento se realizó tras la extracción de las 68 características, estas fueron ingresadas a dos modelos diferentes de clasificadores a la *App Clasification Learner* de la

Plataforma Matlab® wKNN y SVM; con sus respectivos tipos independientemente. Se modificó inicialmente en el caso de wKNN el número de vecinos para *k=1, k=10, k=15, k=20, k=35, k=40, k=50, k=60, k=70, k=80 y k=96;* y para SVM se modificó el parámetro gamma para γ *=1, =2.1, =8.2,* y la función de costo *C=1* permaneció constante, ya que se según *Matlab*® estos valores son óptimos para cada tipo de clasificador dentro de SVM, posteriormente se extrajo cada modelo de entrenamiento y se lo evalúo con el conjunto de prueba de 60 audios. Como se puede observar en la Tabla 11, se indica los porcentajes de exactitud para cada tipo de clasificador obtenidos de Matlab® los porcentajes de los tres clasificadores más altos son *Quadratic SVM* con 81.3% de exactitud, Medium Gaussian SVM con 78.1% y *Weighted KNN* con k=10 con un porcentaje de 71%.

Tabla 11

Porcentaje de exactitud de diferentes tipos de clasificadores

Entrenamiento			**Evaluación**
Clasificador	**C**	γ	**Exactitud (%)**
Quadratic SVM	1	1	81.3
Fine Gaussian SVM	1	2.1	61.9
Medium Gaussian SVM	1	8.2	78.1
Weighted KNN	**k**		
	1		65
	10		71
	15		68.8
	20		70
	35		59.5
	40		55
	50		56
	60		59
	70		58
	80		52
	96		53

Según los resultados obtenidos en la primera etapa, el modelo de clasificador óptimo para la detección de emociones en el caso de las mujeres es *Quadratic SVM*, con $\gamma=1$ *y* función de costo *C* equivalente a 1 ya que según Matlab®, es el valor eficaz para QSVM.

4.2.2 Segundo experimento - Selección de características con el método FSV

Tras la selección del modelo de clasificador óptimo y con el objetivo de mejorar el rendimiento del sistema, se realizó una selección de características con el método *Feature selection via concave minimization and support vector machines* (FSV), que es tipo *wrapper* y es el método apropiado para el modelo QSVM, por lo cual el método FSV evitó el ajuste excesivo de datos y redujo el costo computacional eliminando características irrelevantes para el estudio. En la Tabla 12 se puede observar que en primer lugar dependiendo del número de características se extrajo el modelo qSVM y se lo evalúo con el conjunto de prueba de 60 audios.

Tabla 12

Selección de características mediante el método FSV

Entrenamiento		**Evaluación**
	N. Características	**Exactitud (%)**
Quadratic SVM	68	81
	63	81
	58	81
	57	90
	56	85
	50	85
	43	80

