SEP TecNM

TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE ACAPULCO

TEMA:

IDENTIFICACIÓN DE ESTADOS EMOCIONALES A TRAVÉS DEL ANÁLISIS ACÚSTICO.

OPCIÓN I:
TESIS PROFESIONAL

QUE PARA OBTENER EL TÍTULO DE:
MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:
ING. VICENTE BELLO AMBARIO

DIRECTOR DE TESIS:
DRA. MIRIAM MARTÍNEZ ARROYO

CO-DIRECTOR DE TESIS:
DR. JOSÉ ANTONIO MONTERO VALVERDE

ACAPULCO, GRO. NOVIEMBRE 2018.

A Dios

Por regalarnos un poco de su sabiduría.

A mis padres

Por su apoyo, esmero y dedicación

para poder salir adelante en ésta etapa de nuestras vidas.

A mis hermanos

Por estar ahí cuando más los necesitamos.

A Magaly

Por entregarme su tiempo, corazón y cada detalle

que me hizo fuerte en los momentos más difíciles.

Agradecimientos

A mi Familia

A mis padres Vicente y Carolina, mis hermanos, Arturo e Hibrain que siempre me han dado su apoyo incondicional y a quienes debo este triunfo profesional, por todo su trabajo y dedicación para darme una formación sobre todo humanista y espiritual. De ellos es este triunfo y para ellos es todo mi agradecimiento.

A mis Profesores

Agradezco infinitamente a la Dra. Miriam Martínez Arroyo por ser una excelente guía en esta tesis y al Dr. José Antonio Montero Valverde por sus valiosas observaciones.

Sinceras gracias al Dr. Eduardo de la Cruz Gámez y al M.T.I. Eloy Cadena Mendoza, quienes me asesoraron y atendieron mis dudas en la realización de esta tesis.

Descargo de responsabilidades

Descargo de responsabilidad institucional.

El que suscribe declara que el presente documento de tesis titulado: “Identificación de estados emocionales a través del análisis acústico” es un trabajo propio y original, el cuál no ha sido utilizado anteriormente en institución alguna para propósitos de evaluación, publicación y/o obtención de algún grado académico.

Además se han reconocido todas las fuentes de información utilizadas, las cuales han sido citadas en la sección de referencias bibliográficas de este trabajo.

Ing. Vicente Bello Ambario

Nombre

26 de noviembre de 2018

Fecha y firma

Resumen

El reconocimiento automático de las emociones humanas mediante el análisis de la voz, es un área de investigación activa debido a la amplia variedad de aplicaciones: telecomunicaciones, aprendizaje, interfaz humano-computadora y entretenimiento. En este trabajo se muestra una metodología para el reconocimiento de emociones analizando segmentos de voz. La metodología se basa principalmente en la transformada rápida de Fourier (FFT) y coeficientes de correlación de Pearson. El tono *(Pitch)*, frecuencia fundamental *(F_0)*, la intensidad de la señal de voz (energía) y la tasa de habla se han identificado como importantes indicadores de la emoción en la voz. El sistema tiene una interfaz gráfica que permite la interacción del usuario por medio de un micrófono integrado en la computadora, la cual procesa automáticamente los datos adquiridos. En nuestro entorno los seres humanos estamos programados para dejar que nuestra voz fluya de múltiples formas para comunicar, y captar a través de ella los estados emocionales propios de la región. Existen diversas investigaciones donde se utiliza la base de datos de Berlín, la cual es gratuita y muchos investigadores han utilizado en sus trabajos donde los resultados reportados no han sobrepasado al 80 % con la cual inicialmente se trabajó. Sin embargo la creación de un *corpus emocional* con frases en español fue necesaria para realizar pruebas que nos ofrezcan resultados más claros. El corpus contiene 16 frases por emoción creada por 11 usuarios (9 mujeres y 2 hombres) con un total de 880 muestras de audio. Se consideran las siguientes emociones básicas: *disgusto, ira, felicidad, miedo y neutral*. El algoritmo de reconocimiento de emociones da ofrece un 80 % de efectividad en los resultados obtenidos.

Índice general

Índice de figuras

Índice de tablas

Índice de Ecuaciones

No de Ecuación Pag.

Capítulo 1

Generalidades

1.1. Antecedentes del Problema a Resolver

Uno de los primeros libros de Darwin, fue *La expresión de las emociones en animales y humanos*, la teoría dice que somos capaces de manifestar las emociones y resalta la importancia en las relaciones sociales, de manera que sugiere que la parte emocional de la voz evoluciona por ese motivo [Darwin, 1872]. La investigación sobre la expresión emocional ha estado prioritariamente dirigida en el canal facial, y aunque ya en 1872 el propio Darwin señalo la importancia de los aspectos no verbales del habla como medios de expresión [Carrera and Fernández, 1988]. Las dificultades de tipo técnico que conlleva a la evaluación de la conducta vocal y la fusión respecto a su estatus como conducta no verbal son dos de las razones que han motivado este desface con respecto a la expresión facial [Scherer, 1982].

Los estudios sobre expresión vocal de la emoción pueden encuadrarse en dos grandes grupos: los estudios de expresión y los de reconocimiento. Los estudios de expresión, tienen como objetivo básico determinar como un estado emocional se expresa o exterioriza a través de la voz del sujeto [Brown and Bradshaw, 1985], su evaluación puede realizarse en dos niveles, el análisis acústico y los juicios perceptivos de los parámetros acústicos.

Los estudios de reconocimiento tratan de conocer en qué medida el receptor es capaz de identificar, a través de los aspectos no verbales de la voz del emisor y el estado emocional expresado por este.

Uno de los principales problemas con lo que se ha enfrentado la investigación sobre reconocimiento de emociones en la voz ha sido la separación del canal verbal del vocal, de modo que el contenido lingüístico del mensaje no condicione a los juicios sobre el estado afectivo del codificador. Se ha establecido que el habla es un evento acústico que contiene información importante sobre el funcionamiento del sistema nervioso central, y por lo tanto acarrea información sobre el estado emocional de un individuo.

Diversos trabajos tratan sobre el hecho de que unas emociones son mejor reconocidas que otras. Fechner plantea que el mejor reconocimiento de las emociones negativas puede explicarse en términos de su mayor valor de supervivencia [Fechner, 1978]. En los estudios con adultos, Scherer propone tras su revisión de estudios de reconocimiento, la hipótesis que señala que el enojo suele ser la emoción mejor reconocida, seguida de tristeza, indiferencia y alegría [Scherer, 1981].

