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: IMPLEMENTACIÓN DE UN SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DE ENGAÑOS MEDIANTE EL ANÁLISIS DE LA SEÑAL DE LA VOZ

AUTOR: BRAVO PAREDES, SANTIAGO ANDRÉS

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, "**IMPLEMENTACIÓN DE UN SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DE ENGAÑOS MEDIANTE EL ANÁLISIS DE LA SEÑAL DE LA VOZ**", fue realizado por el señor **Bravo Paredes, Santiago Andrés** 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. C. 1709775637

AUTORÍA DE RESPONSABILIDAD

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORÍA DE RESPONSABILIDAD

Yo, **Bravo Paredes, Santiago Andrés**, declaro que el contenido, ideas y criterios del trabajo de titulación: **"Implementación de un sistema de reconocimiento automático de engaños mediante el análisis de la señal de la 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:

Santiago Andrés Bravo Paredes

C.I. 1723837868

AUTORIZACIÓN

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORIZACIÓN

Yo, **Bravo Paredes, Santiago Andrés**, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: "**Implementación de un sistema de reconocimiento automático de engaños mediante el análisis de la señal de la voz**", en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 24 de junio del 2019

Firma:

Santiago Andrés Bravo Paredes

C.I. 1723837868

DEDICATORIA

Dedico este trabajo de titulación a cada persona quienes con sus enseñanzas, esfuerzos, apoyo y sobretodo amor, permitieron que pueda alcanzar un logro más, que sin lugar a duda ha sido gracias a cada uno de ellos, amados padres, abuelos y hermanos Javier, Rosita, Luis y Sara que día a día con su apoyo consejos y amor han sabido inculcarme valores infinitas gracias, a Evelyn por todo el apoyo y cariño brindado en todo el proceso.

Santiago Andrés Bravo Paredes

AGRADECIMIENTO

A Dios que supo guiar mi camino en esta ardua carrera para alcanzar esta meta, y supo darme una gran familia que brindo conocimientos, consejos y apoyo durante lo largo de mi vida e hicieron lo posible para que mis metas llegasen a cumplirse.

A mis maestros y director de tesis, por guiarme a lo largo de mi carrera universitaria y haberme compartido conocimientos importantes para la misma.

Santiago Andrés Bravo Paredes

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

En las últimas décadas con la evolución de las telecomunicaciones ha aumentado la inseguridad, estafa y engaño por este medio, por lo que con la finalidad de disminuir tipos de estafa por medio de engaños, en este proyecto de investigación se estudia el estrés en la voz de las personas para la detección de engaños a partir de características específicas de la voz, utilizando para ello la herramienta Matlab®. Para la detección del engaño se evalúa las variaciones en el conjunto de características propias del habla neutral, en comparación con el habla producida bajo el estrés, entre éstas están la calidad de la voz, prosódicas y glóticas, y dentro de estas, frecuencia fundamental *(Pitch)*, micro temblor *(Jitter)*, *Shimmer*, *Sharpes*, entre otras, teniendo un total de 68 características extraídas por señales de audio sin pre-procesamiento y con el uso de transformada de Wavelet. Finalmente un grupo de 40 características, determinadas con selección de características (*feature selection*), aplicado a una base de datos de 94 señales de audio, son las utilizadas para realizar el reconocimiento automático de engaños por medio de clasificadores, corroborando que las características extraídas brindan los datos necesarios para clasificar una señal de voz como engaño o verdad, con baja tasa de error en cuatro parámetros medidos que son exactitud, precisión, sensibilidad y especificidad. La importancia que tiene el sistema es que el engaño puede ser detectado por medio de grabaciones de la voz, lo cual no invade la privacidad de las personas que están siendo entrevistadas o interrogadas.

PALABRAS CLAVE:

- **SEÑALES DE VOZ**
- **ANÁLISIS DE LA SEÑAL DE VOZ**
- **CLASIFICADORES BICLASES**
- **DETECCIÓN DE ENGAÑOS**

ABSTRACT

In recent decades with the evolution of telecommunications has increased insecurity, fraud and deception by this means, so in order to reduce scam types by means of deception, this research project studies the stress in the voice of people for the detection of deception based on specific characteristics of the voice, using the tool Matlab®. For the detection of deception, the variations in the set of characteristics of neutral speech are evaluated, in comparison with the speech produced under stress, among these are the quality of speech, prosodic and glottic, and within these, fundamental frequency (Pitch), micro tremor (Jitter), Shimmer, Sharpes, among others, having a total of 68 features extracted by audio signals without pre-processing and with the use of Wavelet transform. Finally a group of 40 characteristics, determined with feature selection, applied to a database of 94 audio signals, are those used to perform the automatic recognition of deception by means of classifiers, corroborating that the extracted characteristics provide the necessary data to classify a voice signal as deception or truth, with low error rate in four measured parameters that are accuracy, precision, sensitivity and specificity. The importance of the system is that deception can be detected through voice recordings, which does not invade the privacy of the people being interviewed or questioned.

KEYWORDS:

- **VOICE SIGNALS**
- **ANALYSIS OF THE VOICE SIGNAL**
- **BICLASE CLASSIFIERS**
- **DECEPTION DETECTION**

CAPÍTULO I

INTRODUCCIÓN DEL PROYECTO DE INVESTIGACIÓN

1. Introducción del proyecto de investigación

1.1 Antecedentes y justificación del proyecto

El engaño, en la mayoría de las personas, provoca estrés psicológico que tiene un reflejo evidente en la cara, la voz, las extremidades y los movimientos para caminar. Las técnicas de análisis de voz para la detección de engaños tienen mucho interés en el desarrollo de muchas aplicaciones para sectores muy diferentes, como banca y seguros, servicios al cliente, análisis de evidencia forense e incluso las compañías de telecomunicaciones utilizan esta tecnología para detectar fraudes en llamadas entrantes. Los departamentos de policía y bomberos, hospitales y otros centros de llamadas de emergencia también podrían beneficiarse de esta aplicación de tecnología porque reciben una gran cantidad de llamadas telefónicas que deben analizarse y filtrarse cuidadosamente para tomar decisiones rápidas de acuerdo con su confianza y relevancia para priorizar las actuaciones y optimizar los recursos (Liu, 2005).

El reconocimiento de engaño por medio de la voz ha sido solicitado por varias instituciones de Defensa Nacional a nivel mundial, tal es el estudio realizado para el Departamento de Justicia de los Estados Unidos (Haddad, Walter, Ratley, & Smith, 2002), en el cual se hace un análisis para detectar el engaño mediante una ligera oscilación en la contracción de los músculos, incluidas cuerdas vocales que vibran en un rango de 8 a 12 [Hz] (micro temblores), de aproximadamente 10 ciclos por segundo. Se afirma que estos micro temblores en la voz, cambian cuando una persona está mintiendo. El sistema presentado detecta una variación en la oscilación de las cuerdas vocales, es entonces cuando se asume que la persona entrevistada está mintiendo.

En el trabajo de investigación de Liu (Liu, 2005), se examina la frecuencia fundamental (Pitch) y el micro temblor (Jitter), de grabaciones de voz de varias personas, y se centra en la posibilidad de detectar el engaño a través de la voz estresada del ser humano. El experimento que presenta está diseñado para analizar la probabilidad de corrección de detección mediante el uso de las características antes mencionadas por separado, entre otras.

En el laboratorio de procesamiento del habla neuromórfica determinan el engaño mediante el análisis de 72 características de la voz, en las cuales están el Pitch, *Jitter*, *Shimmer*, *Sharpness*, entre otras, que se consideran importantes para este estudio, el estudio concluye que los centros de procesamiento neurológico se alteran cuando una persona se ve obligada a fabricar una opinión artificial, que está en contra de sus ideas, de igual manera la entrevista se debe realizar una sola vez a la misma persona ya que esta puede perder la espontaneidad al repetir la entrevista. (Rodellar, Palacios, Nieto, & Gómez, 2015).

En base a estudios realizados se ha percibido la importancia de determinar la veracidad de la información que una persona transmite, por lo que el presente análisis sobre el engaño a partir de la voz, tiene un gran aporte a la sociedad debido a que la aplicación puede ser empleada a nivel gubernamental y social tales como peritajes judiciales, estafas telefónicas, llamadas falsas, entre otras.

Existen varias bases de datos de habla emocional espontánea y simulada. (El Ayadi, Kammel, & Karray, 2011) Pero con respecto al estrés del habla, existen muy pocas bases de datos por el nivel de confidencialidad de cada entidad. Por lo que el presente trabajo trata sobre el estudio del estrés en la voz de hispanohablantes. No hay bases de datos disponibles para cumplir con este objetivo por lo que mediante las características obtenidas en estudios anteriores y bases de datos

pequeñas en otros idiomas se entrenará el software por el cual se realizará una base de datos de hombres y mujeres etiquetada en el idioma español.

1.2 Objetivos de la investigación

1.2.1 Objetivo general

Implementar un sistema de reconocimiento automático de engaños mediante el análisis de la señal de la voz.

1.2.2 Objetivos específicos

- Obtener una base de datos actuada, clasificada y etiquetada de verdades y mentiras, mediante el estudio del estado del arte.
- Identificar las características principales de la voz, mediante el estado del arte para la detección del engaño.
- Evaluar el sistema en comparación con los diferentes tipos de clasificadores (KNN, SVM, Árboles de decisión).
- Reducir el costo computacional, mediante técnicas de selección de características.
- Detectar el engaño a partir de las características extra-lingüísticas de la voz y entrenar el sistema mediante la teoría de *Machine Learning* a través del uso de la técnica de clasificación supervisada.

CAPÍTULO II

MARCO TEORICO

2. Marco teórico

2.1 Sistema de producción vocal

Para entender la producción del habla humana, primero se debe estudiar la anatomía del sistema de producción vocal. Es justo decir que se debe comprender el proceso de producción del habla, modelando un sistema que permita entenderlo y de esta manera crear sistemas que reconozcan sus características y matices distintivos de cada hablante (Beigi, 2018).

La voz es el resultado de un proceso físico que realiza el aparato fonador con la finalidad de satisfacer la necesidad de comunicación. Además es una señal acústica, es decir, una onda de presión longitudinal formada por la compresión y expansión de las moléculas de aire que se transmite de forma paralela a la aplicación de la energía (Duque Sanchez & Morales Perez, 2007).

2.1.1 Aparato fonador

El término fonética está referido al estudio de los sonidos que son producidos por el sistema vocal humano independientemente del idioma que se hable y la fonación trata con la energía acústica generada por las cuerdas bucales en la laringe, los diferentes tipos de fonación sin voz, con voz y susurro (Beigi, 2018).

La fonación no vocalizada puede ser en forma de fonación nula que se corresponde con energía cero o fonación de respiración que se basa en pliegues vocales relajados que provocan una corriente de aire turbulenta. La mayoría de los sonidos se generan a través de la fonación sonora normal, que es cuando las cuerdas vocales vibran a una frecuencia periódica y generan cierta resonancia en la cámara superior del tracto vocal. Otra categoría de la fononización de la voz se llama

laringalización (voz chirriante). Es cuando los cartílagos aritenoides fijan la porción posterior de las cuerdas vocales, permitiendo que la parte anterior de las cuerdas vocales vibre. Otro tipo de fonación con voz es un falsete que es la creación poco natural de una voz de tono alto al apretar la forma básica de las cuerdas vocales para lograr un tono alto falso (Beigi, 2018).

Para la producción de la voz intervienen órganos del sistema respiratorio y digestivo mismos que son controlados por el sistema nervioso central (Hogset, 1995), (Moulines & Laroche, 1995). Principalmente generada por la excitación en las cuerdas vocales que se propaga a través de la faringe y cavidades nasal y bucal (Huang, 2001). El sistema fonador se puede clasificar en tres bloques (Duque Sanchez & Morales Perez, 2007):

- **Sistema resonante:** está conformada por tres cavidades articulatorias: faríngea, oral y nasal. Los sonidos producidos por este sistema se desplazan hasta los orificios nasales y la boca desde las cuerdas vocales, haciendo que el sonido se modifique y amplifique.
- **Sistema de generación:** este sistema se caracteriza por producir un exceso en la corriente de aire por medio de los músculos abdominales y torácicos aumentando la presión en los pulmones. Y de esta manera excitar al sistema de vibración.
- **Sistema de vibración:** es conformado por las cuerdas vocales las cuales se clasifican en superiores e inferiores, de estas solo las inferiores participan en la producción de la voz.

2.2 El engaño en el habla

La mentira no es un aspecto raro en la vida diaria del ser humano, por lo general las personas tienden a mentir conscientemente para obtener cierto tipo de beneficio con ello, además las personas creen que mentir en situaciones necesarias no es ofensa, como por ejemplo en entrevistas, citas, cuando se les pregunta sobre sus pasatiempos o debilidades, entre otras. Esta tendencia de

los humanos se convierte en un gran problema durante la investigación de un crimen, la prueba del polígrafo ha sido una herramienta muy utilizada para la detección del engaño (Jithin, Nandan, Eldho, Akhil, & Sreehari, 2017), pero es un dispositivo invasivo a la privacidad de la persona por lo que en las últimas décadas la detección del engaño a través de la voz se ha vuelto cada vez más popular por parte de entes gubernamentales, compañías de seguros y de telecomunicaciones utilizando esta tecnología en llamadas telefónicas entrantes para detectar fraudes. (Xianfeng, 2005)

Cuando una persona se encuentra en tensión la vibración muscular aumenta, incluidas las cuerdas vocales, vibran en el rango de 8 a 12 Hz, y cuando la persona miente la vibración aumenta más que la frecuencia normal (Jithin, Nandan, Eldho, Akhil, & Sreehari, 2017) y esto crea estrés en la voz, estas vibraciones inaudibles son conocidas como micro temblores en la voz. También se toman en cuenta características audibles de la voz humana, tales como la frecuencia fundamental (*Pitch*), y el micro temblor (*Jitter*), entre otras con el propósito de detectar el engaño a través de grabaciones de la voz estresada del ser humano (Haddad, Walter, Ratley, & Smith, 2002).

2.2.1 El estrés

La definición común del estrés involucra un problema mecánico (Xianfeng, 2005): "... cuando un estrés se aplica a un cuerpo... se produce una tensión correspondiente" (Murray, Baber, & South, 1996).

Sin embargo, esta definición no es tan útil porque, a pesar de que esta relacionando el estrés con el cuerpo, el estado emocional de las personas es desconocida, por lo que otra definición mencionada en (Murray, Baber, & South, 1996) es:

"El estrés es una variabilidad observable en ciertas características del habla debido a una combinación de respuesta inconsciente a factores estresantes y / o control consciente".

Esta definición no solo enfatiza que la naturaleza del estrés es variable, sino que también se refiere a un estado sin estrés, el problema es que no está claro la condición del estado sin estrés, no está claro que características o combinaciones de variaciones de características se correlacionan con diferentes factores del estrés (Murray, Baber, & South, 1996).

En ESCA-NATO taller sobre el habla bajo estrés, se propuso una definición en la cual se separa la causa del estrés y el efecto del estrés (Xianfeng, 2005).

"El estrés es un efecto en la producción del habla (manifestada a lo largo de un rango de dimensiones), causado por la exposición a un factor estresante"

De acuerdo con esta definición, el estrés en la voz puede ser definida como una causa y también como un efecto, en otras palabras el estrés es un efecto sobre el humano causada por un factor estresante (Hollien, Geison, & Hicks, 1987).