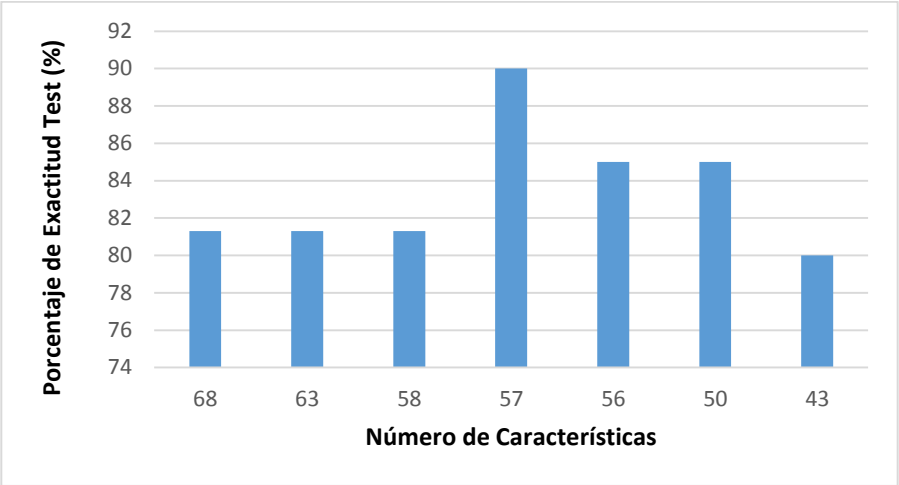

Figura 18. Selección de características utilizando el método FSV

Como se observa en la Figura 18 se realizó una evaluación y reducción paulatina del número de características irrelevantes en el cual se determinó que emplear 57 características (Anexo C) en lugar de las 68, ofrece el rendimiento del 90% y menora el costo computacional del sistema, caso contrario de lo que sucede al utilizar 56 características, que ofrece un rendimiento del 85%. Para realizar estas pruebas se trabajó con el parámetro *gamma=1* y función de costo C=1. La Tabla 13 indica las 10 características más importantes obtenidas a través del método FSV. En la cual el *pitch* tiene mayor importancia para la clasificación de emociones en el caso de mujer.

Tabla 13

Orden de características obtenidas a través del método FSV

Orden	Característica
1	Pitch
2	Pitch Wavelet
3	Media Curtosis Wavelet
4	Media de la Asimetría
5	Media centroide
6	Media Flujo spectral Wavelet
7	Jitter Wavelet
8	Media SNR
9	Media Curtosis
10	Jitter

4.2.3 Tercer experimento

Se realizaron pruebas con datos sin procesamiento, con datos normalizados los cuales se obtienen con la fórmula "*x= x / max(abs(x))",* la cual transforma a la matriz original en otra de mismo tamaño, normalizando los valores con respecto al más alto y finalmente se analizó con datos estandarizados que se centran para tener una media 0 y desviación estándar 1, en la cual se empleó la función de Matlab® "*zscore*". Se determinó con qué tipo de dato se maximizan todos los parámetros de evaluación del clasificador, como se observa en la Tabla 14. Para realizar estas pruebas se trabajó con el parámetro *gamma=1* y función de costo C=1. Una vez realizado la extracción del modelo de clasificador para cada tipo de características se obtienen los parámetros de evaluación mediante los resultados presentados en los 60 audios que son parte del conjunto de pruebas.

Tabla 14

Parámetros de evaluación en función del número de vecinos después de evaluar 53 características sin procesar, normalizadas y estandarizadas

Entrenamiento			**Evaluación**			
Q SVM	**57 C.**	**Tipo Datos**	**Exactitud**	**Precisión**	**Sensibilidad**	**Especificidad**
		Sin Procesar	90%	90%	90%	96.5%
		Normalizados	75%	78.25%	73%	81.75%
		Estandarizados	85%	85.8%	85%	94.75%

En la Tabla 14 se indica los resultados obtenidos tras la comparación de los diferentes tipos de datos, para lo cual se puede observar que el tipo de datos sin procesamiento maximiza todos los parámetros de evaluación obteniendo de esta manera en exactitud un porcentaje del 90%, precisión 90%, sensibilidad 90% y en especificidad el porcentaje más alto con 96.5%. Caso contrario lo que sucede con los datos normalizados que presentaron porcentajes por debajo del 80% en todos sus parámetros.

4.2.4 Análisis - Matriz de confusión

Tras la selección del clasificador, número de características correctas, los resultados se evaluaron con el conjunto de prueba de 60 audios.

4.2.4.1 Características sin procesamiento

En el caso de características sin procesamiento, se obtuvo la matriz de confusión, como se indica en la Tabla 15 el porcentaje de acierto más alto se obtuvo para miedo y enojo con un 93% de acierto, respectivamente, mientras que los más bajos se obtuvo para feliz y triste con un 87% de acierto. Para realizar estas pruebas se trabajó con el parámetro *gamma=1* y función de costo C=1.