El problema aparece cuando se pretende que un sistema automático sea capaz de manipular un conjunto grande de estados emocionales, tal y como hacen los humanos, es decir, cuando se intenta conseguir un alto grado de resolución en la clasificación (Cowie et al., 2001). Reducir este conjunto a uno más pequeño que recoja unas cuantas emociones, que puedan considerarse *básicas*, es bastante complicado, y la determinación de cuántas y cuáles serían estas emociones depende mucho de los autores consultados. No obstante, es importante hacer esta reducción para poder conseguir una clasificación más sencilla y, en consecuencia, más correcta de las emociones detectadas por parte del sistema reconocedor. De hecho, esta reducción no supondrá perder el conjunto global de emociones a reconocer

ya que se puede afirmar, según Plutchik, que se pueden formar por combinación de un conjunto de estados emocionales básicos [Plutchik, 1980]. Ekman y Friesen indican cuatro emociones básicas: *alegría, tristeza, miedo e ira* poniendo de manifiesto la posibilidad de diferenciar emociones a partir de la parametrización de la voz [Ekman and Friesen, 1971].

Al expresar de manera vocal una emoción, nos hemos encontrado con varias aproximaciones descriptivas de la existencia de un patrón acústico posible en su caracterización, en especial en emociones específicas [Sundberg et al., 2011] [Parent, 2005] [Bachorowski, 1999], aún cuando no parece existir evidencia indiscutible respecto de qué variables describen la señal sonora que refleja la emoción.

Algunos autores como Scherer puntualizan que los parámetros de la voz permiten descripciones en términos de la intensidad sonora asociada a emociones diferentes independiente de las características propias de cada lengua, si bien se conoce que los rasgos acústicos de la señal vocal resultan modificados de manera al menos parcial durante una locución emocionada, aún se desconoce el impacto específico de la emoción como generador de cambios en la señal acústica vocal [Scherer, 2000].

Con el progreso de las nuevas tecnologías y los sistemas interactivos, la importancia de reconocer emociones en la voz se ha incrementado enormemente, dado que la voz es el medio de comunicación más natural para los humanos, es necesario proporcionar interfaces para generar, reconocer y clasificar emociones en el habla. En la actualidad, los estudios se centran en encontrar nuevas combinaciones de clasificadores que aumenten la eficiencia de estas clasificaciones en aplicaciones de tiempo real.

1.2. Planteamiento del Problema

Las primeras preguntas que surgen al involucrarse en el reconocimiento de emociones a partir de la voz son:

1. ¿Qué evidencias existen de que en realidad los estados emocionales de las personas se reflejan en sus voces?
2. ¿Las emociones se reflejan de manera semejante en todas las personas?
3. ¿De qué depende la manera en que expresamos emociones con nuestra voz?

A pesar de los muchos intentos tratando de establecer una correspondencia entre emociones y voz no existe un conjunto definido de emociones universalmente aceptado. Hay varios modelos para representar las emociones los cuales son usados para su categorización y organización. Estas categorías difieren dependiendo de las diferentes tareas y aplicaciones.

El trabajo hecho a la fecha se ha centrado principalmente en características relacionadas con aspectos prosódicos. Sin embargo, se ha descubierto que entre más se acerca a un escenario realista, menos fiable es la prosodia como un indicador del estado emocional del hablante [Batliner et al., 2003], por lo tanto, es necesario encontrar características que complementen la información que proporciona el aspecto prosódico de la voz.

El reconocimiento de emociones en la voz es un problema que puede abordarse desde distintos frentes. Por una parte, es necesario elegir un sistema de reconocimiento de emociones que se adapte a nuestras necesidades. Por otro lado, la elección de las características acústicas de las muestras de voz incluidas en el proceso, así como los métodos utilizados para la extracción de las mismas es otro de los puntos críticos del reconocimiento de emociones.

La principal dificultad en los sistemas de reconocimiento de emociones en la voz es que son poco eficientes, solo reconoce la señal de voz que se registra en condiciones favorables. Sin embargo, cuando un sistema de reconocimiento se pone a funcionar en situaciones reales se encuentra con condiciones adversas tales como cambios en el hablante (condiciones fisiológicas, emocionales, cambio en el modo de articulación debido a un fuerte ruido ambiental, entre otras) y en el entorno acústico (ruidos, reverberación y ecos) o eléctrico (como ruidos o distorsiones de la señal provocados por el micrófono o el canal de transmisión), que son irrelevantes desde el punto de vista lingüístico pero que pueden degradar en gran medida la tasa de reconocimiento.

La preocupación por los aspectos afectivos en el desarrollo de los procesos de enseñanza aprendizaje ocupa especial importancia en los investigadores educativos y en los gestores de los centros de educación [Vargas et al., 2017].

La necesidad de un corpus emocional es evidente. En la actualidad en México, existen pocos repositorios de datos por lo tanto es probable que los sistemas actuales tarden algún tiempo en madurar lo suficiente como para presentarse como una alternativa de solución para el reconocimiento de estados de ánimo de alumnos por medio del análisis de la voz.

1.3. Objetivos

1.3.1. Objetivo General

Diseñar un sistema de reconocimiento de emociones a través del análisis de voz mediante un modelo estadístico.

1.3.2. Objetivos Específicos

- Identificar las características determinantes de los estados afectivos en el habla que nos permita reconocer las emociones partiendo de una señal acústica.
- Diseñar una interfaz gráfica de usuario para realizar el proceso de grabación de voz.
- Crear una base de datos de voz emocional.
- Diseñar el algoritmo de reconocimiento de emociones (preprocesamiento, extracción de características y clasificación).
- Evaluar los resultados obtenidos.

1.4. Hipótesis

La identificación de un conjunto de características acústicas y modelos estadísticos, permitirá clasificar emociones en la voz con un porcentaje de más del 80 % de eficiencia.

1.5. Justificación

Los sistemas educativos interactivos y, en particular, los sistemas tutor inteligente (STI), son concebidos como herramientas de apoyo a la enseñanza que permiten adaptarse a las necesidades específicas del estudiante en un dominio particular de conocimiento, mediante la provisión de ayudas específicas para la consecución de los objetivos pedagógicos propuestos. Estos sistemas se diseñan con la intención de simular el comportamiento de un profesor o tutor tradicional, ofreciendo de forma personalizada al alumno las pautas, recomendaciones y ayudas más adecuadas a su nivel de conocimiento y de aprendizaje en un contexto educativo, creando y evaluando en todo momento un modelo del estudiante típicamente fundamentado en su nivel de conocimiento y en su forma de aprendizaje.

De modo similar a un entorno presencial, donde el profesor o tutor dispone de la capacidad para valorar aspectos adicionales a los exclusivamente cognitivos que podrían influir en el aprendizaje, como el estado emocional del alumno, es factible pensar que dotar a los sistemas educativos interactivos de estas capacidades podría suponer una mejora en su rendimiento. Estudios previos han reportado sólidas evidencias de que el estado emocional del estudiante puede tener un impacto significativo en su motivación y, en consecuencia, en el rendimiento de su aprendizaje.

Se considera muy significativo evaluar el impacto de las estrategias instruccionales implementadas en un STI sobre las variaciones del estado afectivo de sus estudiantes, especialmente en entornos educativos interactivos realistas, donde es esencial que los dispositivos necesarios para la captura de la información que pueda dilucidar el estado emotivo sean poco intrusivos. En este sentido, la información que puede resultar relevante para determinar el estado afectivo del estudiante puede ser de naturaleza física, fisiológica o de comportamiento [Marco, 2017].