No es sencillo tener una sola definición aceptable del estrés que satisfaga todas las perspectivas basadas en diferentes dominios de investigación, el concepto del estrés es amplio y se basa en estudios específicos, dependiendo del área de investigación se debe hacer énfasis en una particular definición del estrés y no se considerará como incorrecto, por lo tanto no es necesario establecer una definición del estrés de la voz unificada (Xianfeng, 2005) .

2.2.2 Análisis del estrés en la voz (VSA)

El análisis del estrés en la voz originada del concepto de que cuando una persona está bajo estrés, específicamente si el hablante está bajo peligro, el cuerpo se prepara para la lucha poniendo en tensión sus músculos. Los cambios en la señal acústica del habla debido al estrés son causados principalmente por estas reacciones de estrés. Estos cambios también afectan los órganos que producen el habla como la respiración y la tensión muscular. Por lo tanto debería ser posible

establecer si la persona está estresada simplemente analizando su voz (Beers & Berkow, 1999). La terminología de la vibración del músculo es temblores de micro-músculos (MMT) o micro temblor. Los micros temblores se producen en los músculos que forman el tracto vocal que se transmiten a través del habla. Se describe como una ligera oscilación a varios ciclos cada segundo, lo que se afirma que es la única fuente de detección si una persona está mintiendo (Haddad D. , 2002).

En varios trabajos sobre el análisis del estrés se centra en la comunicación bajo condiciones peligrosas, y en muchos de estos estudios se determina que la frecuencia fundamental tiene un aumento considerable. William y Stevens (News, 2003) concluyeron que bajo estas condiciones existe un aumento en el rango de la frecuencia fundamental y fluctuaciones abruptas del contorno de frecuencia fundamental, con un aumento del estrés.

De acuerdo con las teorías antes mencionadas, se han inventado dispositivos de detección del engaño. Estos dispositivos ofrecen grandes y potenciales ventajas frente al polígrafo (ver Tabla 1). Primero, el tiempo de capacitación es menor que el del polígrafo, y no requiere aspectos académicos para recibir esa capacitación. El análisis del estrés en la voz (VSA) toma poco tiempo, con un promedio de 45 minutos por sesión. Y el aspecto más importante es que el método no es invasivo con la persona debido a que no se colocan sensores en su cuerpo, solo un pequeño micrófono que graba la entrevista. Para el análisis no se requiere la presencia de la persona solo la grabación de la voz.

Tabla 1

Comparación de costos entre VSA y Polígrafo

	Analizador de estrés de voz de computadora	Polígrafo computarizado
Costo inicial del sistema	$9250,00	$13000,00
Matrícula para 1 estudiante.	$1215,00	$3000
Tiempo de capacitación.	6 días	8 semanas
Costo de habitación $70 por día.	$420,00	$3920,00
Salario para el estudiante durante la capacitación	$769,23	$6153,84
Numero de exámenes que un examinador puede dirigir por día.	7 exámenes	2 exámenes
Porcentaje promedio de resultados inconclusos en los exámenes.	0%	20%
El dispositivo puede analizar audios para verificar la verdad	Si	No
Las drogas, condiciones médicas o edad afectan el análisis	No	Si

Fuente: (Haddad D. , 2002)

Los sistemas VSA se dividen en dos categorías que son: sistemas basados en energía y sistemas basados en frecuencia. La mayoría de los sistemas evaluados se basan en la detección del micro temblor (Hopkins, Benincasa, Ratley, & Grieco, 2005). En este estudio se realizó un experimento con el cual se afirmó que cuando los datos de voz se procesan a través de un banco de filtros se puede obtener una serie de formas de onda que representan el estrés en la voz. Si el resultado del análisis de voz muestra una respuesta sin estrés, la forma de onda se muestra como un árbol de navidad (ver Figura 1), y a medida que el estrés aumenta la forma de onda se muestra más plana (ver Figura 2 y 3), de esta manera la forma de onda que muestra signos de estrés significativo se puede etiquetar como engañosa. Este tipo de tecnología es conocida como VSA basada en energía.

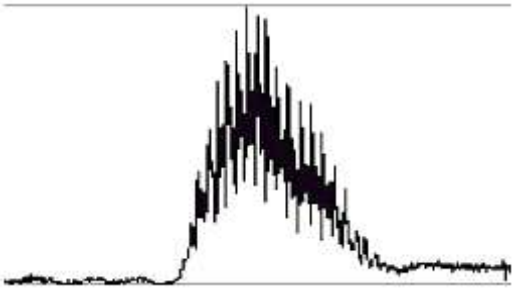

Figura 1. Forma de onda utilizada para medir la tensión basada en la energía en la forma de onda (Poco o sin estrés)
Fuente: (Hopkins, Benincasa, Ratley, & Grieco, 2005).

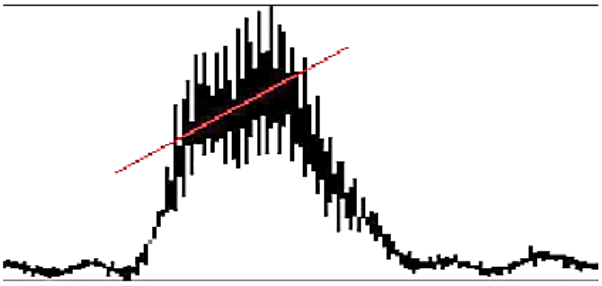

Figura 2. Forma de onda utilizada para medir la tensión basada en la energía en la forma de onda (Estrés medio)
Fuente: (Hopkins, Benincasa, Ratley, & Grieco, 2005).

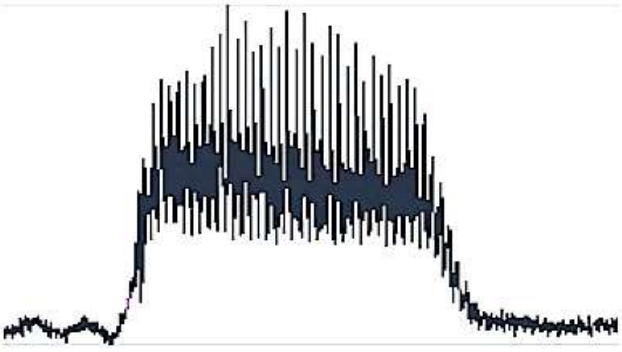

Figura 3. Forma de onda utilizada para medir la tensión basada en la energía en la forma de onda (Mucho estrés)
Fuente: (Hopkins, Benincasa, Ratley, & Grieco, 2005).

El otro tipo de sistemas son los basados en frecuencia, los cuales pueden identificar cambios dentro de las bandas de frecuencia y la distribución de frecuencias dentro de esta banda (Hopkins, Benincasa, Ratley, & Grieco, 2005). En este tipo de sistemas, se puede identificar un rango

continuo de estrés y se utiliza una comparación de la posición del estrés relevante en este rango para determinar si las respuestas son o no engañosas. El propósito de VSA es analizar los niveles de vibraciones que son medidas del estrés. Sin embargo, el estrés puede ser provocado por un pensamiento o una memoria, como recordar alguna situación peligrosa. Debido a la diferencia del entorno y situación, las diferentes personas pueden presentar diferentes niveles de estrés y este además puede variar con el estado de ánimo de la persona (Scherer & Oshinsky, 1977).

2.2.3 Detección del engaño

Tecnologías del análisis del estrés de la voz (VSA), han sido introducidas en para la detección de engaños o mentiras, las cuales son consideradas más convenientes y precisas en comparación con el método tradicional del polígrafo. En la actualidad agencias de policía están utilizando esta herramienta para detectar fraudes. Por ejemplo en Reino Unido, una aseguradora de automóviles (*Highway Insurance*) la cual introdujo detectores de engaños o mentiras telefónicas, se dice que la cuarta parte de todos los reclamos por robo de vehículos se han retirado desde que se comenzó a emplear este análisis (News, 2003). Otro ejemplo es el de España país donde La Policía Nacional ya tiene implementado un sistema informático para detectar las denuncias denominadas como falsas, este tiene una precisión del 90%, que es un porcentaje considerablemente alto comparado con la asertividad de un agente policial del 75%. Este último sistema se denomina “VeriPol”, y se encuentra habilitado en todas las comisarías de España y con ella el índice de resolución de este tipo de delitos mejora en un 80% (Español, 2018).

La detección de engaños tiene la finalidad de determinar si la información procesada es o no engañosa. Existen dos tipos de análisis y enfoques para detectar el engaño: el enfoque manual y el enfoque automático, este último más eficiente y sencillo de usar. Sin embargo en ciertas

situaciones, aun es útil y necesario el uso de la detección manual por parte de expertos en el área (DePaulo, Stone, & Lassiter).

Un estudio basado en el análisis del estrés de la voz (VSA) por el Instituto Nacional para la Verificación de la Verdad (NITV), determina que VSA se basa en el principio de determinar cambios en los parámetros asociados con la disipación involuntaria del componente FM en la voz. Estos cambios que están relacionados directamente con el estrés psicológico inducido por el miedo, la ansiedad, la culpa o el conflicto pueden ser útiles para la detección del engaño (A real case solved by CVSA , 2001).

Un ejemplo es la aplicación práctica de VSA: un caso real de abuso infantil resuelto por *Computer Voice Stress Analyzer* (CVSA) fue publicado por el NITV (A real case solved by CVSA , 2001). La víctima fue un niño de dos años que vivía con su madre y una compañera de habitación de su madre, los dos fueron sometidos al análisis de CVSA y se obtuvieron las gráficas mostradas en la Figura 4, las cuales al ser comparadas con las Figuras 1-3, se concluyó que los dos sospechosos estaban muy estresados en el momento de la entrevista (note las figuras con círculos rojos). Se informó que: el análisis del estrés en la voz debe ser considerado como una ayuda de investigación que debe usarse una vez que los investigadores hayan recopilado toda la información sobre el caso. Y para determinar si la información es engañosa o no el dispositivo debe tener un alto nivel de precisión y el profesional debe tener una gran habilidad de interrogación.

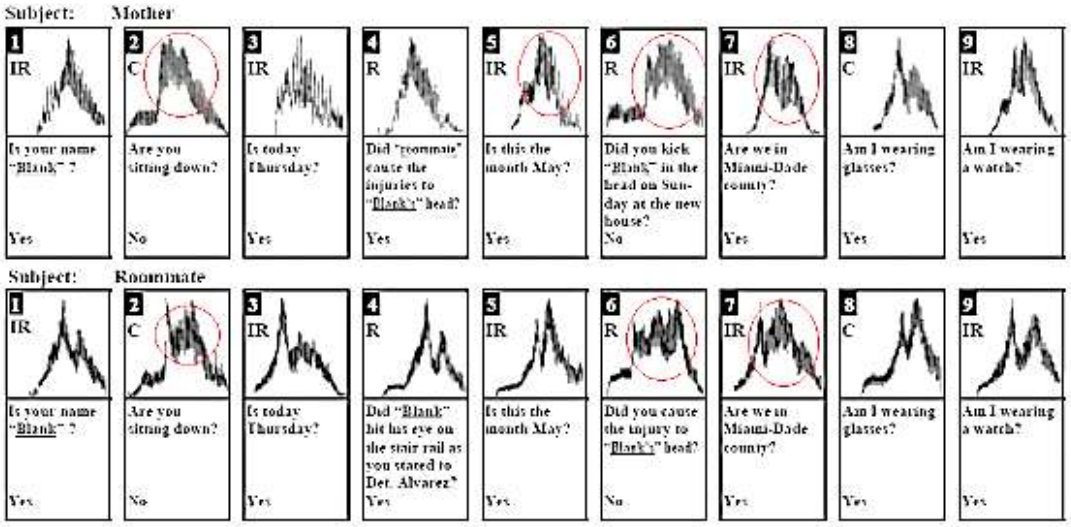

Figura 4. Caso real de detección de engaños resuelto por CVSA
Fuente: (A real case solved by CVSA , 2001)

2.3 Procesamiento de la señal de la voz y extracción de características

La detección de estrés se realiza observando las variaciones en el conjunto de características propias del habla neutral, en comparación con el habla producida bajo estrés. El número de características que se usan en un estudio deben ser bien analizadas porque un número grande de las mismas puede contener mucha redundancia y alto costo computacional, por el contrario una pequeña cantidad puede no tener suficiente información para caracterizar el engaño (Rodellar, Palacios, Nieto, & Gomez , 2015).

La información es acústica cuando la extracción se realiza únicamente de la señal de la voz, y se agrupan en (Erro & Tonantzintla, 2010):

- **Espectrales** son las que describen las propiedades de una señal en el dominio de la frecuencia por medio de harmónicos y formantes.
- **De calidad de Voz** estas definen los diferentes estilos de hablar como son: neutral, susurrante, jadeante, sonoro, ruidoso, estrepitoso, resonante.

- **Prosódicas** estas describen fenómenos suprasegmetales como: entonación**,** volumen, velocidad, duración, pausas y ritmo.

Otro tipo de información que se puede obtener es lingüística, esta proviene del texto transcrito de la señal de voz obtenida, se debe tener en cuenta que en una aplicación en tiempo real para obtener información lingüística es teniendo un sistema de reconocimiento automático del habla. A este tipo de información se le asignan dos enfoques diferentes (Erro & Tonantzintla, 2010):

- **Bolsa de Palabras:** en este tipo de enfoque (Liscombe, Riccardi, & Hakkani-Tur, 2005) (Polzehl, 2009) la información en texto se representa a través de vectores lingüísticos. Estos últimos aumentan en dimensión con cada palabra agregada, representando la frecuencia dentro de la elocución.
- **Palabras Clave:** en este tipo de enfoque (Lee, 2009) (Wollmer , 2009), para mejorar la clasificación la estrategia empleadas es detectar palabras clave, y para determinar esta se suele utilizar el concepto de "palabra relevante" y esta es la palabra que aparece con más frecuencia.

Para el procesamiento de características existen dos enfoques el Modelado Dinámico y Estático de características (Erro & Tonantzintla, 2010).

- **Modelo dinámico:** En este tipo de modelado se emplean características como tono, energía, MFCCs y sus derivativas, entre otras. Con modelos de clasificación dinámicos como Modelos Ocultos de Markov (*Hidden Markov Models*) (Pittermann & Pittermann, 2006) o Modelos Mixtos Gaussianos (*Gaussian Mixture Models*). El análisis se hace por medio de

ventanas del mismo tamaño por lo tanto para elocución se tienen diferentes vectores de características de diferentes tamaños dependiendo de su duración (Vlasenko, 2007).

- **Modelo estático:** Con este tipo de modelado se clasifica usando métodos estáticos como *Support Vector Machine* o redes neuronales. La clasificación se la realiza a nivel de elocución completa por lo tanto los segmentos de análisis son de diferentes tamaños. Las características son obtenidas de la extracción de LLDs (*Low Level Descriptors*), por ejemplo entonación, energía o coeficientes espectrales, y características estadísticas; desviación estándar, media, mediana; por lo que el resultado son vectores de características del mismo tamaño para todas las elocuciones (Vogt & André, 2009) (Lee, 2009) (Planet, 2009).

2.3.1 Transformadas Tiempo-Frecuencia

Las representaciones tiempo-frecuencia (TFR), se ha comprobado que son muy útiles ya que mejoran los resultados de los métodos espectrales y temporales clásicos, puesto que son capaces de mostrar cambios en frecuencia con respecto al tiempo. Los métodos que se basan en características temporales no son capaces de detectar con certeza las características esenciales de la señal. Es por eso que un uso combinado de los dos dominios permite un aumento del aprovechamiento de la señal, obteniendo características útiles (Duque Sanchez & Morales Perez, 2007).