Tabla 15

Matriz de confusión para características sin procesamiento

	Salida				
Entrada	Feliz	Triste	Miedo	Enojo	**Acierto (%)**
Feliz	13	0	1	1	87
Triste	1	13	1	0	87
Miedo	0	1	14	0	93
Enojo	0	0	1	14	93
Acierto Global (%)					**90**

4.2.4.2 Características normalizadas

En el caso de características normalizadas, se obtuvo la matriz de confusión, como se indica en la Tabla 16 el porcentaje de acierto más alto se obtuvo para miedo y enojo con un 80% de acierto, respectivamente, mientras que el más bajo se obtuvo para triste con un 66% de acierto. Para realizar estas pruebas se trabajó con el parámetro *gamma=1* y función de costo C=1.

Tabla 16

Matriz de confusión para características normalizadas

	Salida				
Entrada	Feliz	Triste	Miedo	Enojo	**Acierto (%)**
Feliz	11	1	2	1	73
Triste	1	10	3	1	66
Miedo	0	1	12	2	80
Enojo	0	0	3	12	80
Acierto Global (%)					**75**

4.2.4.3 Características estandarizadas

En el caso de características normalizadas, se obtuvo la matriz de confusión, como se indica en la Tabla 17 el porcentaje de acierto más alto se obtuvo para enojo y miedo con un 93% de acierto, respectivamente, mientras que el más bajo se obtuvo para feliz con un 73% de acierto. Para realizar estas pruebas se trabajó con el parámetro *gamma=1* y función de costo C=1.

Tabla 17

Matriz de confusión para características estandarizadas

	Salida				
Entrada	Feliz	Triste	Miedo	Enojo	**Acierto (%)**
Feliz	11	1	1	2	73
Triste	1	12	2	0	80
Miedo	0	0	14	1	93
Enojo	0	1	0	14	93
Acierto Global (%)					**85**

Nota: Todos los resultados presentados previamente, se obtuvieron en base al programa desarrollado en Matlab®, detallado en el Anexo D.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5 Conclusiones y recomendaciones

- Las características prosódicas, espectrales y calidad de voz presentadas en el estado del arte se extrajeron de dos maneras datos sin procesamiento y mediante la transformada wavelet *Daubechies N=5,* puesto que ofrece una mejor localización de frecuencia y se aplicaron 3 niveles, dando como resultado un total de 68 características.
- Se concluyó que, con las 68 características extraídas, se obtuvo que para el caso del hombre el modelo con el porcentaje de exactitud más alto fue *Weighted k-Nearest-Neighbors (wkNN),* con 68.8% al utilizar *k*=10. Para el caso de las mujeres el modelo de entrenamiento óptimo resultó ser *Quadratic SVM* con 81.3% al utilizar gamma *=1 y* función de costo *C*=1.
- Se realizó la selección de características para cada género, en el caso de hombre se efectuó mediante el método *Feature Selection via Eigenvector Centrality* (ECFS), que es tipo *Filter* y es el apropiado para el modelo *wKNN*, del cual se concluyó que al utilizar 53 características en lugar de las 68, ofrece un rendimiento del 90% y menora el costo computacional del sistema.
- En el caso de las mujeres se utilizó el método *Feature selection via concave minimization and support vector machines* (FSV), que es tipo *wrapper* y es el método apropiado para el modelo QSVM del cual se determinó que emplear 57 características en lugar de las 68, ofrece el rendimiento del 90%.