Existen diversas aplicaciones donde se puede aprovechar el conocimiento del estado emocional de los usuarios para tomar decisiones sobre qué acciones debe seguir un sistema; en el sector académico un tutorial interactivo en el que se podría adaptar la carga emocional de la respuesta del sistema buscando motivar y captar el interés dependiendo del estado emocional del alumno [Hernández et al., 2008].

1.6. Alcance del Proyecto

Se abarcan los estudios relacionados con parámetros acústicos como referencia en el reconocimiento de emociones y se desarrollo un prototipo para grabar muestras de audio utilizadas para realizar pruebas en la identificación del estado anímico de estudiantes.

Se creó un corpus de emociones grabada por jóvenes en el municipio de San Marcos, Guerrero. Se implementaron las fases de preprocesamiento, extracción de características y clasificación que dieron como resultado el reconocimiento de las emociones primarias. Se probó con más de 800 muestras de audio en la clasificación de estados emocionales.

1.7. Limitaciones

- El corpus se creó con emociones actuadas para probar el algoritmo de reconocimiento. Se pretende posteriormente realizar pruebas con emociones espontáneas.
- En la fase de pruebas se generaliza el reconocimiento de emociones en las voz en hombres y mujeres.
- La interfaz gráfica incluye únicamente el módulo de grabación.
- El sistema reconoce emociones básicas o primarias.

Capítulo 2

Antecedentes de la Investigación

El Reconocimiento de emociones de la voz (REV) es un campo de investigación de creciente relevancia que día a día se gana más adeptos.

El desarrollo de mejores algoritmos y de modelados más precisos, junto con la aparición de sistemas informáticos más potentes y asequibles, posibilita la integración de los sistemas de diálogo hombre-máquina a través de la voz en numerosos ámbitos de la sociedad actual. Estos sistemas de diálogo permiten el acceso a una gran cantidad de información a través de una forma de comunicación tan natural como es el habla, facilitando un elevado número de servicios interactivos utilizando el teléfono, la televisión o computadora como elementos de acceso.

El propósito de este capítulo es presentar los principales avances tecnológicos obtenidos en los últimos años en el ámbito de los sistemas de reconocimiento de emociones en él habla; dentro de la investigación científica se muestran diferentes características y técnicas utilizadas. La revisión de la literatura representa la importancia de elegir entre diferentes modelos de clasificación y características utilizadas por los autores. También se revisan los corpus emocionales, el lenguaje utilizado y la clasificación de emociones hasta la fecha.

Las principales líneas de trabajo están orientadas al campo de la inteligencia artificial con sistemas capaces de identificar un estado emocional de un hablante de forma precisa. Dentro de este contexto la computación emocional está siendo integrada en los robots con el propósito de establecer una interacción más natural y unida con los humanos.

2.1. Estado del Arte

En la Universidad Tecnológica de la Mixteca, Huajuapan de León, Oaxaca, en México, se estudia la integración de optimización evolutiva para el reconocimiento de emociones en voz [Pérez-Gaspar et al., 2015]. Se presenta el desarrollo de un sistema de reconocimiento de emociones basado en la voz. Se consideraron las siguientes emociones básicas: Enojo, Felicidad, Neutro y Tristeza. Para este propósito una base de datos de voz emocional fue creada con ocho usuarios mexicanos con 640 frases (8 usuarios x 4 emociones x 20 frases por emoción). Los Modelos Ocultos de Markov (Hidden Markov Models, HMMs) fueron usados para construir el sistema de reconocimiento. Basado en el concepto de modelado acústico de vocales específicas emotivas con un total de 20 fonemas de vocales (5 vocales x 4 emociones) y 22 fonemas de consonantes fueron considerados para el entrenamiento de los HMMs. Un Algoritmo Genético (Genetic Algorithm, GA) fue integrado dentro del proceso de reconocimiento para encontrar la arquitectura más adecuada para el HMM para cada vocal específica emotiva. Las frases emocionales fueron grabadas en un salón a puerta cerrada con la herramienta Wavesurfer en formato .WAV [Sjölander and Beskow, 2018] con una frecuencia de muestreo de 48000 Hz. La distancia entre el micrófono (micrófono interno de una computadora tipo laptop) y el usuario fue de alrededor de 60 cm. A cada voluntario se le pidió pronunciar cada una de las 20 frases por emoción llegando a un total de 80 muestras de voz por voluntario (80 frases x 8 usuarios = 640 frases).

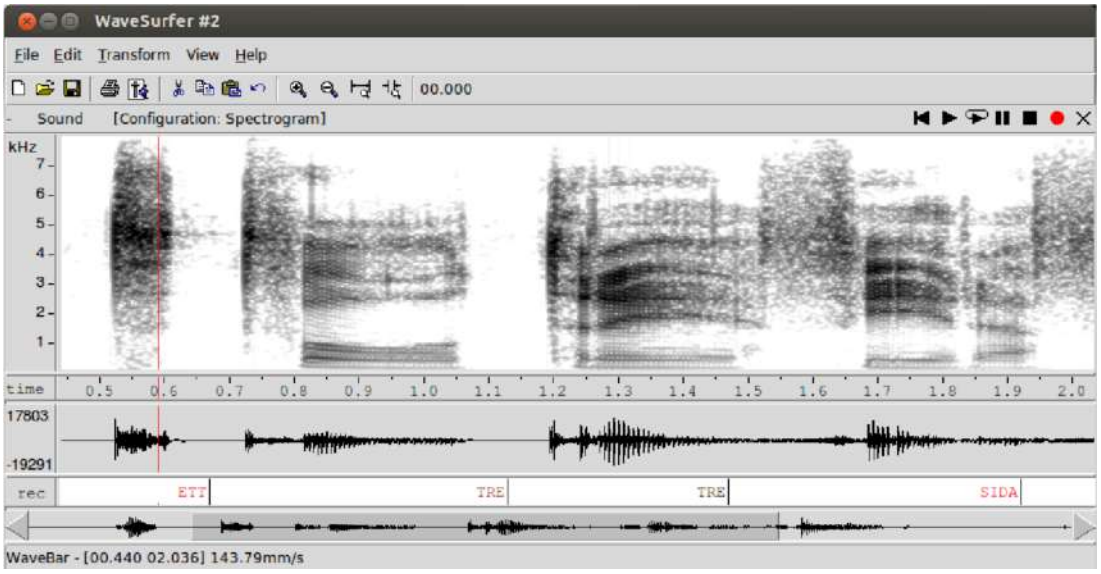

Figura 2.1: WaveSurfer usado en estudios de fonética acústica[Sjölander and Beskow, 2018].