Dado que el tiempo es una variable intrínseca en todo tipo de información y señal es importante el tratamiento temporal de la señal y conocer sus principales características (forma, amplitud, pendientes, cruces por cero, energía, entre otras.) ya que con esto podemos extraer la mayor cantidad de información útil para el estudio (Duque Sanchez & Morales Perez, 2007).

2.3.1.1 Transformada discreta de Fourier (DFT) y Estimación Espectral

Esta transformada de tiempo – frecuencia permite calcular las características espectrales para el procesamiento de los datos de voz y para ello se puede aplicar la ecuación 1 (Beigi, 2018).

$$H_k = \sum_{n=0}^{N-1} \tilde{h}_n \, e^{-i\frac{2\pi kn}{N}} = \sum_{n=0}^{N-1} h_n w(n) \, e^{-i\frac{2\pi kn}{N}} \qquad (1)$$

Donde, $k = \{0, 1, \dots, N-1\}$ es el índice del dominio de la frecuencia.

2.3.1.2 Transformada Gabor

Tipo de transformada utilizada para el procesamiento de señales basadas en empleo de ventanas temporales, es una clase de representaciones tiempo-frecuencia. Esto es de funciones bien localizadas en un intervalo. La ventana denominando $g(t)$, contiene una cierta porción de la señal y permite aplicar localmente la Transformada de Fourier a esa porción de señal. De esta manera se puede revelar la información de frecuencia localizada temporalmente en el dominio efectivo de la ventana. Para obtener la mayor y completa información tiempo-frecuencia se desplaza temporalmente la ventana y se cubre el dominio de la señal (Duque Sanchez & Morales Perez, 2007). Y se representa de la siguiente manera (ver ecuación 2).

$$\widehat{s_g}(\tau, \omega) = \int_{-\infty}^{\infty} s(t) g(t-\tau) e^{-j\omega t} \, dt \qquad (2)$$

Con esta ecuación obtenemos un completo mapa en el dominio tiempo-frecuencia que despliega la información de la señal, y esta puede recuperarse con la fórmula de inversión (ecuación 3) (Duque Sanchez & Morales Perez, 2007).

$$s(t) = \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \widehat{s_g}(\tau, \omega) e^{j\omega t} \qquad (3)$$

Considerando un par de ventanas estas actúan como filtros pasabanda, con definición de fase. De modo que la transformada de Gabor puede entenderse como un tratamiento localizado de la señal mediante filtros pasabanda deslizante, de ancho de banda constante (Serrano, 2000).

La transformada de Gabor en aplicaciones de análisis de señales acústicas es muy eficiente en la localización y caracterización de eventos con patrones de frecuencia bien definida, no superpuestos y relativamente largos, respecto a la ventana de análisis. Sin embargo es inapropiada para detectar detalles de corta duración (Serrano, 2000).

2.3.1.3 Transformada Wavelet

Utilizar ventanas moduladas es una de las alternativas de la transformada de Gabor, ventanas de dimensión variable, ajustada a la frecuencia de oscilación. Es decir que mantenga un mismo número de oscilaciones en el dominio de la ventana. Esto es tener una única ventana modulada y generar una completa familia de funciones elementales mediante sus dilataciones o contracciones y traslaciones en el tiempo, a esto se le denomina transformada de Wavelet Continua (ecuación 4) (Mallat, 1998) (Duque Sanchez & Morales Perez, 2007).

$$W_\Psi s(a, b) = \int_{-\infty}^{\infty} s(t) \Psi_{a,b}(t) dt \qquad (4)$$

Donde $a \neq 0 \; y \; b$ son parámetros de escala y de translación. La transformación así definida mantiene la energía de la señal, y la forma de inversión es:

$$s(t) = C_{\Psi} \int_{0}^{\infty} \int_{-\infty}^{\infty} W_{\Psi} s(a,b) \Psi_{a,b}(t) \frac{dbda}{a^2} \qquad (5)$$

La fórmula expresa la síntesis de la señal como la superposición integral de las funciones elementales. El mapeo sobre domino tiempo-frecuencia, parametrizado por (a, b), esto es la Transformada Wavelet Continua, representa una novedosa alternativa a la Transformada de Fourier por ventanas (Duque Sanchez & Morales Perez, 2007).

Las transformadas de Wavelet Continua y discreta son muy optimas en el procesamiento de señales y particularmente en las señales de emisión acústica, solucionando la problemática tiempo – frecuencia que estas presentan. La Continua facilita el mapeo gráfico de la información de la señal, donde se puede visualizar las estructuras, patrones y fenómenos (Serrano, 2000).

Una de las Wavelet más usadas es la de Daubechies, las cuales son un conjunto de señales de base ortogonal y de fácil implementación en el filtrado digital. Son una amplia familia de funciones ortogonales es decir si dos funciones $f_1(x)$ y $f_2(x)$, su integral en el límite $a \leq x \leq b$, del producto de ambas funciones es igual a cero. Y la principal característica de las Wavelet Daubechies es que se adaptan a las señales o imágenes que posean cierta suavidad (figura) (Hernández, 2018).

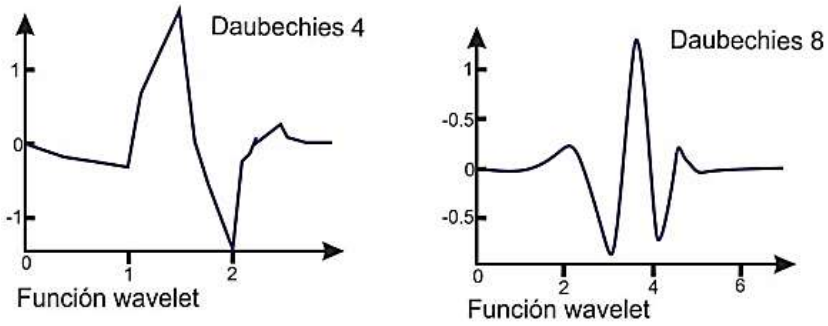

Figura 5. Función wavelet de Daubechies 4 y 8.
Fuente: (Hernández, 2018)

2.3.2 Extracción de Características

La caracterización de la señal consiste en la obtención de parámetros, que conforme a su relevancia permitan de forma completa o parcial la descripción de la misma. El principal objetivo de la caracterización es con una reducción en la dimensionalidad y realce de aspecto de la señal poder obtener información que permita ser analizada y con esto realizar estudios de reconocimiento, segmentación o clasificación (Duque Sanchez & Morales Perez, 2007). Las características pueden ser dividas en varios grupos, según (Duque Sanchez & Morales Perez, 2007) (Vargas, 2003) (Mesa & Morales , 2004) se dividen en dos tipos de características: las acústicas, aquellas que tienen un sentido físico, y las de representación, son aquellas que corresponden a valores calculados a partir de alguna forma de representación de la voz.

Y según (Erro & Tonantzintla, 2010), las características de la voz pueden ser divididas en tres grupos principales: Prosódicas, Espectrales y de Calidad de la voz; las características prosódicas se subdividen en tiempos de elocución, contorno melódico y energía. Estas tienen que ver con la prosodia como duración, pitch, energía, entre otros. Es apropiado decir que se debe considerar un número adecuado de características para que no exista redundancia, ni falta de información.

Características Acústicas: como ya se mencionó las características acústicas son aquellas que poseen un significado físico**,** por lo que permiten una calificación de cualidades vocales. Estas características se clasifican en:

- Parámetros cuasiperiódicos, estos son los que presentan las formas de periocidad presentes en la señal de la voz: frecuencia fundamental
- Parámetros de perturbación, estos reflejan una variación relativa de cierto parámetro: *Jitter, Shimer, HNR.*

Características de Representación: estas características describen el comportamiento dinámico de señales, estas son calculadas a partir de algún método de representación de la señal y que generalmente no se le asocia a algún sentido físico (Duque Sanchez & Morales Perez, 2007).

Características Prosódicas: La prosodia es una fuente de información de la señal que contiene características muy importantes del habla ya que complementa el mensaje lingüístico con una intensión determinada, esta contiene actitudes o estado emocional del hablante. De este tipo de características también se puede extraer funciones extralingüísticas del hablante como su edad, sexo, entre otras (Erro & Tonantzintla, 2010).

Características de Calidad de la Voz: estas definen los diferentes estilos de hablar como son: neutral, susurrante, jadeante, sonoro, ruidoso, estrepitoso, resonante. Y estas son: *Jitter*, *Shimmer*, *Noise to harmonics ratio mean*, entre otras (Erro & Tonantzintla, 2010).

Características Espectrales: son las que describen las propiedades de una señal en el dominio de la frecuencia por medio de harmónicos y formantes. Y estas son: Varianza, media, mediana, mínimo, máximo, *kurtosis, centroid, Skewness*, entre otras (Erro & Tonantzintla, 2010).

2.3.3 Detección de actividad de voz (VDA)

La detección de actividad de la voz o VDA por sus siglas en inglés (*Voice Activity Detection*), es importante en la caracterización de la señal de voz. Normalmente cuando hay silencio, la relativa proporción de ruido puede ser pequeña y consecuentemente todos los tipos de resultados pueden ocurrir (Beigi, 2018).

Estos detectores son implementados en diferentes áreas de procesamiento de voz, como comunicación móvil (Evangeopulos & Maragos, 2006), transmisión de voz en tiempo real

(Sangwan & Chiranth, 2002), reducción de ruido en dispositivos de audición (Nasibov, 2012). Y el objetivo común de VAD es detectar las zonas de voz o silencio de una señal, basando en las características de la señal y ciertas reglas de decisión basadas en análisis estadístico (Rodriguez, Castañeda , & Ballesteros , 2015).

El proceso de programación que el VAD sigue se describe en la Figura 6, en la cual se presenta cuatro entradas principales, que corresponden a las muestras de la trama tomada, y sobre el bloque de detección se realiza la comparación de los valores de retardo, y condiciones sobre el umbral que detecta el tono mediante el uso de banderas que indican la presencia del mismo.

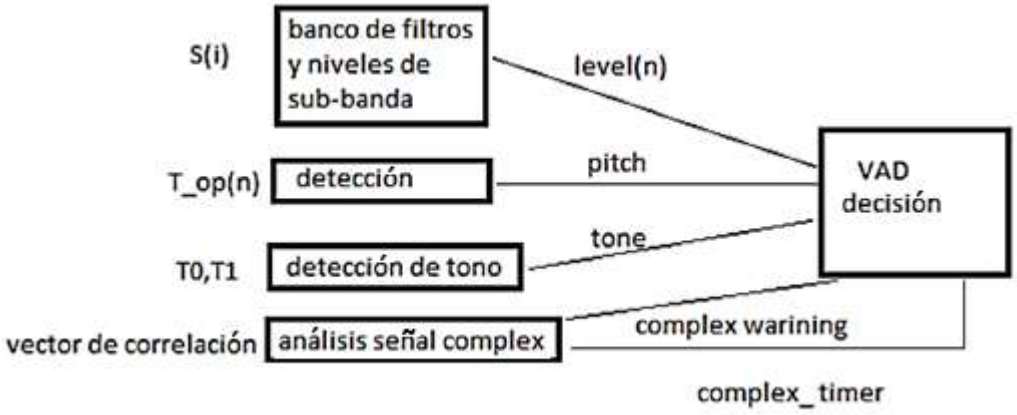

Figura 6. Diagrama de bloques algoritmo VDA.
Fuente: (Chen, 2010)

2.4 *Machine Learning*

El desarrollo de las nuevas tecnologías es imparable en los diferentes tipos de procesos, con mejora de costes operativos de alta complejidad técnica y que manejen gran volumen de datos. Por lo que el aprendizaje automático o *machine learning*, generalmente se refiere a los cambios en los sistemas que realizan tareas asociadas con la inteligencia artificial. Dichas tareas incluyen reconocimiento, diagnóstico, planificación, control, predicción, etc. Por lo tanto *machine learning*

permite el manejo de datos y permite aprovechar al máximo los mismos con la premisa de reducción de tiempo y de costes operativos (Nilsson, 2005) (Carreño, 2017).

Machine learning permite encontrar patrones en los datos de forma relativamente sencilla y además puede adaptarse a nuevos patrones según las necesidades, y puede ser aplicado a cualquier conjunto de datos (Nilsson, 2005).

Para el correcto entendimiento se va a dar una introducción sobre los paradigmas de aprendizaje supervisado y no supervisado.

2.4.1 Aprendizaje no supervisado

En el aprendizaje automático, el aprendizaje no supervisado o *clustering* se basa en el agrupamiento de instancias que tienen cierta similitud entre sí. El objetivo de este proceso es extraer información relevante o conocimiento de los datos. Sin embargo, el concepto de parentesco tiene connotaciones diferentes y es por ello que existen diferentes técnicas de *clustering* (Estivill-Castro, 2002). El aprendizaje no supervisado es utilizado en áreas como la minería de datos, el aprendizaje automático, reconocimiento de patrones, biomedicina entre otros. Es por ello que en los últimos años esta técnica de aprendizaje tiene grandes aportaciones (Carreño, 2017).

El concepto de cluster trata de maximizar la distancia entre los diferentes grupos de parentesco y, a su vez, minimizar la distancia entre los elementos del mismo cluster, la función de clasificación realiza la asignación de una etiqueta a una instancia (Carreño, 2017).

2.4.2 Aprendizaje Supervisado

Los sistemas de aprendizaje automático supervisado parten de conocimientos previos (conjunto de entrenamiento), para por medio de esto poder predecir o asignar una etiqueta a casos nuevos. Es

decir, el sistema de clasificación aprenderá de experiencias previas para después poder etiquetar la nueva información a analizar. Es por ello que en un sistema de clasificación se pueden diferenciar dos fases. La primera fase se conoce como de aprendizaje, y esta etapa se encarga de aprender los patrones o características de cada instancia. La segunda fase es conocida como test, esta es donde el clasificador es puesto a prueba haciendo que clasifique otro conjunto de instancias de las cuales ya se conoce la etiqueta correcta. De esta manera se podrá conocer la calidad o el comportamiento del sistema.

En la figura 7, se presenta un cuadro de resumen de los diferentes paradigmas de *Machine Learning.*

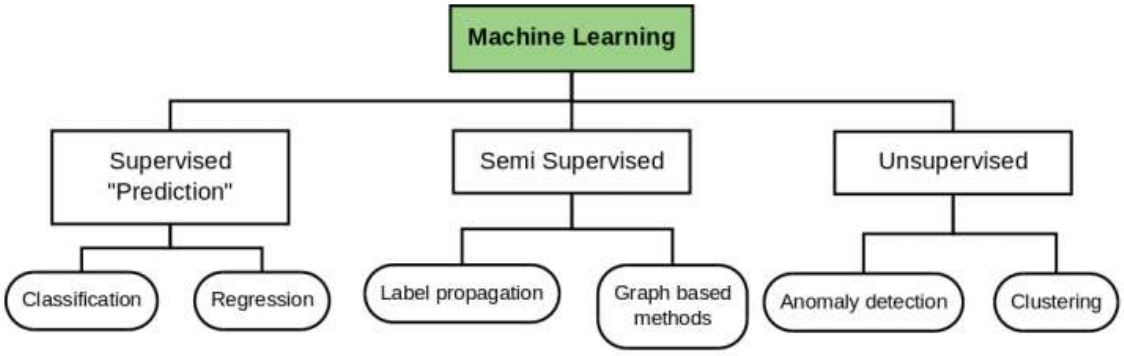

Figura 7. Esquema resumen de los diferentes paradigmas de Machine Learning.
Fuente: (Carreño, 2017)

2.4.1 Máquinas de Soporte Vectorial (*Support Vector Machine (SVM)*)

El paradigma de vector de máquina de soporte (SVM) para el aprendizaje de predictores lineales en espacios de características de alta dimensión. La alta dimensionalidad del espacio de características plantea tanto la complejidad de la muestra como los desafíos de la complejidad computacional. El paradigma algorítmico de SVM aborda el desafío de la complejidad de la muestra buscando separadores de "margen grande". En términos generales, un medio espacio

separa un conjunto de entrenamiento con un gran margen si todos los ejemplos no solo están en el lado correcto del plano de separación, sino que también están muy lejos de él. La restricción del algoritmo para generar un separador de margen grande puede producir una pequeña complejidad de muestra, incluso si la dimensionalidad del espacio de la característica es alta (Sahi & Shai, 2014).