- Se observó mediante métodos de selección de características propios de cada modelo de clasificador, que las características que más se destacaron en el estudio son el *pitch* para el caso de las mujeres y *shimmer* para el caso de los hombres.
- Para el caso de hombres se analizó dos tablas de entrenamiento con las 53 características, cada una con datos sin procesamiento y con datos normalizados, de las cuales independientemente se realizaron pruebas con un rango de vecinos *k* desde 1 hasta 96, y se determinó que con el modelo de datos normalizados y número de vecinos *k=20* se maximizaron todos los parámetros de evaluación exactitud 88.5%, precisión 92%, sensibilidad 90% y especificidad 96%, obteniendo así el mejor rendimiento.
- Para el caso de mujeres se realizó tres tablas de entrenamiento con 57 características, cada una con datos sin procesamiento, normalizados y estandarizados, de los cuales se concluyó que para mujeres el modelo de datos sin procesamiento maximiza todos los parámetros de evaluación obteniendo de esta manera en exactitud un porcentaje del 90%, precisión 90%, sensibilidad 90% y en especificidad el porcentaje más alto con 96.5%.
- El porcentaje de acierto más alto para mujeres se obtuvo para el estado emocional de enojo y miedo con el 93%, mientras que los más bajos fueron para feliz y triste con el 87%. Para el caso de hombres el más alto es para enojo con el 93% y los porcentajes más bajos son para felicidad, miedo y tristeza con el 87%. Según estudios psicológicos y culturales los hombres tienden a expresar menos sus emociones y en especialmente la de miedo, razón por la cual las mujeres presentaron un porcentaje de exactitud global más alto que el de los hombres.

- Se recomienda usar herramientas que permitan obtener un número menor de características, para evitar que el sistema aprenda el ruido de los datos y entorpezca la clasificación de las emociones.
- Tras realizar la selección de características, se evidenció que las características que no son recomendables usar en posteriores estudios, son en su mayoría las características espectrales, ya que para el caso de hombres las 15 características eliminadas presentaron un débil coeficiente de correlación y un bajo umbral de varianza, datos obtenidos mediante el método ECFS; y para el caso de mujeres las 11 características eliminadas presentaron un significativo número de características recursivas y sus valores eran muy cercanos a cero, por los que se los consideró despreciables, datos obtenidos mediante el método FSV.

CAPÍTULO VI

TRABAJOS FUTUROS

6. Trabajos Futuros

- Se propone incrementar el número de emociones del detector automático, con la finalidad de dar a conocer que por medio de la voz se puede determinar todos los estados de ánimo de una persona.
- Realizar un solo programa para detectar niveles de tristezas con fines médicos y psicológicos, con el fin de determinar el nivel de depresión que tiene una persona y poder brindarle la ayuda necesaria.
- Como trabajo futuro se propone aumentar el porcentaje de asertividad del sistema implementado para que este pueda ser utilizado individualmente sin ayuda de un especialista.
- Crear una base de datos espontánea y actuada, en el idioma español, para realizar posteriores estudios y validaciones del presente trabajo.
- Realizar una interfaz gráfica con fines lucrativos y comerciales.

CAPÍTULO VII

BIBLIOGRAFÍA

7. Bibliografía

Arroyo, V. (2013). Las emociones básicas y la ira. *Universidad San Pablo*.

Arsuaga, J. (2000). *La especia elegida.* Barcelona.

Azor, R. (2017). La transformada de wavelet. Revista de la Universidad De Mendoza.

Bair, D. (2011). Un análisis acústico de las vocales del K'ichee.

Bericat, E. (s.f.). Emociones. *Sociopedia*, 1-13.

Bolaños, J. (2018). Estadística: conceptos básicos y definiciones.

Bradley, S. (2015). Feature Selection via Concave Minimization and Support Vector Machine. Madison.

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

Carmona, J. (s.f.). El habla con emociones. 23-39.

Casado, A. (2002). La evaluación clínica de la voz: fundamentos médicos y logopédicos. Málaga: Ediciones Aljibe.

Castañeda, P. F. (2015). Aparato Fonador. 124-140.

Cavalcanti, A. N. (2010). Sistema Inteligente para Diagnóstico de Patalogias na Laringe Utilizando Máquinas de Vector de Suporte. Brasil.