Existen varias plataformas de reconocimiento automático de emociones. Emospeech es una plataforma que se adaptará a diferentes áreas de aplicación, tal como el análisis de las llamadas telefónicas recibidas en los centros de recepción o Call Centers. Esta plataforma dará consejos en tiempo real a los trabajadores para que dispongan de más información para incrementar las ventas y relacionarse con los clientes. Si un agente dispone de información acerca del estado anímico en el que se encuentra el usuario que ha contactado con el (enfadado o alegre), podrá atenderle de forma más adecuada. Emospeech está orientada en un principio a cubrir las necesidades de centros de llamadas en México, y su meta es crear estrategias que incrementen y mejoren la efectividad y la calidad de los servicios interpretando las emociones de los usuarios. Seria de utilidad, por ejemplo, que una persona que trabaje realizando llamadas comerciales tenga información sobre si el receptor de la llamada tiene cierto nerviosismo por que tiene prisa, por lo que no estará receptivo. Estos aspectos podrán ser analizados por Emospeech para que el software interprete las emociones tanto emisor como receptor en una llamada y así generar datos que ayuden a mejorar el acercamiento que las empresas deben tener hacia los usuarios [Noroozi et al., 2017].

Género	Emoción	Neutro	Alegria	Rabia	Ternura	Amor	Miedo	Tristeza
Hombres	Intensidad Media (dB)	40	53.7	50.6	50	51	55.8	58
	Tono Medio (Hz)	70	208	153	91	162	247	191
Mujeres	Intensidad Media (dB)	32	54.6	52	47	50.6	55	54
	Tono Medio (Hz)	157	236	211	138	188.6	276	244

Tabla 2.1: Patrones de intensidad y tono medios identificados.

También se llevan a cabo estudios en Chile, donde se establece la correlación intensidad/frecuencia para la expresión vocal de cada emoción básica, con el fin de determinar si la voz puede ser un reflejo de cada emoción. Por medio de un registro de audio dicho en español, por actores entrenados previamente en el método AE (Alba Emoting), durante un periodo de 4 meses. El texto fue expresado en las 6 emociones básicas más la neutralidad, mientras se reproducían los patrones efectores emocionales propuestos por el método, quedando registrada la voz en cada emoción interpretada por cada actor. De esta forma, 18 actores (varones y mujeres) dijeron el texto en cada una de las emociones más la neutralidad, modificando además el Nivel de Intensidad Emocional (NIE). Inicialmente y en forma promedio, en la tabla 2.1 se reflejan las emociones estudiadas con los patrones de intensidad y tono medio.

La intensidad promedio en varones se incrementa respecto de los valores obtenidos en una condición de habla neutra en todas las emociones, sin embargo, lo hace de manera más notoria en alegría, miedo y tristeza. En el caso de las mujeres sucede de manera similar. Las emociones con menor intensidad promedio correspondieron en ambos géneros a la ternura y el amor.

En el Instituto Tecnológico de Culiacán, Sinaloa, México; han realizado diversos estudios con PREMOC, una plataforma que brinda un servicio web para el reconocimiento de

emociones en texto, imágenes de rostros, sonidos de voz y señales electroencefalográficas (EEG) de manera mono-modal y multimodal. El reconocedor de emociones en voz que utiliza PREMOC identifica la valencia del audio recibido (positiva o negativa) y también consiste en dos fases: entrenamiento y ejecución [Zatarain-Cabada et al., 2016]. En la fase de entrenamiento se utilizaron un total de 45 audios de 9 sujetos diferentes, donde cada audio está clasificado en positivo o negativo. De los audios se extraen las características utilizando su espectrograma para obtener los parámetros para entrenar una máquina de soporte vectorial (SVM) de la librería LibSVM [Chang and Lin, 2011]. La instancia de la SVM ya entrenada es guardada para utilizarla online en la clasificación de sonidos de PREMOC. La fase de ejecución en línea se muestra en la Figura 2.2.

Figura 2.2: Algoritmo del reconocedor en voz [Zatarain-Cabada et al., 2016].

El desarrollo de las tecnologías de información y comunicación ha posibilitado la incorporación a diferentes áreas de la actividad humana, negocios, áreas académicas e incluso en la medicina. La Enfermedad de Parkinson (EP) es la segunda condición clínica neurodegenerativa más prevalente después del Alzheimer y para el sistema de salud mundial es fundamental identificar marcadores tempranos; sin embargo, en la actualidad es un campo nuevo de estudio que necesita un mayor desarrollo. Con una evaluación de métodos de Fourier y máxima entropía para La detección automática de la enfermedad de Parkinson [Arias Mejía et al., 2015], se ha demostrado que cerca del 90 % de los pacientes con EP también desarrollan deficiencias en la voz, mostrando síntomas como un habla monótona, baja intensidad en el tono, pausas aisladas, pronunciación imprecisa de consonantes y problemas en la prosodia.

En el ámbito de la investigación, el análisis tiempo-frecuencia ha demostrado ser una herramienta poderosa en el procesamiento de señales acústicas, más específicamente, el procesamiento de la voz. Con el ánimo de tener representaciones limpias del espectro que ayuden a mejorar la extracción de características y mitigar otros problemas resultantes al usar métodos clásicos de estimación, se busca estudiar el comportamiento del Método de Máxima Entropía (MEM) comúnmente usado en ciencias oceanográficas y astronomía, en el estudio en señales de voz [Arias Mejía et al., 2015]. Se estima que los déficits en el habla están presentes entre el 60-80 % de los enfermos con EP, se caracterizan por alteraciones en frecuencia, duración e intensidad. Se reportan pacientes con afectaciones del tono y la prosodia. En estos pacientes las alteraciones observadas se solapan con los cambios naturales que se producen en el anciano: modificaciones de la laringe, alteraciones en el sistema respiratorio, en las cavidades de resonancia y en los órganos articulatorios, consecuentes al deterioro en los músculos, cartílagos, articulaciones, ligamentos y mucosa laríngea. Existen variaciones patológicas en los parámetros acústicos de la voz que involucran principalmente la frecuencia fundamental y el Voice Onset Time (VOT). Su objetivo primordial es el de mostrar la alteración de parámetros acústicos de la voz y el habla en la EP [Martínez-Sánchez, 2010]. La figura 2.3 muestra la metodología utilizada para la detección automática de la EP.

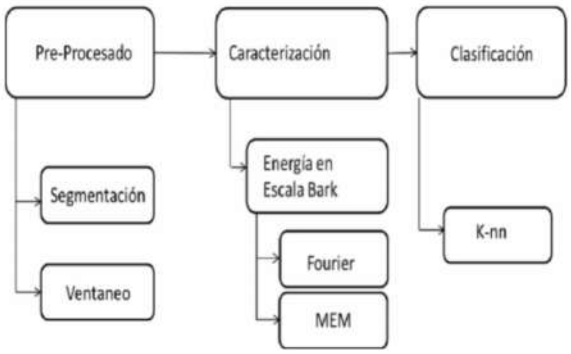

Figura 2.3: Diagrama de bloques de la metodología [Arias Mejía et al., 2015].

Anavoz 1.0 es una herramienta usada para la extracción y procesamiento de los parámetros acústicos. Fue llevado a cabo por miembros del Grupo de Procesamiento de Voz del Centro de Estudios de Neurociencias, Procesamiento de Imágenes y Señales, de la Facultad de Ingeniería Eléctrica de la Universidad de Oriente [Escobedo et al., 2008]. Anavoz 1.0 es un programa desarrollado en ambiente `MATLAB`[1] (The Mathworks, Inc), con las facilidades gráficas asociadas a este ambiente. Este programa se instala en computadoras con multimedia y periféricos.