El objetivo del algoritmo de *Support Vector Machine*, es encontrar un hiperplano en un espacio N-dimensional que clasifique claramente los puntos de datos. Como se muestra en la Figura 8 (Gandhi, 2018).

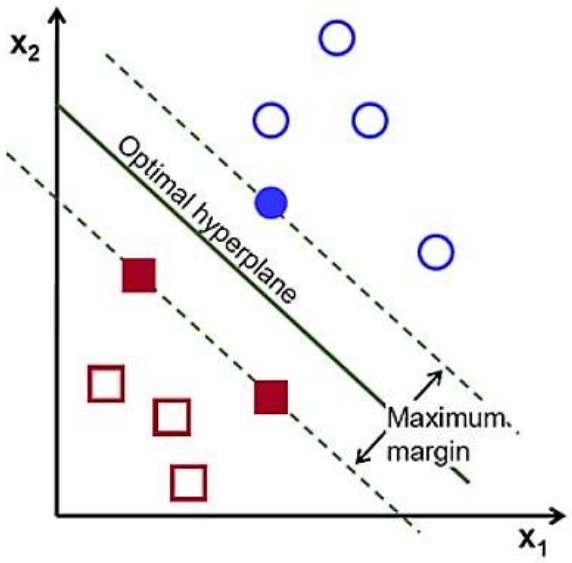

Figura 8. Posible Hiperplano
Fuente: (Gandhi, 2018)

Para separar las dos clases de puntos de datos, existen varios posibles hiperplanos, lo principal es encontrar un plano que tenga el margen máximo, es decir, la distancia máxima entre los puntos de datos de ambas clases. Maximizar la distancia del margen proporciona cierto refuerzo para que los puntos de los datos nuevos a analizarse puedan clasificarse con mayor exactitud (Gandhi, 2018).

2.4.2 *K-Nearest Neighbors (*KNN*)*

Es un algoritmo de aprendizaje automático supervisado, es decir que se basa en los datos de entrada etiquetados para realizar nuevas predicciones en función de lo aprendido. El paradigma se fundamenta en clasificar un nuevo caso, y etiquetarlo según la clase más frecuente a la que pertenecen sus vecinos más cercanos, esto hace que el clasificador sea simple e intuitivo. El algoritmo KNN asume que existen cosas similares en la proximidad (García & Gómez, 2012).

En la Figura 9, se observa el método de clasificación de KNN, el cual los puntos similares están agrupados por secciones o vecinos más cercanos o similitud (García & Gómez, 2012).

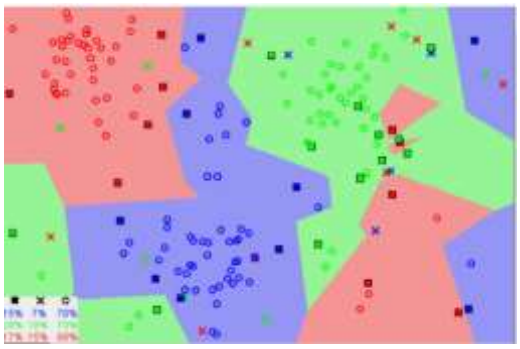

Figura 9. Método de Clasificación de KNN
Fuente: (García & Gómez, 2012)

2.4.3 Árboles de decisión

Es un tipo de algoritmo muy utilizado en Inteligencia Artificial (IA), su estructura se basa en nodos (*nodes*), ramas (*branches*) y hojas (*leaves*) en la Figura 10 se muestra la estructura básica del algoritmo. El primer nodo es denominado nodo raíz, correspondiéndose cada nodo a una única característica (*feature*) y cada rama representa a un rango de valores de dicha característica para dividir los datos. Los nodos finales son denominados hojas, a partir de los cuales termina la división de los datos, obteniéndose así un histograma de los datos existentes correspondientes a una etiqueta y el cual será utilizado como resultado para una predicción futura (Aguirre, 2017).

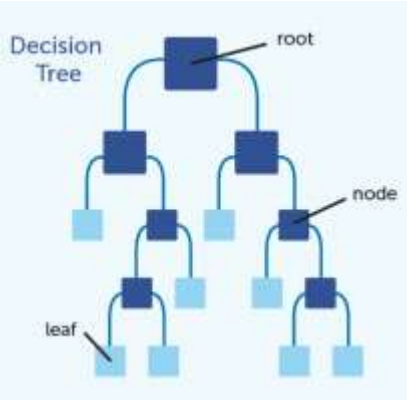

Figura 10. Estructura básica del árbol de decisión.
Fuente: (Aguirre, 2017)

La predicción de una nueva clase se realiza clasificando de acuerdo con la que tiene mayor presencia en las hojas

CAPÍTULO III

METODOLOGIA DEL PROYECTO DE INVESTIGACIÓN

3. Metodología del proyecto de investigación

3.1 Descripción general del proyecto de investigación

Con el propósito de identificar el engaño por medio del habla de una persona, se utilizó el software Matlab®, para el análisis de la señal, extracción de características, grabación de la señal del voz y clasificador automático (*machine learning*) (Figura 11).

La extracción de características se las realizó en dos instancias, la señal de audio sin procesamiento y con transformada de Wavelet, realizando un análisis en cada una de ellas y descartando las características que no son relevantes en cada análisis con el propósito de reducir el procesamiento computacional. Una vez realizado la extracción de características en cada una de las transformadas se utilizó el clasificador que posee el software antes mencionado y con este realizar la predicción.

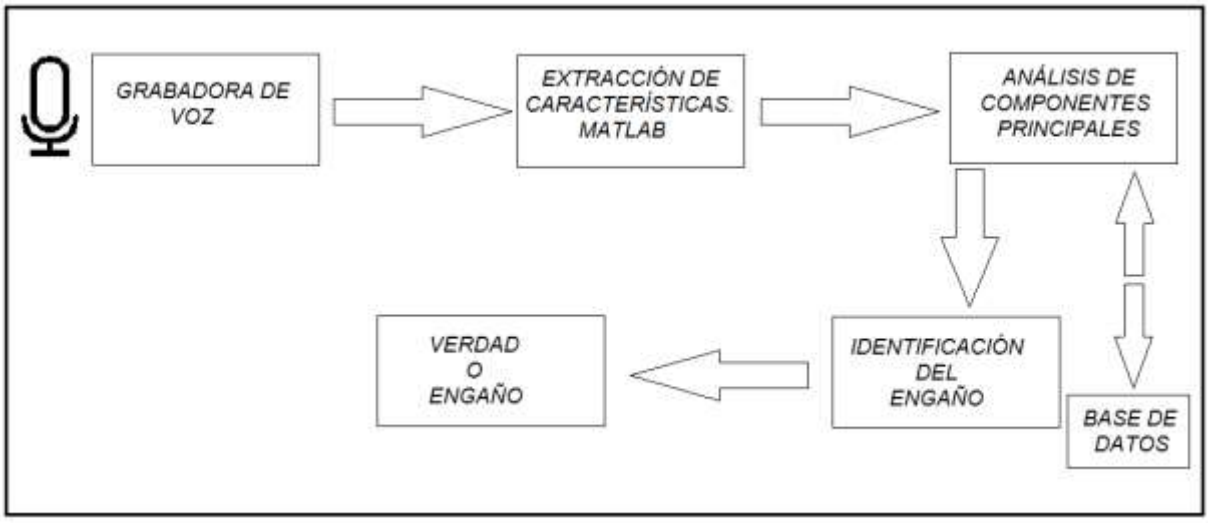

Figura 11. Diagrama de bloques del método propuesto

3.2 Extracción de Características

En esta sección se presenta el desarrollo del algoritmo para la identificación del engaño por medio del reconocimiento automático, para lo cual primero se extraen las características relevantes de la señal de la voz, de una base de datos etiquetada. Estas muestras servirán para realizar el procesamiento y entrenamiento del sistema de detección automático, más adelante se explica el procedimiento para la extracción de cada característica.

Se utilizó como herramienta computacional Matlab®, ya que es un software matemático que ofrece el procesamiento y análisis de la señal de voz por medio de un lenguaje de programación.

Para ello se realizó la normalización de la señal de audio de entrada, y se dividió en ventanas (cuadros) no superpuestos a corto plazo de 20 milisegundos, para cada cuadro se calculan sesenta y ocho características, las cuales posteriormente serán analizadas para seleccionar las más relevantes para el estudio. A continuación se describen las características consideradas como principales.

3.2.1 Frecuencia Fundamental (*Pitch*)

La frecuencia fundamental corresponde a la frecuencia de señal de excitación proveniente de la glotis, siendo caracterizada por el número de vibraciones de las cuerdas bucales por segundo. Esta frecuencia se denomina primer harmónico y varia alrededor de 113 [Hz] para los hombres y 220 [Hz] para las mujeres. Se sabe que la frecuencia fundamental se da por una interacción entre la compresión, masa y tensión de las cuerdas vocales (Cavalcanti de Almeida, 2010).

La presencia de diferentes patologías en la voz puede ser analizada con la frecuencia fundamental, ya que estas pueden hacer variar considerablemente este valor.

Para la extracción de la frecuencia fundamental se utilizó el método de Cepstro, un análisis cepstral de una señal de voz permite trabajar con una señal de la glotis y del tracto vocal por separado, lo que facilita el análisis y estudio de los cambios en las cuerdas vocales (Teixeira , Barbosa, & Moreira, 2011). El método de Cepstro es una operación matemática que consiste en extraer la transformada de Fourier del espectro de la señal. El significado físico de esta transformación se puede interpretar como la información de los cambios de ritmo en las diferentes bandas del espectro de una señal cualquiera.

El análisis de una señal acústica, a través del análisis cepstral en la detección de modificaciones en la señal relacionadas con alteraciones laríngeas, las cuales pueden variar al estar bajo tensión, permite obtener una herramienta no invasiva (Teixeira , Barbosa, & Moreira, 2011).

3.2.2 Micro temblores (*Jitter*)

El *jitter* se da en un ciclo de vibraciones de las cuerdas vocales, de variabilidad involuntaria de la frecuencia fundamental, con lo cual se puede determinar el grado de estabilidad del sistema fonador (Titze, 1994).

Las razones por las que estas pueden varias son: neurológicas (signos neurológicos descoordinados que causan variaciones en la contracción muscular), biomecánicas (pulsaciones debido a la circulación sanguínea en los capilares del tejido de las cuerdas vocales), aerodinámicas (inestabilidad en el flujo del aire) y acústicas (por la interacción entre la glotis y el tracto vocal). Al analizar este parámetro es aceptable la presencia de un pequeño grado de perturbación e irregularidad en la señal de la voz, pero si este ya presenta cambios significativos puede presentarse una irregularidad (Guimaraes, 2007).

Existen varios tipos de medidas de *Jitter*, relativamente estos valores se pueden encontrar de la siguiente manera (Boersma & Weenink, 2010):

***Jitter* (local):** Representa la diferencia media absoluta entre dos periodos consecutivos, dividido por el periodo medio, ecuación 6.

$$jitt = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \qquad (6)$$

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

***Jitter* (absoluto):** Representa la diferencia media absoluta entre dos periodos consecutivos y a este parámetro se le denomina *Jitta*, ecuación 7.

$$jitta = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}| \qquad (7)$$

El principal interés de la medida del *jitter* de una señal de voz, es el diagnóstico diferencial entre la voz normal y la voz bajo estrés, ya que está científicamente comprobado diferencias estadísticas significativas en estos valores en los dos casos antes mencionados. En cuanto a los umbrales para poder diferenciar, en la literatura no existe mucho acuerdo, debido a la variedad de aplicaciones usadas, a las diferentes formas de calcular este valor, y a las diferentes bases de datos que se utilizan para realizar estos estudios, sin embargo estos umbrales pueden estar comprendidos entre 1% y 3% (dos Santos Lopes, 2008).

3.2.3 *Shimmer*

El *Shimmer* al igual que el *Jitter* se mide en un ciclo de vibraciones de las cuerdas vocales, es de variabilidad no voluntaria de la amplitud de cada ciclo, cuantificando los cambios mínimos de la amplitud de la señal, con base en cada ciclo fonatorio. El *Shimmer* puede ser medido en dB, como una evaluación de la variación logarítmica entre la amplitud en ciclos consecutivos, a través de la ecuación 8.

$$Shimmer(dB) = \frac{20 \times \sum_{i=0}^{i=N-1} |log_{10} \frac{A_i}{A_{i+1}}|}{N-1} \qquad (8)$$

Relativamente el *Shimmer* se puede calcular de las siguientes formas (Boersma & Weenink, 2010):

***Shimmer* (local):** Representa la diferencia media absoluta entre las amplitudes de dos periodos consecutivos, dividido para la amplitud media, ecuación 9.

$$Shimm = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \qquad (9)$$

N corresponde al número de ciclos evaluados, *i* corresponde al índice de ciclo, A_i corresponde a la amplitud de muestra de cada ciclo.

En cuanto a los valores umbrales del *Shimmer* al igual que el *Jitter* estos valores no están bien consensuados en la literatura, pero sin embargo estos pueden estar comprendidos entre 3% y 5% (dos Santos Lopes, 2008).

3.2.4 Energía

En el procesamiento digital de señales de voz, la intensidad sonora puede ser obtenida por medio del cálculo de la energía de la señal, para lo cual se utilizan técnicas en el dominio del tiempo (análisis temporal) o en el dominio de la frecuencia (análisis espectral) (Cavalcanti de Almeida, 2010).

Las propiedades estadísticas de las señales de voz se pueden considerar como invariantes en el tiempo, para intervalos cortos, hasta 32 ms, siendo un valor típico, 16 ms. Con esto se busca obtener los parámetros temporales de la señal a partir de los segmentos que estén situados en el intervalo de interés.

La energía segmental está definida como:

$$E_{seg} = N_A \cdot E\{[s(n) - \mu_s(n)]\text{^}2\} \qquad (10)$$

Donde N_A es el número de muestras del segmento a analizar, $s(n)$ es la señal de voz y $\mu_s(n)$ es la media de la señal de voz.

3.2.5 Entropía de la Energía

La entropía de una señal es una medida de la cantidad de información que tiene una señal ya sea este audio o imagen. La entropía de la energía a corto plazo se puede interpretar como una medida de cambios bruscos en el nivel de energía de una señal de audio (Jithin, Nandan, Eldho, Akhil, & Sreehari, 2017) (Djebbar & Ayad, 2017).

La señal de audio se denota como $x(t)$, donde $(t = 0, 1, 2, 3 \dots, N-1)$ y la entropía de la señal se denota $H(x)$, Se emplea la entropía de Shannon (Shannon, 1948) (ecuación 11). Las

pequeñas perturbaciones en los valores de la muestra de la señal producen perturbaciones más pequeñas en la entropía medida, la entropía se define de la siguiente manera:

$$H(x) = -\sum_{i} p(x_i) \cdot \ln(p(x_i)) \qquad (11)$$

La entropía es la suma de dos variables aleatorias discretas independientes.