Chávez, N. (2012). Aprendizaje no supervisado y el algoritmo wake-sleep en redes neuronales. *Universidad tecnologica de la mixteca*.

Diaz, S. (Noviembre de 2015). Características acústicas de las vocales del español de Chile producidas por sujetos residentes en la ciudad de Santiago. Revista Chilena de Fonoaudiología.

Dubuisson, T. (2009). On the Use of the Correlation between Acoustic Descriptors for the Normal/Pathological Voices Discrimination. *EURASIP*.

Dupuis, K. (2010). *Browsing "Toronto emotional speech set (TESS)" by Title.* Obtenido de https://tspace.library.utoronto.ca/handle/1807/24487/browse?type=title&submit_browse=Title

Duque, C. (2007). Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones. *Universidad Tecnológica de Pereira - UTP*.

Eldho, G. (5 de Mayo de 2017). Lie Detection Through Voice Stress Analysis. United States.

Fatiha, D. (Septiembre de 2016). Energy and Entropy Based Features for WAV Audio Steganalysis. Dubai.

Forbes, R. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. *HLT-NAACL 2004*, 201-208.

Garcia, I. (Junio de 2015). El otorrinolaringólogo ante el profesional de la voz. Madrid.

Gaya, G. (1999). *Elementos de fonética general.* Madrid.

Gervás, P. (2010). Expresión de emociones en la síntesis de voz en contextos narrativos. *Universidad Complutense de Madrid*.

Gómez, M. (2017). Procesamiento digital de señales. *Estadística, probabilidad y ruido*.

Gonzáles, R. (2013). *Producción de la voz y el habla.* Valdecilla.

Hechenbichler. (2015). *Weighted k-Nearest-Neighbor Techniques and Ordinal.* Sonderforschungsbereich.

Larrañaga, P. (2015). *Clasificadores K-NN.*

Liscombe, R. (2005). Using context to improve emotion detection in spoken dialog systems. *Eurospeech*.

Matlab. (2019). *Mathworks*. Obtenido de https://la.mathworks.com/help/wavelet/ref/wavedec.html

Mora, S. (2015). Manipulación de señales de audio. México.

Morales, M. (Abril de 2007). Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones. Pereira.

Nam, U. (28 de Abril de 2001). Special Area Exam Part II. Stanford.

Nilsson, N. J. (1998). Introductioin to Machine Learning.

Nuñez, L. E. (2010). Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo. *INAOE*.

Oliveira, N. (27 de 9 de 2019). Wavelets de Haar y Daubechies y sus aplicaciones.

Petisco, J. (2010). A veces la voz dice más que las palabras.

Planet, S. (2009). Contribution to the Interspeech 2009 Emotion Challenge . *Interspeech. - Brighton, U.K.*

Rabiner, L. (1978). Digital Processing of Speech Signals. New Jersey: Upper Saddle River .

RAVDESS. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *A dynamic, multimodal set of facial and vocal expressions in North American English*. Rusia.

Roffo, G. (6 de 8 de 2018). Feature Selection Library (MATLAB Toolbox). Glasgow.

Shwartz, S. (2014). Understanding Machine Learning: From Theory to Algorithms. *Cambridge University Press*.

Sintas, P. (29 de 09 de 2013). Análisis Acústico de la Voz PRAAT. Valdivia.

Soler, J. (2010). La casa de las emociones. *Ecología emocional*, 1-2.

Teixeira, P., Barbosa, D., & Moreira, S. (2011). Análise Acústica Vocal. *Escola Superior de Tecnologia e Gestao*.

Vivas, M. (2001). Educar las emociones. 23-32.

Vizzarri, P. (2016). Transformada Wavelet en el An´alisis de Se˜nales.

Vuletich, J. (2005). Nuevas bases para el procesamiento de música en el dominio de tiempo-frecuencia. *Universidad de Bueno Aires*.