El programa Anavoz 1.0 permite la adquisición, almacenamiento, edición y reproducción de diversos tipos de señales de voz, lenguaje (habla), etc. Igualmente, este programa facilita la medición, estimación y extracción de parámetros o atributos acústicos cuantitativos y cualitativos de la voz en Amplitud o Intensidad, Tiempo y Frecuencia. Anavoz 1.0 es una herramienta especializada de interfaz amigable y por sus características puede ser utilizada en aplicaciones en áreas relacionadas de diversas formas con la voz, el habla, el análisis del llanto infantil, entre otras [Pacheco et al., 2015] [Beceiro et al., 2018].

Para estos experimentos fue usada una base de datos de pacientes que presentan la enfermedad de Parkinson usando 50 registros de voces patológicas y la misma cantidad de voces sanas, donde se analizarán segmentos sonoros y no sonoros obtenidos de la palabra "PA-TA-KA", sobre los cuales fueron estimadas energías provenientes tanto de la transformada de Fourier, como del método de máxima entropía.

Con el fín de evaluar el rendimiento de los métodos clásicos y MEM se usó un clasificador del vecino más cercano (K-nn) y se encontraron tasas de acierto cercanas al 60 % al considerar MEM tanto en fonemas como en sílabas.

[1] https://es.mathworks.com/products/matlab.html

El proyecto Speech Emotion Recognition se enmarca dentro el campo de la Inteligencia Artificial (*IA*), en concreto del reconocimiento de emociones por voz, y propone implementar un sistema capaz de reconocer y clasificar un conjunto específico de emociones por medio del análisis de las características de la señal de voz. Para la realización de dicho análisis, se plantea el uso de Ivectors+PLDA (Probabilistic Linear Discriminant Analysis)[Pérez Pascual, 2017], una técnica de reconocimiento del locutor. Esta nueva técnica, que se considera una extensión del JFA (Joint Factor Analysis)[Kenny et al., 2007], se basa en definir un único espacio que contenga conjuntamente información del locutor y del canal, en lugar de dos espacios separados. Este nuevo espacio, llamado espacio de Variabilidad Total, contiene la variabilidad del locutor y del canal de una manera simultánea, lo que provoca que no se haga ninguna distinción entre el efecto de ambos componentes en el supervector GMM (Gaussian Mixture Models), que se construye concatenando las medias de las diferentes Gaussianas que forman el UBM (Universal Background Model).

ALIZE [Larcher et al., 2013] es una plataforma de código abierto escrita en C++ diseñada para lidiar con las tareas propias del área del reconocimiento del locutor. Puede usarse para tareas de reconocimiento de emociones, dado que las funcionalidades que contiene el programa son extrapolables a este campo.

ALIZE está construido con una arquitectura multicapa, como se puede observar en la figura 2.4, donde la capa base (ALIZE-Core) contiene las funciones básicas de entrada/salida; mientras que la capa superior (LIA_ RAL) presenta las funciones de más alto nivel requeridas para los problemas de reconocimiento del locutor.

Su funcionamiento está basado en binarios a los que el usuario le pasa un archivo de configuración con los parámetros que desea para sus experimentos correspondientes.

Figura 2.4: Estructura de ALIZE [Pérez Pascual, 2017].

La preocupación por los aspectos afectivos en el desarrollo de los procesos de enseñanza aprendizaje ocupa especial importancia en los investigadores educativos y en los gestores de los centros de educación.

Las investigaciones procuran hacer una revisión de los estados emocionales presentes en el proceso enseñanza-aprendizaje, en como incide el adecuado manejo de la afectividad para el desarrollo de habilidades y destrezas, así como el desarrollo de madurez emocional para la comprensión por el otro en la convivencia, con la finalidad de lograr estudiantes competentes, profesionales con fácil adaptación a los cambios y motivados al aprendizaje y a la construcción del conocimiento, constructores de nuevas realidades, capaces de mejorar el mundo que los rodea con valores de civismo, pluralismo, comprensión mutua y paz.

La afectividad consciente, la motivación, el interés, la buena disposición, los estímulos positivos, la empatía, son variaciones pedagógicas del principio que articula la cabeza con el corazón, la razón con el sentimiento, lo cognitivo con lo afectivo.

Emoción	Aciertos
Felicidad	93.44 %
Enojo	86.67 %
Tristeza	96.67 %
Miedo	70 %

Tabla 2.2: Porcentaje de detección de emociones entrenadas [Bustamante et al., 2015]

Un trato afectivo de parte del docente hacia los estudiantes tiene una gran repercusión en el desarrollo de la personalidad equilibrada y estable, que incide en el éxito académico. Considerando que la educación debe orientarse al pleno desarrollo de la personalidad de los alumnos, el desarrollo cognitivo, debe complementarse con el desarrollo emocion[Vargas et al., 2017].

Otros estudios utilizaron los enfoques: temporal, frecuencial y prosódico, para extraer características descriptivas de la señal de voz. Se desarrolló un sistema de clasificación mediante redes neuronales utilizando para el entrenamiento, cuatro de las seis emociones de la base de datos Berlín: felicidad, enojo, miedo y tristeza [Burkhardt et al., 2005]. Posteriormente, se procedió a localizar las emociones detectadas en plano excitación-valencia. El porcentaje de detección de emociones entrenadas se muestra en la tabla 2.2.

La mejor emoción detectada por este clasificador resultó ser la tristeza con una tasa de aciertos del 96,67 %. El miedo fue la emoción peor detectada, con un porcentaje de aciertos del 70 %. Esta emoción se confunde con la felicidad. Ambas emociones, a pesar de tener valencias opuestas, tienen alta excitación, por tanto, pueden ser fácilmente confundibles debido a que el clasificador utiliza características relacionadas con la potencia o energía de la señal.

Capítulo 3

Marco Teórico

En el siguiente capítulo se presentan las consideraciones teóricas referentes a esta investigación. Se explican los conceptos básicos en el proceso del habla y el reconocimiento de emociones en la voz. Se mencionan los parámetros del habla, la transmisión de emociones primarias, análisis de las características de la voz emotiva y las dimensiones emocionales.

3.1. Conceptos Básicos

Es necesario conocer algunos conceptos empleados en el estudio de las señales de voz, así como en el funcionamiento de los sistemas generadores de voz, de tal forma que se puedan establecer las características que sirvan para realizar un correcto reconocimiento de emociones a través de la voz.

En un entorno cotidiano las personas expresan sus emociones y estados afectivos mediante información procedente del rostro (expresiones faciales), del habla tanto con información explícita o lingüística (el mensaje), como implícita o paralingüística (características prosódicas como el tono de la voz, la intensidad, la velocidad o el ritmo) y del cuerpo (gestos de las manos y posturas o movimientos del cuerpo) [Marco, 2017].