3.2.6 Taza de Cruce por cero

La tasa de cruce por cero de un cuadro de audio es la velocidad de los cambios de signo de la señal durante la ventana. Es decir, es el número de veces que la señal cambia de valor, de positivo a negativo y viceversa. La tasa de cruce por cero (ZCR) se define de acuerdo a la ecuación 12:

$$Z(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} |sign(x_i(n)) - sign(x_i(n-1))| \qquad (12)$$

Donde W_L es la secuencia de muestras de audio y $sign(\,)$ representa la función signo con valor 1 para valores positivos y -1 para valores negativos.

3.2.7 Roll-off espectral

Es un tipo de característica que mide la frecuencia por debajo de la cual se concentra la mayor parte de la energía de espectro. El roll-off considera el espectro como una distribución de probabilidad, siendo este el valor para el cual se supera una determinada probabilidad en la distribución de probabilidad acumulada, esta característica se define como la ecuación 13 (Sobreira & Rodríguez, 2008).

$$SR = max_m\{\sum_{k=0}^{m}|X_t(k)| \leq TH \cdot \sum_{k=0}^{N-1}|X_t(k)|\} \quad (13)$$

Donde el umbral TH toma valores entre 0.85 y 1.

3.2.8 Centroide Espectral

El centroide espectral es el centro de gravedad de la distribución de la densidad espectral de potencia, tiene relación con el *Shimmer* ya que las dos se relacionan con la brillantez de la señal de audio (mayor contenido en altas frecuencias que en medias y graves). Cuanto mayor es el valor del centroide espectral, más desplazada se encuentra la energía hacia las frecuencias altas. El centroide espectral está definida por la ecuación 14 (Sobreira & Rodríguez, 2008).

$$Centroid_t = \frac{\sum_{k=0}^{N-1}|X_t(k)| \cdot k}{\sum_{k=0}^{N-1}|X_t(k)|} \quad (14)$$

3.2.9 Flujo espectral

Esta característica representa los cambios espectrales entre dos cuadros sucesivos y se calcula como la diferencia entre las magnitudes normalizadas de los espectros de las dos ventanas sucesivas de corto plazo (Jithin, Nandan, Eldho, Akhil, & Sreehari, 2017). El flujo espectral se define en la ecuación 15, obteniendo el flujo correspondiente a una trama t, que depende de la amplitud de la trama anterior $a(t-1,k)$.

$$flux(t) = 1 - \frac{\sum_k a(t-1,k)\, a(t,k)}{\sqrt{\sum_k a(t-1,k)^2} \cdot \sqrt{\sum_k a(t,k)^2}} \quad (15)$$

Este parámetro toma valores entre 0 y 1, siendo casi nulo cuando existe gran similitud entre espectros, y viceversa (Aguirre, 2017).

Las características que se describen a continuación se obtuvieron con librerías propias de la herramienta utilizada (Matlab®), el cual posee programas para obtener estos valores ingresando parámetros particulares de la señal.

3.2.10 *Skewness* (Sesgo)

El *Skewness* da una medida de la asimetría de una distribución en torno a su valor medio (Peeters, 2004). Esta asimetría se calcula según la expresión de la ecuación 16.

$$\begin{cases} m_3 = \int (x-\mu)^3 \cdot p(x)\delta(x) \\ \gamma_1 = \dfrac{m_3}{\sigma^3} \qquad (skewness) \end{cases} \qquad \textbf{(16)}$$

En función al valor de *Skewness*, el espectro será asimétrico o simétrico con mayor energía hacia la izquierda (menor frecuencia) o hacia la derecha (mayor frecuencia) del centroide (Aguirre, 2017):

$$\begin{cases} \gamma_1 = 0 \qquad\qquad\qquad\qquad (Distribución\ simétrica) \\ \gamma_1 < 0 \quad (Distribución\ asimétrica\ hacia\ la\ derecha) \\ \gamma_1 > 0 \ (Distribución\ asimétrica\ hacia\ la\ izquierda) \end{cases}$$

3.2.11 *kurtosis*

La *kurtosis* da una medida del grado de planicie de una distribución y su valor medio, Esta se calcula a partir del momento de cuarto orden como se describe en la expresión de la ecuación 17 (Aguirre, 2017) (Peeters, 2004).

$$\begin{cases} m_4 = \int (x-\mu)^4 \cdot p(x)\delta(x) \\ \gamma_2 = \dfrac{m_4}{\sigma^4} \qquad (Kurtosis) \end{cases} \tag{17}$$

Al igual que el *skewness*, la *Kurtosis* indica lo plano o picudo que es el espectro en función de los valores obtenidos, respecto a una distribución normal ($\gamma_2 = 3$):

$$\begin{cases} \gamma_2 = 3 \qquad (Distribución\ normal) \\ \gamma_2 < 3 \quad (Distribución\ más\ plana) \\ \gamma_2 > 3\ (Distribución\ más\ picuda) \end{cases}$$

3.2.12 Media

Es una medida de tendencia central y resulta al efectuar una serie determinada de operaciones con un conjunto de números. En estadística es conocida también como esperanza matemática, sea X una variable aleatoria con distribución de probabilidad f(x). La media o valor esperado se describe en la ecuación 18 (Pabón Ángel , 2010).

$$\mu = E(x) = \int_{-\infty}^{\infty} x f_x(x) dx \tag{18}$$

3.2.13 Mediana

Es la observación de la mitad después de que se han colocado los datos en una serie ordenada. Si los datos están agrupados, la mediana se define como el valor dentro del intervalo que divide la distribución en dos partes iguales (Pabón Ángel , 2010).

3.2.14 Transformada de Wavelet

La transformada de Wavelet realiza un filtrado de la señal de audio en el dominio del tiempo, mediante filtro pasa bajos y pasa altos, los cuales eliminan componentes de alta o baja frecuencia de la señal (Hernández, 2018).

Para realizar esta transformada, se utilizó el comando de Matlab® "wavedec", el cual ingresando la señal a ser transformada, el número de niveles a ser dividida y el tipo de wavelet a utilizarse, en el caso de este trabajo se utilizó tres niveles de división, y el tipo "db5", que es wavelet Daubechies ya que una de las principales características es que se adaptan bien a las señales (Hernández, 2018). En la Figura 12, se puede visualizar los niveles de descomposición de la señal.

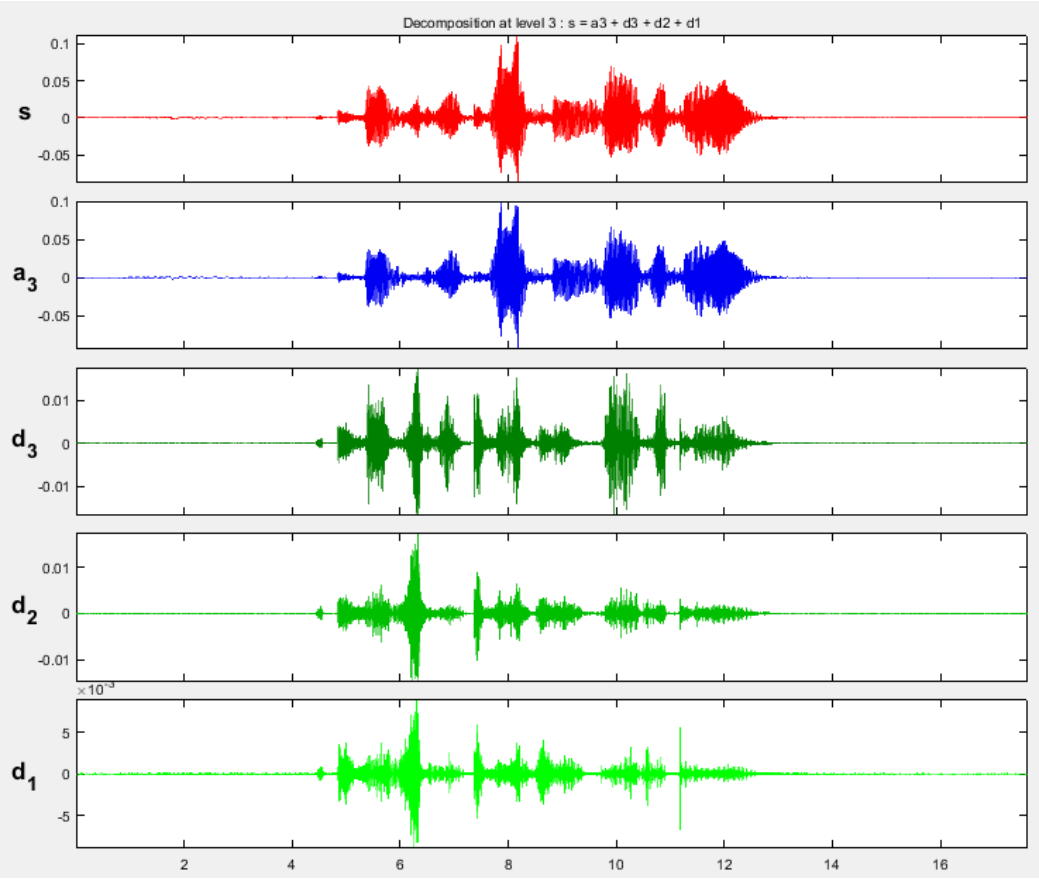

Figura 12. Niveles de descomposición de Wavelet con db5

3.3 Procesamiento de las características extraídas

Una vez realizado la programación para la extracción de cada característica, estas son agrupadas de manera que estas se presenten en un vector, el cual contenga la etiqueta de cada característica y el valor calculado de la señal de voz, esto se realiza para para grupos de señales de audio almacenados en archivos "wav", obteniendo de esta manera la tabla con las características extraídas para el grupo de señales que se vaya a analizar, en el caso de este trabajo, estas tablas están divididas en 64 señales de audio para el aprendizaje del sistema y 32 para la prueba del sistema.

3.4 Base de datos.

Para el presente estudio se utilizó una base de datos, realizada a partir de las diferentes características principales que una persona puede presentar en su voz, al momento de encontrarse bajo estrés, es decir cuando este procesando una mentira o engaño a una pregunta realizada. Por otro lado, las características que se presenta cuando la persona está en un estado neutro o sin estrés. De esta manera se utilizó una base de datos de un total de 96 audios grabados por actores hombres y 96 audios grabados por actrices mujeres, debidamente etiquetada para el estudio a realizarse.

Según (Ekman, 2007), las tres emociones, más frecuentes, que experimentan las personas cuando mienten son el miedo, sorpresa y en otros aspectos ira. El miedo a ser descubierto es la emoción más común en las personas, y este temor puede observarse en el rostro (en una micro expresión o pérdida de color en el rostro) y en la voz (con el miedo existe el nerviosismo, y con esto la persona titubea o la voz se quiebra). La reacción de sorpresa se tiene cuando se realiza un juicio directo hacia la persona y esta para no ser descubierta reacciona sorprendida, de igual manera que el miedo, esta puede observarse en el rostro y en la voz con las mismas características. La ira

se presenta con la necesidad de la persona de ocultar la mentira, y se puede identificar en la voz (voz elevada, combinada con nerviosismo).

Por el contrario una persona que no engaña o miente con respecto a un cuestionamiento, reacciona con un comportamiento normal o calma, esta puede estar acompañada de nerviosismo pero a diferencia de un engaño esta mantiene un timbre de voz normal y no realiza demasiadas pausas en dar su respuesta.

Tomando en cuenta los aspectos antes mencionados la base de datos utilizada es la que se presenta de manera pública, *RAVDESS*, la cual contiene 7356 archivos de audio, etiquetado, dentro de los cuales están emociones, engaños y verdades, entre otras (Livingstone SR, Russo FA, 2018).

La finalidad de la elaboración de la base de datos (RAVDESS), fueron estudios neurocientíficos, psicológicos, psiquiátricos y ciencias de la computación, con el objetivo de detección de trastornos neurológicos (Livingstone SR, Russo FA, 2018). La licencia y permisos de Copyright de la base de datos se encuentra en el ANEXO A.

3.5 Aprendizaje Automático Supervisado

El aprendizaje automático de las máquinas (*Machine Learning*) tiene como objetivo desarrollar algoritmos que permitan aprender de un conjunto de observaciones, de esta manera dando la capacidad al sistema de establecer hipótesis o predicciones de nuevas observaciones, de tal manera que se puedan tomar decisiones automáticamente, dando así al sistema inteligencia artificial (Aguirre, 2017).

El aprendizaje automático es utilizado en gran variedad de ámbitos, en este trabajo en la clasificación de señales de audio. Este tipo de algoritmos tienen dos fases principales: la de entrenamiento y la de evaluación.

La fase de entrenamiento consiste en proporcionar ejemplos al sistema, con los cuales este debe aprender, en este caso se utilizó un aprendizaje automático supervisado, el cual consiste en proporcionar al sistema un conjunto de datos etiquetado. Para esta fase se utilizó la aplicación "*Classification Learner*" de la herramienta Matlab® (Figura 13), la cual permite seleccionar un archivo el cual contenga los predictores (características) y las clases, y de esta manera realizar un modelo de aprendizaje, para con el posteriormente realizar predicciones con nuevas señales de audio. Este proceso se realizó con una base de datos de 96 audios, los cuales 64 se tomaron para el entrenamiento y 32 para la evaluación, representando así el 70 y 30% respectivamente del total de la base de datos, estos porcentajes fueron tomados de referencia de trabajos previos (Martínez Rodríguez, 2018) (Bagnato, 2017).

La fase de evaluación se lleva a cabo una vez que el sistema ha sido entrenado. Esta fase consiste en entregar al sistema un subconjunto de datos, que el sistema debe etiquetar de acuerdo con lo aprendido anteriormente, y teniendo las verdaderas etiquetas se puede verificar los resultados de la evaluación.