El Reconocimiento de Emociones de la Voz (REV) es un sistema de identificación de emociones a través de un locutor humano. Este proceso permite reconocer el impulso emocional causado por un estímulo temporal llamado emoción interacción persona-computadora, a diferencia del estado emocional, la voz emotiva suele durar pocos minutos.

Los diferentes estados emocionales de un hablante producen cambios fisiológicos en el aparato fonador, lo que se ve reflejado en la variación de dichas características. Las técnicas empleadas en el análisis de la señal de voz se pueden dividir en dos categorías: Transformadas Tiempo - Frecuencia y Análisis Paramétrico. La primera de estas categorías hace referencia a la representación de la señal en espacios conjuntos del tiempo y la frecuencia, permitiendo conocer la ubicación temporal del contenido espectral, esta técnica es efectiva en el tratamiento de señales no estacionarias como es la señal de voz. El análisis paramétrico busca estimar un modelo matemático que de forma aproximada represente el sistema de producción vocal [Duque and Morales, 2007].

3.1.1. Voz

La voz es el sonido producido voluntariamente por el aparato fonatorio humano. El aparato fonador es el conjunto de órganos del cuerpo humano encargado de generar y ampliar el sonido que se produce al hablar. Éste está formado por los pulmones como fuente de energía en la forma de un flujo de aire, la laringe, que contiene las cuerdas vocales, la faringe, las cavidades orales (o bucales) y nasal, además de una serie de elementos articulatorios: los labios, los dientes, el alvéolo, el paladar, el velo del paladar y la lengua [Miyara, 1999].

El sistema fonador [Sánchez et al., 2007] se puede dividir en tres bloques:

- *Sistema de generación*: Los músculos abdominales y torácicos aumentan la presión en los pulmones produciendo un exceso en la corriente de aire, ésta sale por los

bronquios y la tráquea hasta llegar a la laringe donde es excitado el sistema de vibración.

- *Sistema de vibración* : Está conformado básicamente por las cuerdas vocales, las cuales se dividen en dos pares: superiores e inferiores, de estas, sólo las últimas participan en la producción de voz. En el caso de la respiración las cuerdas se abren y se recogen a los lados permitiendo el libre paso del aire, si por el contrario se encuentran juntas y tensas el aire choca haciendo que se produzcan los diferentes sonidos.

- *Sistema resonante*: Lo componen tres cavidades articulatorias: cavidad faríngea, cavidad oral y cavidad nasal. Los sonidos producidos por el sistema de vibración se desplazan desde las cuerdas vocales hasta los orificios nasales y la boca, la articulación de las cavidades modifica y amplifica los sonidos que finalmente son expulsados al exterior.

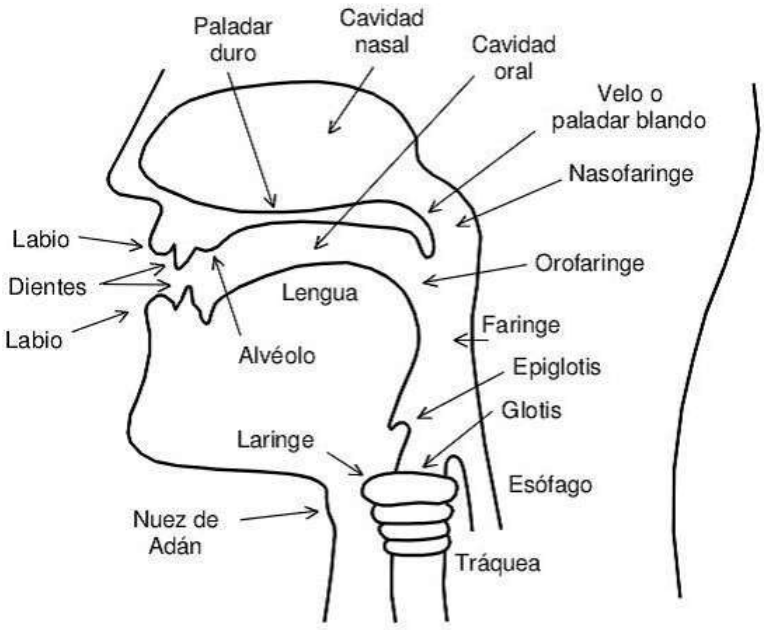

Figura 3.1: Sistema fonador [Rowden, 1992]

La voz está compuesta por una secuencia de sonidos, estos sonidos y la transición entre ellos sirven como una representación simbólica de la información [Rabiner and Juang, 1993].

Los sistemas de reconocimiento de voz se pueden clasificar de acuerdo al tipo de voz, el tamaño de vocabulario y la dependencia del locutor. Esto determina qué algoritmo se tiene que utilizar [Vrinda and Shekhar, 2013].

De acuerdo al tipo de voz existen dos clases:

- *Voz continua* : Permite al usuario hablar casi naturalmente, sin necesidad de establecer silencios entre las palabras, mientras la computadora determina el contenido.
- *Voz discreta* : Consta de palabras aisladas que están separadas por silencios. La ventaja es que el límite de la palabra se puede ajustar.

Dependiendo del tipo de expresiones que pueda reconocer el sistema se clasifican en:

- *Palabras aisladas*: Donde el reconocedor generalmente requiere que cada expresión tenga una pausa entre sí. El reconocedor acepta una sola palabra a la vez. Estos sistemas cuentan con estados "Listen/Not listen"[Hernández, 2016].
- *Palabras conectadas* : Se utilizan como unidades de reconocimiento pero pueden ser emitidas secuencialmente con pausas entre ellas.

3.1.2. La Naturaleza del Sonido

Las dos propiedades básicas de todo sonido son la frecuencia e intensidad. La frecuencia es simplemente la velocidad a la que se producen las vibraciones. Se mide en Hertz (Hz) o ciclos por segundo. Un ciclo es una vibración completa. La cantidad de hertz es la frecuencia; entre más alta sea, mayor será el tono.

No es posible escuchar en todas las frecuencias posibles. Muy pocas personas pueden oír menos de 20 Hz o más de 20 KHz. De hecho la nota más baja en un piano tiene una frecuencia de 27 Hz y la más alta un poco más de 4 KHz. Una estación de radio FM (Frecuencia Modulada) transmite notas de hasta 15 KHz. A la intensidad del sonido se le denomina amplitud (está asociado con el volumen del sonido). Esta intensidad depende de la fuerza de las vibraciones que producen el sonido. Por ejemplo, una cuerda de piano vibra levemente cuando la tecla se pulsa con suavidad. La cuerda oscila de arriba hacia abajo en un arco angosto y el tono que emite es suave. Sin embargo, si la tecla se pulsa con fuerza, la cuerda oscila en un arco más amplio. El volumen de los sonidos se mide en decibeles (db). El susurro de las hojas secas está clasificado en 20 db, el ruido promedio en la calle es 70 db y un trueno cercano en 120 db [Tintaya, 2005].

3.1.3. El Proceso Digital de Señales

El procesado digital de señales (Digital Signal Procesing, por sus siglas en ingles DSP) son técnicas matemáticas que se utilizan para medir magnitudes físicas que contienen información sobre un fenómeno natural. Las más habituales en aplicaciones informáticas son:

- Temperatura
- Presión
- Desplazamiento
- Sonido/Voz
- Imagen

La digitalización de la voz es la acción de convertir en digital información analógica (convertir cualquier señal de entrada continua en una serie de valores numéricos).

Las señales digitales, en contraste con las señales analógicas, no varían en forma continua, sino que cambian en pasos o en incrementos discretos. Las señales en tiempo discreto son aquellas que se representan matemáticamente como una secuencia de números. Además del carácter de estar definidas en tiempo discreto, la amplitud de la señal puede ser también discreta.

3.1.4. Muestreo

El muestreo es el proceso de convertir las ondas originales de sonido analógico en señales digitales que puedan almacenarse y reproducirse después se toman instantáneas de los sonidos analógicos y se almacenan.

La velocidad con que el ADC (Analog to Digital Converters) tome las muestras se llamará frecuencia de muestreo (Sample Rate, por sus siglas en ingles SR), y estará expresada en Hertz o Kilo Herzt (Hz, Khz). 1 Hz será 1 muestra por segundo y 10 Khz son 10 000 muestras por segundo. A cada una de esas muestras le asignará un valor correspondiente a la amplitud de ese instante en la señal original (Cuantización).

En la mayoría de los casos, las señales en tiempo discreto surgen de tomar muestras de una señal analógica. De esta forma, el valor numérico del n-ésimo número de la secuencia es igual al valor de la señal analógica $X_a(t)$,en el instante temporal nT_s, es decir:

$$\hat{x}(n) = x_a(nT_s, -\infty < n < \infty) \tag{3.1}$$

La cantidad T_s se denomina perodio de muestro y su inversa es la frecuencia de muestreo F_s.

Figura 3.2: Frecuencia de Muestreo

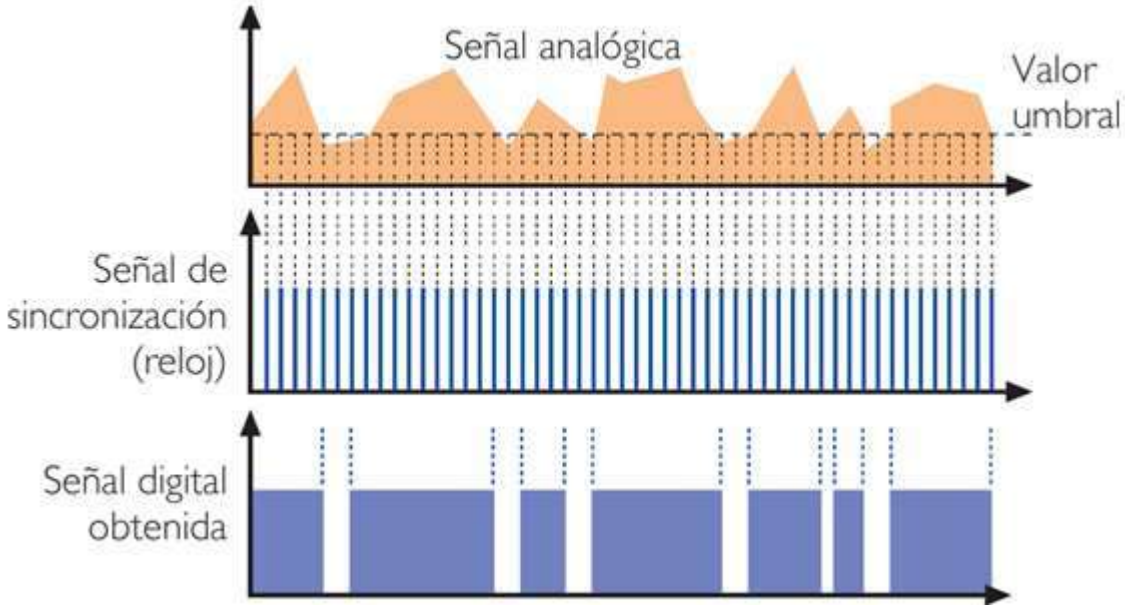

Figura 3.3: Digitalización por muestreo de una señal analógica

El muestreo asigna un valor numérico a la señal en unidades discretas de tiempo constante dependiendo de la frecuencia Nyquist que específica la frecuencia máxima a la que una señal puede reproducirse completamente [Pérez Badillo et al., 2013]. El teorema de Nyquist garantiza que, para poder reconstruir una señal a partir de sus muestras, se debe utilizar una frecuencia $N_s \geq 2_f N$, o sea al menos el doble de f_N . Siendo f_N la componente de más alta frecuencia de la señal.

En la etapa de muestreo se obtiene una señal en tiempo discreto cuyas amplitudes $\hat{x}$(n) son valores continuos. Para digitalizar la señal resta discretizar esos valores (cuantizarlos).

3.1.5. Cuantización

La cuantización consiste en que cada muestra se representa con un valor digital limitando el rango de valores discretos correspondiente a la original. El propósito del cuantizador es transformar la muestra de entrada $\hat{x}$(n) en un valor x(n) de un conjunto finito de valores preestablecidos. Esto se realiza redondeando los valores de las muestras hasta el nivel de cuantización más próximo.

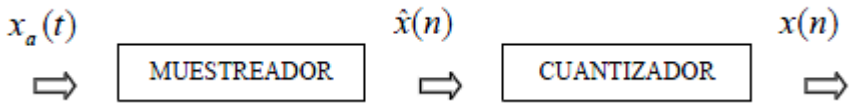

Figura 3.4: Representación conceptual de la digitalización de una señal analógica.

La precisión de los datos dependerá del número de bits con que se codifiquen los niveles de cuantización. Por tanto, se introduce un ruido de cuantización que se asume como ruido blanco.

3.2. Reconocimiento de emociones

La definición del término emoción es la base para cualquier tipo de investigación en esta área. Una definición común permite comparar resultados entre diferentes grupos de investigación y evitar malentendidos. La manera en que las emociones son definidas también determina el tipo de fenómenos estudiados en la investigación sobre emociones. Según Scherer [Scherer, 2000] las emociones son definidas como:

Episodios de cambios coordinados en varios componentes (incluyendo al menos activación neuropsicológica, expresión motriz, y sentimientos subjetivos, pero posiblemente también tendencias a la acción y procesos cognitivos) en respuesta a eventos externos o internos de mayor significancia para el organismo.