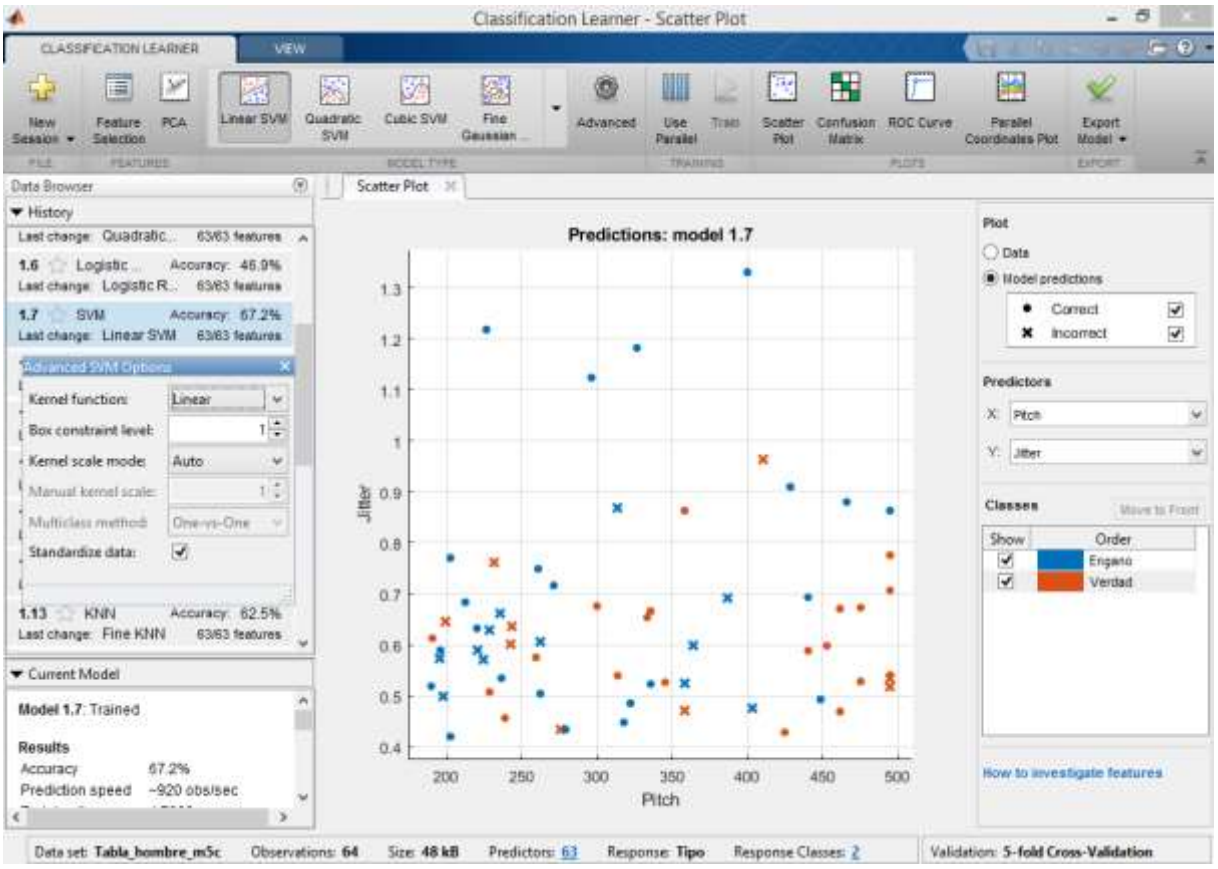

Figura 13. Aplicaccion Classification Learner Matlab®

Mediante la aplicación se puede determinar el mejor clasificador por medio de porcentaje de exactitud. En el caso de SVM, se puede cambiar el orden polinomial de la función núcleo (kernel), estas pueden ser: Lineal, Cuadrática, Cubica y Gaussina. Además que permite variar la Función de Costo (C, Regularización, BoxConstraint en Matlab®), valor que se varía dependiendo de cuan separables son los datos, ya que el algoritmo de entrenamiento debe permitir alguna clasificación errónea en el conjunto de entrenamiento, por lo que se aplica un costo a la clasificación errónea. Este valor debe ser alto cuando el margen de separabilidad de los datos es pequeño, haciendo que la separación sea mucho más estricta (Figura 14) (Hamm, 2019). En el caso de este trabajo el valor se establece en 1, siendo este constante para todos los clasificadores.

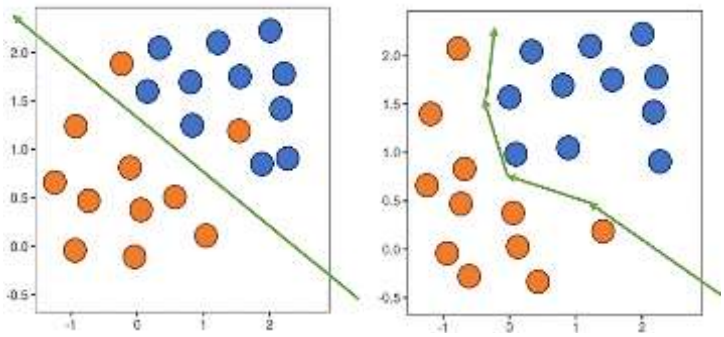

Figura 14. Izquierda: bajo valor Regularización, Derecha: alto valor de regularización
Fuente: (Patel, 2017)

Otro parámetro que puede variarse es el conocido como Gamma (Kernel Scale), y define hasta donde llega la influencia de un solo ejemplo de entrenamiento, con valores bajos significa "lejos" y valores altos, "cerca" (Figura 15) (Patel, 2017).

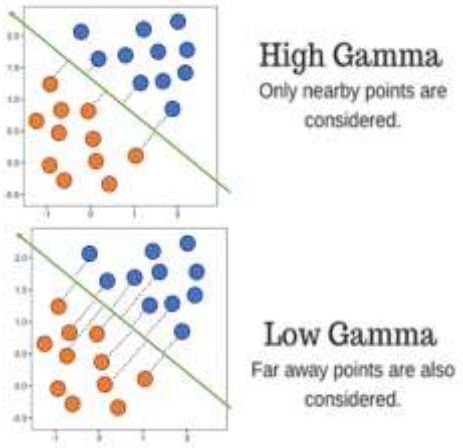

Figura 15. *Valores de Gamma e influencia.*
Fuente: (Patel, 2017)

En el proceso de entrenamiento se obtuvo mejores resultados con el clasificador "*Suport Vector Machine*" (SVM) lineal con C y Gamma igual a 1, siendo este modelo el utilizado e implementado para la detección de engaños y verdades de una señal de audio.

3.6 Selección de Características

El proceso de selección de características (*feature selection*), normalmente va asociado a un conocimiento previo del problema a resolverse, utilizando así características que permitan diferenciar las clases. Sin embargo, es común realizar el proceso extrayendo todas las características posibles para posteriormente identificar cuáles de ellas en conjunto tienen mejores resultados, en el presente trabajo se utilizó el ultimo método antes mencionado, extrayendo todas las características y posteriormente identificando las más relevantes para el estudio. La clasificación y selección de funciones para la clasificación tiene como objetivo reducir la dimensionalidad y el ruido en los conjuntos de datos.

Este procedimiento se realizó con la librería para Matlab® "*Feature Selección Library*" (FSLib), elaborada por Roffo Giorgio, esta selección de características realiza un filtrado de la información, eliminando datos redundantes o no deseados de un flujo de información (Roffo, 2018).

Las técnicas de selección de características se dividen en tres clases: *wrappers,* que usan clasificadores para puntuar un conjunto dado de características; *embedded*, este método integra el proceso de selección en el aprendizaje del clasificador; y *filters*, los cuales identifican las propiedades intrínsecas de los datos, ignorando el clasificador. En el presente trabajo se utiliza el método *wrapper,* ya que es el método establecido para SVM y este sigue el modelo presentado en la Figura 16 (Roffo, 2018).

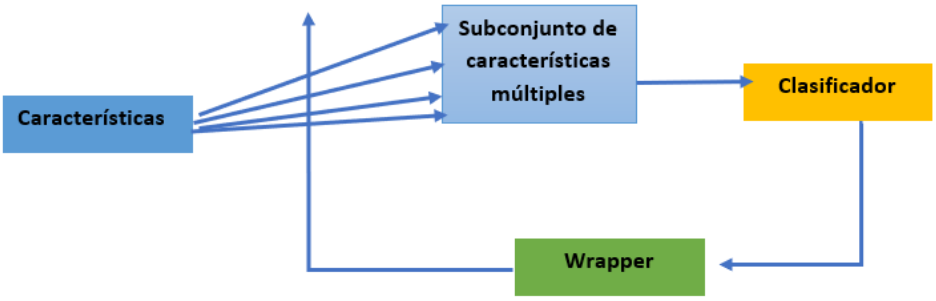

Figura 16. Modelo Wrapper de Selección de Características
Fuente: (Roffo, 2018)

De acuerdo con la librería de *Features Selection* antes mencionada se aplica el código de FSV (*feature Selection via concave minimization*), el cual corresponde para uno de tipo *wrapper* y método supervisado, según los acrónimos asignados por el autor para cada método.

Una vez realizado la selección de características mediante la librería, se realizó pruebas descartando características consideradas menos importantes y se obtuvo una reducción de 68 a 40 características, obteniendo mejores resultados de exactitud en la predicción del sistema, determinando así que dichas características no aportaban al estudio, sino por el contrario entorpecieron el correcto aprendizaje del sistema. Este caso se presentó en los dos géneros analizados, siendo eliminadas las características menos relevantes en cada uno de los casos.

CAPÍTULO IV

ANÁLISIS DE RESULTADOS OBTENIDOS

4. Análisis de Resultados obtenidos

Una vez realizado el proceso de extracción de las 68 características totales (ANEXO B), se selecciona un grupo de la base de datos para realizar el entrenamiento del sistema y el otro para las pruebas, el grupo para el entrenamiento conformado por 64 señales de voz, y el de pruebas por 32 señales. Esto se realizó tanto para el análisis de hombre y mujer, evaluando los resultados por separado.

Los resultados del clasificador son analizados por cuatro parámetros principales que son: Exactitud, Precisión, Sensibilidad y Especificidad, descritos en las ecuaciones 19 – 22 (Enríquez Fustillos, 2017).

$$Exactitud(\%) = \frac{N_E}{N_T} \times 100 \quad \textbf{(19)}$$

$$Presición(\%) = \frac{N_{VP}}{N_{VP} + N_{FP}} \times 100 \quad \textbf{(20)}$$

$$Sensibilidad(\%) = \frac{N_{VP}}{N_{VP} + N_{FN}} \times 100 \quad \textbf{(21)}$$

$$Especificidad(\%) = \frac{N_{VN}}{N_{VN} + N_{FP}} \times 100 \quad \textbf{(22)}$$

Donde, N_E es el número de señales predichas correctamente, N_T es el número total de señales a ser predichas, N_{VP} es el número de verdaderos positivos, N_{FP} números de falsos positivos, N_{VN} número de verdaderos negativos, N_{FN} numero de falsos negativos.

El análisis se realizó por separado el género masculino y femenino debido a que ciertas características para detectar un engaño en un hombre se encontraban en el rango normal de una mujer..

4.1 Análisis de resultados de clasificación de hombres

Teniendo la matriz con los valores de las 68 características de las 64 señales de audio se procede a realizar la clasificación con la aplicación "*Classification Learner*" de Matlab®, la cual de manera muy intuitiva, permite al usuario realizar una clasificación, ya que esta evalúa el entrenamiento con diferentes clasificadores, y optimizando valores propios de cada clasificador, en el caso de KNN, el valor de k (número de vecinos), en SVM (*Support Vector Machine*), los valores de C (Función de Costo) y Gamma y en el caso de árboles de decisión el número de divisiones (ramas). Los resultados de este procedimiento se muestran en la Tabla 2, en la cual se presenta el Clasificador, el tipo y el porcentaje de exactitud de cada clasificador.

Tabla 2

Porcentaje de Exactitud de cada clasificador utilizando la aplicación de "Clasification Learner" (Hombres).

CLASIFICADOR	TIPO	PARÁMETROS DEL CLASIFICADOR		% EXACTITUD
		Núm. Divisiones (ramas)		
Árboles de Decisión	Fine	100		71,90%
	Medium	20		71,90%
	Coarse	4		71,90%
		Número de vecinos (K)		
K-nearest neighbors (KNN)	Fine	1		59,40%
	Medium	10		59,40%
	Coarse	100		46,90%
		Función Costo (C)	**Gamma (γ)**	
Support Vector Machine (SVM)	Lineal	1	-	78%
	Cuadrático	1	1	67,20%
	Cubico	1	1,6	70,30%
	Gaussiano	1	6,3	67,20%

Los datos presentados en la Tabla 2, se representan en la Figura 17, teniendo un mejor análisis del clasificador con mayor porcentaje de exactitud

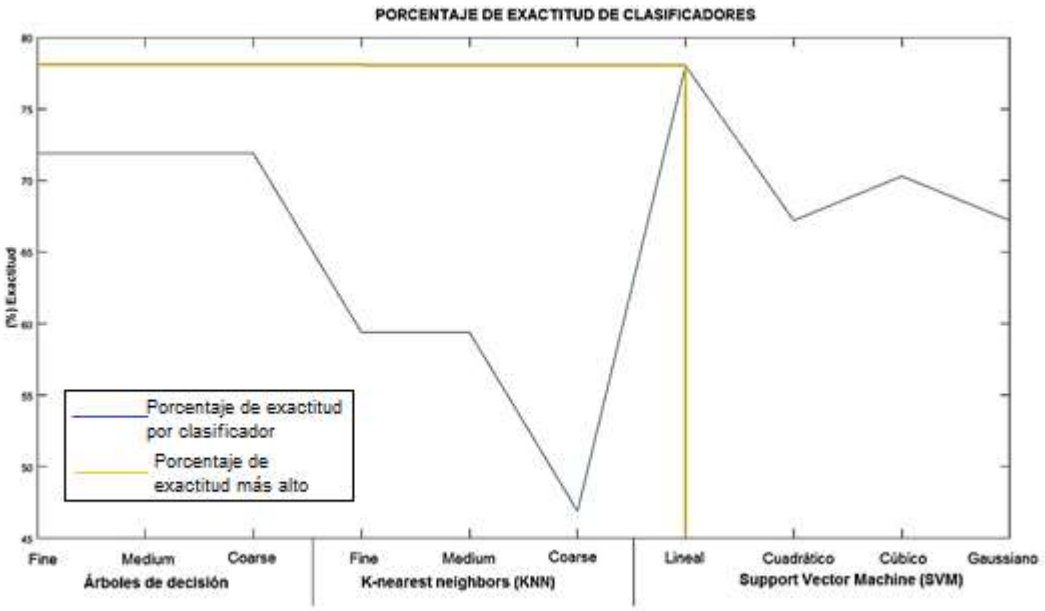

Figura 17. Porcentaje de exactitud modelos de clasificación Hombres.

A partir de la Figura 17 se determina que el clasificador con mayor porcentaje de exactitud es el de *Support Vector Machine (SVM)* tipo lineal, con 78% de exactitud en las predicciones con la tabla de pruebas. Por lo tanto es este clasificador el que se utilizará para el análisis de señales de audio y determinar el engaño o verdad. Como se mencionó anteriormente la aplicación optimiza valores propios de cada clasificador con el fin de obtener el mayor porcentaje de exactitud, en el caso del clasificador a utilizarse, el valor establecido por Matlab®, para SVM lineal es: $C = 1$, en un sistema donde el número de predictores es grande, los modelos son proclives a sufrir *overfitting* (ajuste excesivo o aprendizaje a detalle). Esto hace que entre los diferentes tipos de *kernels*

(núcleos), sea adecuado emplear el de menor flexibilidad, el *kernel* lineal. El único parámetro de un SVM de tipo lineal es el valor de función de costo *C.*

4.1.1 Selección de Características

Con la finalidad de reducir el costo computacional e identificar características que no sean relevantes o a su vez estén entorpeciendo al sistema de clasificación se realizó una selección de características (*feature selection*), utilizando el método correspondiente para el clasificador SVM, FSV (*feature selection via concave minimization*) de tipo "*wrapper*", se obtuvo los resultados de la Tabla 3, eliminando características consideradas no importantes hasta obtener un número de las mismas que mantengan o mejoren el porcentaje de exactitud del sistema.

Tabla 3
Porcentaje de exactitud reduciendo características menos importantes.

Núm. Características	% Exactitud
68	78,13
63	78,13
58	84,37
53	84,37
48	84,37
43	81,25
40	90,63
38	87,5
33	88,37

Los datos presentados en la Tabla 3, son representados en la Figura 18 para una mejor visualización de los resultados.

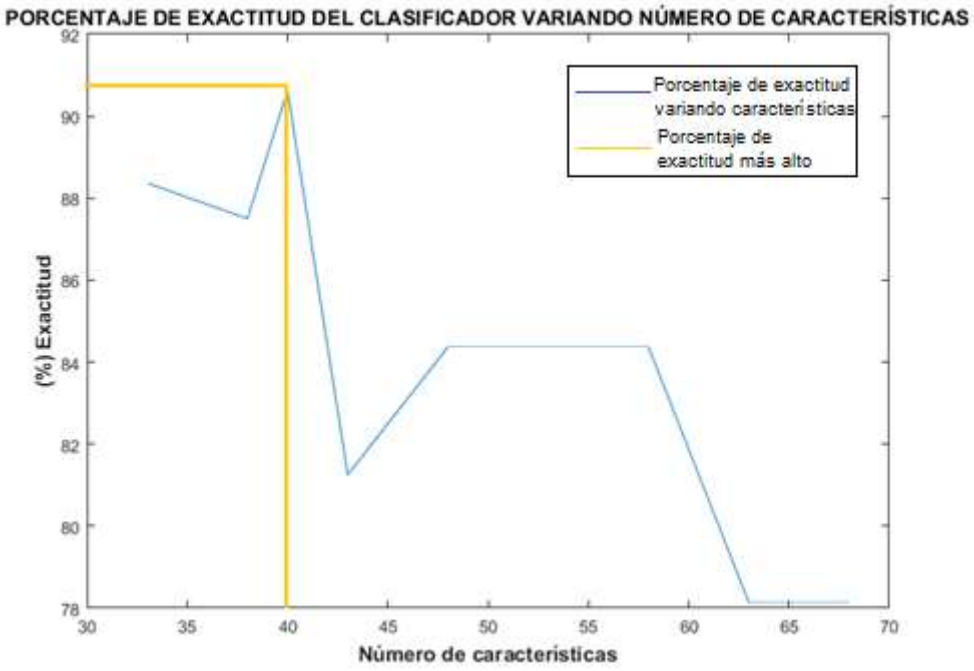

Figura 18. Porcentaje de exactitud del sistema variando el número de características

De la figura 18 se determina que con 40 características (ANEXO C) se obtiene un porcentaje de exactitud de 90.63%, sistema del cual se eliminaron 28 características menos importantes, con esto se puede decir que el número de características eliminadas no aportaban al entrenamiento del sistema, sino por el contrario estaban entorpeciendo, de tal manera que el porcentaje de exactitud disminuya.

A continuación se realiza un análisis con los datos sin procesamiento, datos estandarizados y datos normalizados realizando una comparación del sistema entrenado que presenta un mejor rendimiento.

4.1.2 Aprendizaje del sistema con datos sin procesar, estandarizados y normalizados

El aprendizaje automático se realizó una vez obtenido la extracción y selección de características, siendo estas 40 en total, y el análisis se realiza con los datos obtenidos directo de la extracción de características sin procesamiento. Los datos estandarizados se obtiene con la función

de Matlab® "*zscore*", la cual transforma la matriz original en otra del mismo tamaño, y cada columna (columna de características) de esta nueva matriz tiene media cero y desviación estándar uno. Y la matriz normalizada, al valor más alto de cada columna le asigna uno y normaliza al resto de valores con respecto a ese valor, los parámetros de evaluación se muestran en la Tabla 4.

Tabla 4

Parámetros de evaluación de sistema entrenado con datos sin procesar, estandarizados y normalizados

	Parámetros de evaluación			
Tipo Datos	**Exactitud (%)**	**Precisión (%)**	**Sensibilidad (%)**	**Especificidad (%)**
Sin Procesar	90,63	88,24	93,75	87,5
Estandarizados	81,25	77,88	87,5	75
Normalizados	78,13	90,91	62,52	93,75

Para una mejor visualización y apreciación de los datos presentados en la Tabla 4, estos son representados en la Figura 19.

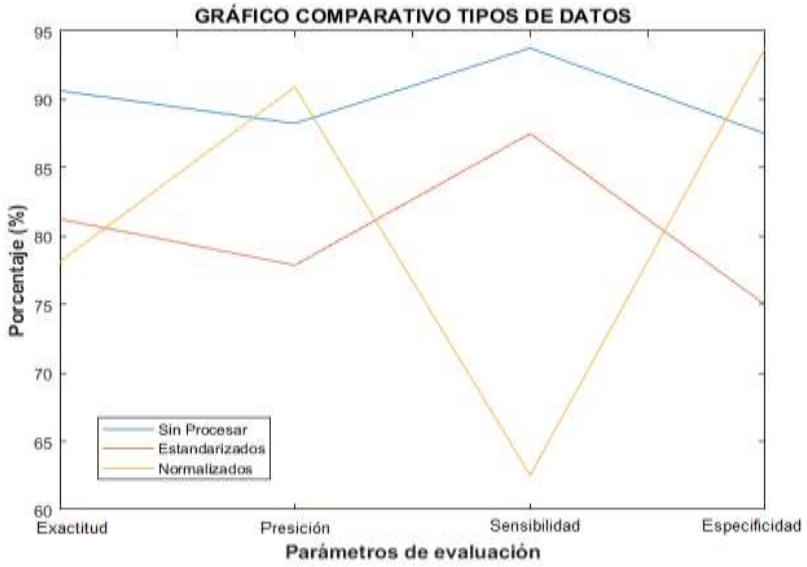

Figura 19. Gráfico comparativo entre los tipos de datos, sin procesar, estandarizados y normalizados.

La Figura 19 muestra el resultado de la comparación entre los diferentes tipos de datos utilizados para el entrenamiento del sistema, para lo cual el modelo de entrenamiento con datos sin procesamiento presenta un porcentaje de exactitud del 90.63% siendo el más alto, mientras que en la precisión el porcentaje más alto es el de los datos normalizados con 90.91%, en sensibilidad el porcentaje más alto lo tiene el entrenamiento con datos sin procesar con 93.75% y la especificidad con datos normalizados con el 93.75%. Los datos sin procesar presentan una curva más estable en cuanto a los parámetros de evaluación con respecto al entrenamiento con datos normalizados y el entrenamiento con datos estandarizados se encuentra por debajo de los porcentajes presentados por el sistema con datos sin procesar.

4.1.3 Características más importantes en la clasificación

En el proceso de selección de características, se determina las características con más y menos importancia dentro de la clasificación, en la Tabla 5 se presentan las 10 características más importantes, en la cual la que ocupa el primer lugar es la Frecuencia fundamental de la señal transformada con wavelet.

Tabla 5

Características más importantes en la clasificación, obtenidas mediante el método FSV de selección de características.

Núm. Orden	Características
1	*Wavelet Pitch*
2	*Pitch*
3	Desviación estándar de flujo espectral
4	Wavelet *Jitter*
5	*Kurtosis*
6	*Jitter*
7	*Shimmer*
8	Desviación estándar de la entropía de la energía
9	Media de la entropía de la energía
10	Mediana de la entropía de la energía

4.2 Análisis de resultados de clasificación de mujeres

De igual manera que la clasificación para el género masculino, una vez extraídas las 68 características en una matriz se realiza la clasificación con la aplicación "*Classificarion Learner*" de Matlab®, obteniendo los resultados presentados en la Tabla 6, presentando los clasificadores y el porcentaje de exactitud que cada uno presenta ante el *test*.

Tabla 6

Porcentaje de Exactitud de cada clasificador utilizando la aplicación de "Clasification Learner" (Mujeres).

CLASIFICADOR	TIPO	PARÁMETROS DEL CLASIFICADOR		% EXACTITUD
		Núm. Divisiones (ramas)		
Árboles de Decisión	Fine	100		71,90%
	Medium	20		71,90%
	Coarse	4		68,80%
		Número de vecinos (K)		
K-nearest neighbors (KNN)	Fine	1		75,00%
	Medium	10		75,00%
	Coarse	100		46,90%
		Función Costo (C)	**Gamma (γ)**	
Support Vector Machine (SVM)	Lineal	1	-	83%
	Cuadrático	1	1	81.30%
	Cubico	1	1,6	79.70%
	Gaussiano	1	6,3	48.40%

A partir de los datos de la Tabla 6, se obtiene la Figura 20, en la cual se puede visualizar de mejor manera el porcentaje de exactitud de cada uno de los clasificadores.

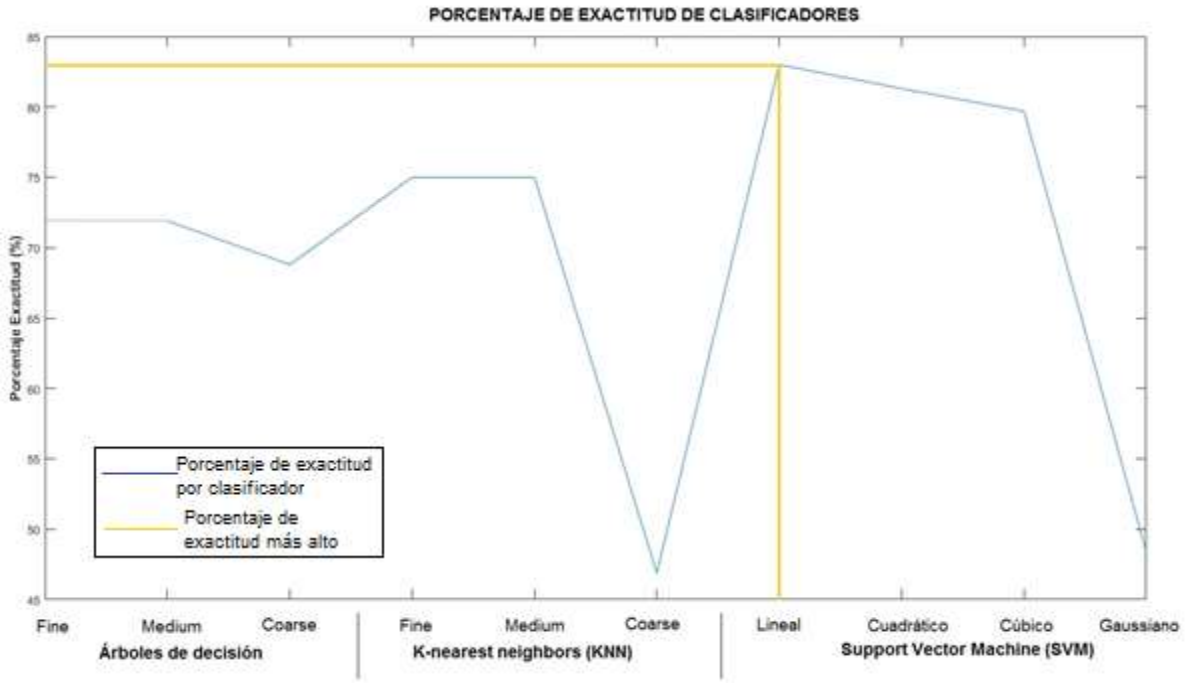

Figura 20. Porcentaje de exactitud modelos de clasificación Mujeres.

De la Figura 20 se determina que el clasificador con mayor porcentaje de exactitud es el de *Support Vector Machine (SVM)* tipo lineal, con 83% de exactitud en las predicciones con la tabla de *test*. Por lo tanto es este clasificador el que se utilizará para el análisis de señales de voz y determinar el engaño o verdad en mujeres. En el caso del clasificador a SVM lineal, el valor establecido por Matlab®, para la función de costo es: $C = 1$, debido a que estos modelos son proclives a sufrir *overfitting* (ajuste excesivo o aprendizaje a detalle).

4.2.1 Selección de Características

La selección de características (*features selection*), busca reducir el costo computacional e identificar características que no sean relevantes o a su vez estén entorpeciendo al aprendizaje del sistema de clasificación. Se utilizó el método correspondiente para el clasificador SVM, FSV (*feature selection via concave minimization*) de tipo "*wrapper*", y se obtuvo los resultados de la

Tabla 7, eliminando características consideradas no importantes hasta obtener un número de las mismas que mantengan o mejoren el porcentaje de exactitud del sistema.

Tabla 7

Porcentaje de exactitud reduciendo características menos importantes en mujeres.

Núm. Características	% Exactitud
68	87.5
63	87.5
58	84,33
53	87.5
48	84,33
43	84.33
40	90,63
38	81.25
33	81.25

Los datos presentados en la Tabla 7, son representados en la Figura 21 para una mejor visualización de los resultados.

Figura 21. Porcentaje de exactitud del sistema variando el número de características

De la figura 21 se determina que con 40 características (ANEXO D) se obtiene un porcentaje de exactitud de 90.63%, sistema del cual se eliminaron 28 características menos importantes, con esto se puede decir que el número de características eliminadas no aportaban al entrenamiento del sistema, sino por el contrario estaban entorpeciendo, de tal manera que el porcentaje de exactitud disminuya.

4.2.2 Aprendizaje del sistema con datos sin procesar, estandarizados y normalizados

Una vez realizado la extracción y selección de características, se realiza la predicción y análisis con tres diferentes tipos de datos. Los datos obtenidos directo de la extracción de características sin procesamiento. Los datos estandarizados obtenidos con la función de estandarización de Matlab® "*zscore*", transformando a cada columna de la matriz (columna de características) con media cero y desviación estándar uno. Y la matriz normalizada, al valor más alto de cada columna le asigna uno y normaliza al resto de valores con respecto a ese valor, los parámetros de evaluación se muestran en la Tabla 8.

Tabla 8

Parámetros de evaluación de sistema entrenado con datos sin procesar, estandarizados y normalizados

	Parámetros de evaluación			
Tipo Datos	**Exactitud (%)**	**Precisión (%)**	**Sensibilidad (%)**	**Especificidad (%)**
Sin Procesar	90,63	88,24	93,75	87,5
Estandarizados	84,37	78.94	93.75	75
Normalizados	84.37	78.94	93.75	75

Para una mejor visualización y apreciación de los datos presentados en la Tabla 4, estos son representados en la Figura 22.

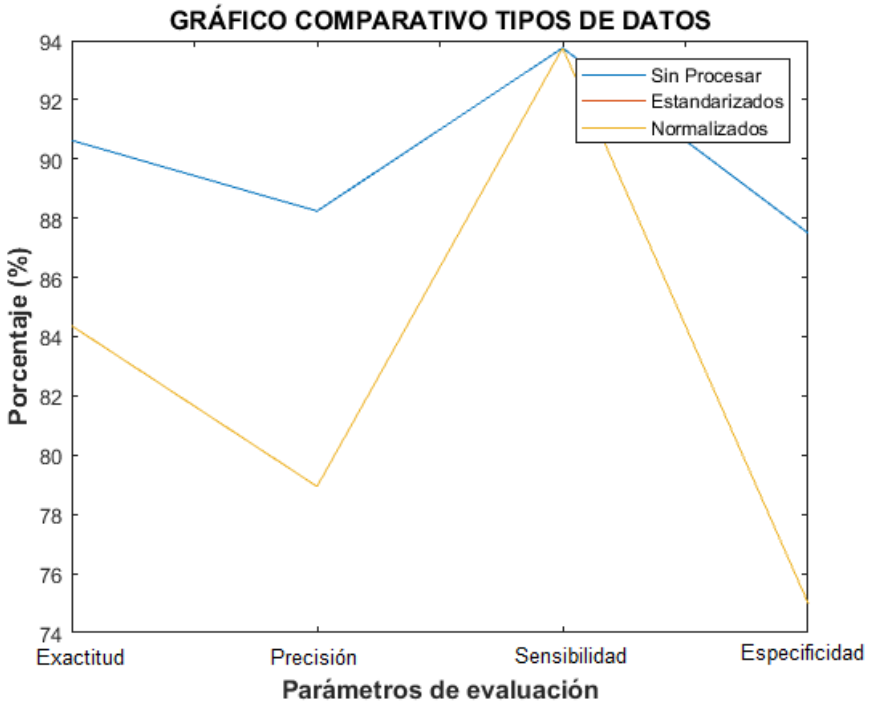

Figura 22. Gráfico comparativo entre los tipos de datos, sin procesar, estandarizados y normalizados (mujer).

En la Figura 22 se muestra el resultado de la comparación entre los diferentes tipos de datos utilizados para el entrenamiento del sistema (sin procesar, estandarizado, normalizado), teniendo como el modelo de clasificación con mejores resultados al que utiliza los datos sin procesar, los modelos con los datos estandarizados y normalizados presentan los mismos valores de los parámetros de evaluación.

4.1.3 Características más importantes en la clasificación

Los datos presentados en la Tabla 9, son las 10 características consideradas como las más importantes para la clasificación de engaños y verdades en mujeres, siendo la más importante la Frecuencia fundamental.

Tabla 9

Características más importantes en la clasificación, obtenidas mediante el método FSV de selección de características para mujeres.

Núm. Orden	Características
1	*Pitch*
2	*Jitter*
3	*Shimmer*
4	Desviación estándar de la entropía de la energía
5	*Media Spectral Roll-off*
6	Desviación estándar de Tasa de cruce por cero
7	Media de la energía
8	*Wavelet Kurtosis*
9	Wavelet Mediana Flujo espectral
10	Mediana *Spectral Centroid*

Todo el código implementado en Matlab®, se presenta en el ANEXO E.

CAPÍTULO V

CONLUSIONES Y RECOMENDACIONES

5. Conclusiones y Recomendaciones

- En el aprendizaje del sistema, se obtuvo que el clasificador con más porcentaje de exactitud es *linear Support Vector Machine* (SVM) con 73.4%, seguido de Árboles de decisión con 71.9%, por lo que el primero mencionado fue el utilizado para el realizar la clasificación de engaños y mentiras para el género masculino.
- Para el género femenino el clasificador que tuvo mayor porcentaje de exactitud es SVM lineal con 82.8% de exactitud, debido a que este clasificador presenta buenos resultados en clasificación de señales de baja frecuencia, se puede concluir que la separación de las dos clases se realizó por un hiperplano lineal, clasificando señales de audio con alto porcentaje de exactitud.
- Realizando la selección de características (*feature selection*), con el método FSV de "*wrapper*", método correspondiente para el clasificador SVM, se determinó que eliminando 28 características menos relevantes, quedando 40 características en total, el porcentaje de exactitud de predicción del sistema subió a 90.63%, concluyendo así que el sistema sufrió *overfitting* (ajuste excesivo o aprendizaje a detalle), esto significa que el ruido o las fluctuaciones aleatorias en los datos de entrenamiento son aprendidos por el modelo, el problema es que estos datos aprendidos no se aplican a nuevos datos y tiene un impacto negativo en la capacidad del modelo para predecir una nueva señal.
- En el caso de la selección de características de la tabla de datos de mujeres eliminando 28 características menos importantes se obtuvo el mismo porcentaje de exactitud de predicción

que el caso de los hombres con 90.63%, concluyendo de igual manera que el sistema sufrió *overfitting* (aprendizaje a detalle).

- Dentro de las características principales del habla de una persona para identificar un engaño o verdad se encuentran el *Pitch* (Frecuencia fundamental), *Jitter, Shimmer,* entre otras mencionadas en la sección 4, determinadas con el método de selección de características.
- Con la tabla de entrenamiento con las 40 características extraídas de cada señal se realiza tres análisis, uno con el clasificador entrenado con la tabla de datos sin procesar, el segundo con la tabla de datos estandarizados y el último con datos normalizados, con los cuales realizando una comparación el modelo entrenado con los datos sin procesar. En el caso de los hombres el modelo de clasificación con mejor rendimiento es el entrenado con datos sin procesar, teniendo porcentajes de exactitud, precisión, sensibilidad y especificidad de 90.63, 88.24, 93.75 y 87.5% respectivamente.
- Para el caso del género femenino se obtuvo mejor rendimiento en el modelo con datos sin procesamiento, teniendo valores de 90.63, 88.24, 93.75 y 87.5% de exactitud, precisión, sensibilidad y especificidad respectivamente.
- Las 28 características eliminadas en cada caso se realizó con el método FSV de "*wrapper*", el cual realiza una eliminación de características recursivas, evitando así el *overfitting.*
- Por lo tanto se recomienda no utilizar en estudios posteriores las características de tipo estadístico con la transformada de wavelet, ya que fueron las que presentaron alta recursividad, haciendo que el sistema disminuya el porcentaje de exactitud

- Se recomienda realizar una investigación de características que puedan aportar significativamente al aprendizaje del sistema con la finalidad de aumentar el porcentaje de exactitud de predicción.

CAPÍTULO VI

LÍNEAS DE TRABAJOS FUTUROS

6. Trabajos Futuros

- Como trabajos futuros se propone aumentar el porcentaje de los parámetros de evaluación de los clasificadores, con la finalidad de que el sistema tenga un mayor porcentaje de exactitud de predicción
- Investigar y aumentar el número de características importantes extraídas para de esa manera mejorar la predicción del sistema teniendo en cuenta no elevar el costo computacional del mismo.
- Realizar un sistema de predicción de engaños y verdades online, es decir que este pueda entrenarse o modificarse con parámetros ingresados en tiempo real.
- Crear una interfaz gráfica de la aplicación para mejorar la interactividad con el usuario.

CAPÍTULO VII

7. Bibliografía

A real case solved by CVSA . (2001). *National Institute for Truth Verification*.

Aguirre, F. (2017). Desarrollo y análisis de clasificadores de señales de audio. *UNIVERSIDAD POLITECNICA DE VALENCIA*.

Arciuli, J., Villar, G., & Mallard, D. (2009). Lies, lies and more lies. *31st Annual Meeting of the Cognitive Science Society(CogSci 2009)*, (págs. 2329-2334).

Bagnato, J. I. (12 de 12 de 2017). *Machine Learning: Underfitting y Overfitting*. Obtenido de https://www.aprendemachinelearning.com/que-es-overfitting-y-underfitting-y-como-solucionarlo/

Beers, M., & Berkow, R. (1999). The Merck Manual of Diagnosis and Therapy,. *John Wiley & Sons*.

Beigi, H. (2018). *Fundamentals of Speaker Recognition* . Springer.

Boersma, P., & Weenink, D. (2010). *Praat: doing phonetics by computer.* Obtenido de University of Amsterdam: http://www.fon.hum.uva.nl/praat/

Carreño, A. (2017). Detección de sucesos raros con Machine Learning. *Universidad Politecnica de Madrid*.

Cavalcanti de Almeida, N. (2010). *Sistema Inteligente para Diagnostico de Patologias na Laringe.* Obtenido de Universidade Federal do Rio Grande do Norte.

Cestaro, V. (1996). A comparison between decision accuracy rates obtained using the polygraph instrument and the Computer Voice Stress Analyzer (CVSA) in the absence of jeopardy. 117-127.

Chen, S. (2010). Improved voice activity detection algorithm using wavelet and Support vector machine. . *Computer Speech & Language*, 531-543.

DePaulo, B., Stone, J., & Lassiter, G. (s.f.). Deceiving and detecting deceit, in The Self and Social Life. *B. R. Schlenker*, 323-370.

Djebbar, F., & Ayad, B. (2017). Energy and Entropy Based Features for WAV Audio Steganalysis. *Journal of Information Hiding and Multimedia Signal Processing, 8*, 168-181.

dos Santos Lopes, J. M. (Julio de 2008). *Ambiente de análise robusta dos principais parâmetros qualitativos da voz.* Obtenido de Faculdade de Engenharia da Universidade do Porto.

Duque Sanchez, C., & Morales Perez, M. (2007). *Caracterización de voz empleando análisis tiempo-frecuencia.*

Ekman, P. (2007). *Emotions Revealed: Telling Lies.* Obtenido de https://zscalarts.files.wordpress.com/2014/01/emotions-revealed-by-paul-ekman1.pdf

Enríquez Fustillos, J. A. (2017). *Extracción de Caracteristicas de señales Sismicas Empleando Wavelets*. Obtenido de Universidad de las Fuerzas Armadas ESPE: http://repositorio.espe.edu.ec/xmlui/bitstream/handle/21000/13107/T-ESPE-057239.pdf?sequence=1&isAllowed=y

Erro, L., & Tonantzintla, M. (2010). Reconocimiento de Emociones a Partir de Voz basado en un Modelo Emocional Continuo. *Coordinación de Ciencias Computacionales INAOE*.

Español, E. (2018). *La Policía desarrolla la primera app a nivel mundial para detectar denuncias falsas*. Obtenido de https://www.elespanol.com/espana/politica/20181027/policia-desarrolla-primera-mundial-detectar-denuncias-falsas/348715373_0.html

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD explorations newsletter*, 65-75.

Evangeopulos, G., & Maragos, P. (2006). Multiband modulation energy tracking for noisy speech detection.

Gandhi, R. (07 de 06 de 2018). *Support Vector Machine—Introduction to Machine Learning Algorithms*. Obtenido de https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47?fbclid=IwAR1glgfqff1P02TeuiMuGNxVgFe_vub3qsT61SJsC255JWlIRh4yWZb5FmI

García, C., & Gómez, I. (2012). *ALGORITMOS DE APRENDIZAJE: KNN & KMEANS*. Obtenido de Inteligencia en Redes de Telecomuncicación.

Guimaraes, I. (2007). *A Ciencia e a Arte da Voz Humana.* Obtenido de Escola Superior de Saúde de ALcoitao.

Haddad, D. (2002). Investigation and Evaluation of Voice Stress Analysis Technology. *Final Report for National Institute of Justice*.

Haddad, D., Walter, S., Ratley, R., & Smith, M. (2002). Investigation and Evaluation of Voice Stress Analysis Technology. *The U.S. Department of Justice* .

Hamm, B. (09 de 04 de 2019). *Box constraint in SVMtrain function*. Obtenido de Mathworks: https://la.mathworks.com/matlabcentral/answers/301213-what-is-box-constraint-in-svmtrain-fucntion?fbclid=IwAR0o1Rly0MU2Jy_botOOB6VUn3v9zCy8vzmlK4nagnvnB2mLXt1BuO1mxeo#toggle-comments

Hernández, H. (12 de 2018). *Aplicación de la Transformada Ondicular y uso de Splines cúbicos para la mejora de la imagen sismica.* Obtenido de Universidad Nacional Autónoma de México.

Hogset, C. (1995). *Técnica vocal.*

Hollien, H., Geison, L., & Hicks, J. (1987). Voice stress evaluators and lie detection. *Journal of Forence Science, JFSCA*, 405-418.

Hopkins, C., Benincasa, D., Ratley, R., & Grieco, J. (2005). Evaluation of voice stress analysis technology. *Hawaii International Conference on System Sciences (IEEE)*.

Huang, X. (2001). *Spoken language processing.* (P.-H. Inc, Ed.)

Jithin, N., Nandan, M., Eldho, G., Akhil, R., & Sreehari, V. (2017). Lie Detection Through Voice Stress Analysis. *IJSTE: National Conference on Technological Trends (NCTT)*, 77-82.

Lee, C. (2009). Emotion Recognition Using a Hierarchical Binary Decision Tree Approach. *Interspeech. - Brighton*.

Liscombe, J., Riccardi, G., & Hakkani-Tur, D. (2005). Using context to improve emotion detection in spoken dialog systems. *Eurospeech. - Lisboa*.

Livingstone SR, Russo FA. (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.* Obtenido de PLoS ONE 13(5): e0196391. : https://doi.org/10.1371/journal.pone.0196391

Mallat, S. (1998). Wavelet Tour of Signal Processing. *Boston: American Press*.

Martínez Rodríguez, J. (2018). *Estudio Comparativo de modelos de Machine Learning para la detección de dianas microARN*. Obtenido de Universitat Oberta de Catalunya (UOC).

Mesa, J., & Morales , P. (2004). Codificacion, Sintesis y Reconociientgo de Voz. *Universidad de las Palmas Gran Canaria*.

Moulines, E., & Laroche, J. (1995). *Speech Comunications.*

Murray, I., Baber, C., & South, A. (1996). Towards a definition and working. *Elsevier Science Publishers*, 3-12.

Nasibov, Z. (2012). Decision fusion of voice activity detectors. *Master thesis. School of computing, University*.

News, B. (2003). *Lie detectors "cut car claims"*. Obtenido de http://news.bbc.co.uk/1/hi/uk/3227849.stm

Nilsson, N. (2005). Machine learning. *Stanford University*.

Pabón Ángel , H. J. (2010). Probabilidad y Estadística con Matlab para investigadores. *Facultad de Ingenieria de Sistemas, Universidad de Cundimarca Seccional Ubaté*.

Patel, S. (03 de 05 de 2017). *SVM (Support Vector Machine) Theory*. Obtenido de https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72?fbclid=IwAR3SZUPAmEl7xvdjzrbkbpiuKUJTlxA-BPqzitmcatWLteQtJlz2ZTQOUd8

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *Technical report, IRCAM*.

Pittermann, A., & Pittermann, J. (2006). Getting Bored with HTK? Using HMMs for Emotion Recognition. *8th International Conference on Signal Processing (ICSP).-Guilin, China*.

Planet, S. (2009). GTM-URL. *Contribution to the Interspeech 2009 Emotion Challenge*.

Polzehl, T. (2009). Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features. *Interspeech* .

Rodellar, B., Palacios, A., Nieto, L., & Gomez , V. (2015). Analysis of emotional stress in voice for deception detection. *International Work Conference on Bio-inspired Intelligence (IWOBI)*, 127-133.

Rodriguez, K., Castañeda , A., & Ballesteros , D. (2015). Voice activity detection based on the discrete wavelet transform. *Fundacion IAI*, 11-16.

Roffo, G. (2018). *Feature Selection Library (MATLAB Toolbox)*. Obtenido de University of Verona: https://www.groundai.com/project/feature-selection-library-matlab-toolbox/?fbclid=IwAR1buDl2_rKbZdjxNtnnU23cys_8gVg-9a-cucgw89br9cmWg-dQlk3YLV8

Sahi, B. D., & Shai, S. S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University. Obtenido de https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Sangwan, A., & Chiranth, M. (2002). VAD techniques for real-time speech transmission on the Internet. . *Proceedings 5th International Conference on High Speed Networks and Multimedia Communications*, 46-50.

Scherer, K., & Oshinsky, J. (1977). Cue Utilization in Emotion Attribution from Auditory Stimuli. *Motivation and Emotion*, 331-346.

Serrano, E. P. (2000). Introducción a la transformada Wavelet y sus aplicaciones al proceamiento de señales de emisión acústica . *Escuela de Ciencia y Tecnología - Universidad Nacional de General San Martín*.

Shannon, C. (July de 1948). A mathematical theory of communication. *Bell System Technical Journal,, 27*, 379423 - 623656.

Smart Lab. (2018). *Ciencia de la musica, investigacion auditiva y tecnologia.* Obtenido de https://smartlaboratory.org/ravdess/

Sobreira, M., & Rodríguez, A. (2008). CLASIFICACIÓN AUTOMÁTICA DE FUENTES DE RUIDO DE TRÁFICO. *Universidad de Coimbra*.

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

Titze, I. (1994). *Workshop on Acoustic Voice Analysis.* Obtenido de National Center for Voice ans Speech.

Vargas, F. (2003). Selección de características en el análisis acústico de voces. *Master's thesis: Universidad Nacional de Manizales*.

Vlasenko, B. (2007). Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. *Affective Computing and Intelligent Interaction (ACII2007). - Lisbon*, 139-147.

Vogt, T., & André, E. (2009). Exploring the benefits of dizcretization of acoustic features for speech emotion recognition. *Interspeech. - Brighton, U.K.*

Wollmer , M. (2009). Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks. *ICASSP. - Taipei, Taiwan.*

Xianfeng, L. (Agosto de 2005). *Voice Stress Analysis: Detection of Deption.*