3.2.1. Análisis de emociones

Emoción y estado emocional son conceptos diferentes: mientras que las emociones surgen repentinamente en respuesta a un determinado estímulo y duran unos segundos o minutos, los estados de ánimo son más ambiguos en su naturaleza, perdurando durante horas o días. Las emociones pueden ser consideradas más claramente como algo cambiante y los estados de ánimo son más estables. Aunque el principio de una emoción puede ser fácilmente distinguible de un estado de ánimo, es imposible definir cuando una emoción se convierte en un estado de ánimo; posiblemente por esta razón, el concepto de emoción es usado como un término general que incluye al del estado de ánimo [Ortego Resa et al., 2009]. Las emociones pueden ser vistas por su valor adaptativo con las tareas fundamentales de la vida. Cada emoción tiene características únicas y otras que son comunes por ser producto de nuestra evolución [Ekman, 1992]. Las emociones básicas son: enojo, miedo, tristeza, alegría disgusto y sorpresa. La voz neutral [Kim et al., 2007] se puede percibir de una forma uniforme, calmada, con un tono más o menos idéntico, sin alteraciones o interrupciones, posteriormente la emoción de enojado se puede apreciar una voz determinante, fuerte, irritable, agresiva y severa. Para el estado de la felicidad, se le puede considerar como una voz cantada, llena de alegría, de alguna forma como si el locutor tuviera una sonrisa en la cara; la forma de expresarse con la emoción del miedo denota una voz cambiante, interrumpida, un tono casi chillón, voz ansiosa, con susurros. Por último, el estado emocional de tristeza puede ser percibido como monótono, depresivo, lento, melancólico y lento [Solís, 2011].

El habla neutra suele caracterizarse por un tono con un rango de variación estrecho y unas transiciones de F_0 suaves, además de una velocidad de locución alta. A continuación plantearemos una de las clasificaciones de las emociones primarias:

- *Enfado*: El enfado se define como "la impresión desagradable y molesta que se produce en el ánimo". El enfado se caracteriza por un tono medio alto (229 Hz), un amplio rango de tono y una velocidad de locución rápida (190 palabras por minuto), con un 32 % de pausas.

- *Alegría*: Se manifiesta en un incremento en el tono medio y en su rango, así como un incremento en la velocidad de locución y en la intensidad.

- *Tristeza*: El habla triste exhibe un tono medio más bajo que el normal, un estrecho rango y una velocidad de locución lenta.

- *Miedo*: Comparando el tono medio con los otras cuatros emociones primarias estudiadas, se observó el tono medio más elevado (254 Hz), el rango mayor, un gran número de cambios en la curva del tono y una velocidad de locución rápida (202 palabras por minuto).

- *Disgusto/odio*: Se caracteriza por un tono medio bajo, un rango amplio y la velocidad de locución más baja, con grandes pausas.

3.2.2. Análisis de Señales

La capacidad auditiva del ser humano varía en un rango de frecuencias de 20 Hz a 20,000 Hz [Herrera, 2006]. Los sonidos emitidos al hablar se encuentran de 100 Hz a 15,000Hz en mujeres y en hombres de 400Hz a 15,000 Hz [Hernández, 2016].

El enfado se caracteriza por un tono medio alto (229 Hz), un amplio rango de tono y una velocidad de locución rápida (190 palabras por minuto), con un 32 % de pausas.

La alegría manifiesta en un incremento en el tono medio y en su rango, así como un incremento en la velocidad de locución y en la intensidad.

El habla triste exhibe un tono medio más bajo que el normal, un estrecho rango y una velocidad de locución lenta.

El miedo se distingue comparando el tono medio con los otros cuatros emociones primarias estudiadas, se observa el tono medio más elevado (254 Hz), el rango mayor, un gran número de cambios en la curva del tono y una velocidad de locución rápida (202 palabras por minuto).

En la figura 3.5 se puede observar las señales de voz que expresa en la palabra en serbio "da", que en castellano se puede traducir como "si"; dichas señales fueron expresadas en 5 diferentes emociones y cabe hacer notar las diferencias en duraciones de tiempo, así como las diferencias en amplitud [Kim et al., 2007].

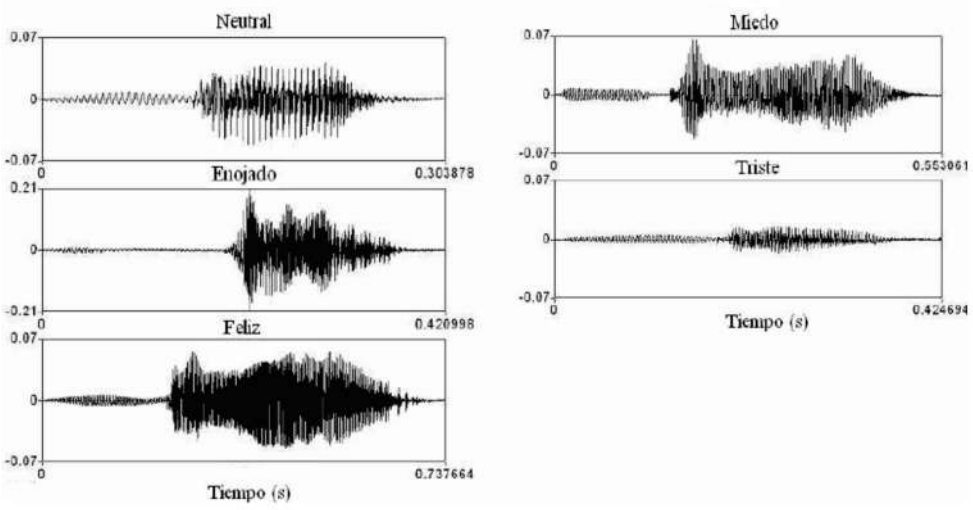

Figura 3.5: Palabra *da* (en serbio, se traduce como *si* en castellano) [Solís, 2011].

3.2.3. Análisis de las características Acústicas

Se han presentado muchos enfoques para reconocer estados afectivos basados en características específicas del habla. Para este propósito se han utilizado características a corto plazo (formantes, ancho de banda de formantes, frecuencia de tono/fundamental y energía de registro) y características a largo plazo (media de tono, desviaciones estándar de tono, envolventes temporales de tono y energía).

Las características a corto plazo reflejan las características del habla local en una ventana de corto tiempo, mientras que las características a largo plazo reflejan las características de la voz sobre un enunciado completo [Li and Zhao, 1998]. El tono (Pitch), frecuencia fundamental *(F_0)*, la intensidad de la señal de voz (energía) y la tasa de habla se han identificado como importantes indicadores de la emoción en la voz [Ververidis and Kotropoulos, 2006].

La información acústica describe sonidos, lenguaje y la expresión emotiva; estos elementos incluyen fonemas, la forma de articulación y en que estado de ánimo se pronuncie. La información es acústica cuando la extracción se hace únicamente sobre la señal de voz, la cual describe los sonidos básicos del lenguaje y trata de explicar cómo se realizan acústicamente en una expresión hablada. De acuerdo al tipo de información las características acústicas suelen agruparse en:

- *Espectrales*: Describen las propiedades de una señal en el dominio de la frecuencia mediante harmónicos y formantes.
- *Calidad de Voz*: Definen estilos al hablar como neutral, susurrante, jadeante, estrepitoso resonante, sonoro, ruidoso.
- *Prosódicas*: Describen fenómenos suprasegmentales como entonación, volumen, velocidad, duración, pausas y ritmo.

3.3. Análisis de los Parámetros de Voz

Los efectos fisiológicos en el habla (acústicos, prosódicos y léxicos), se utilizan para expresar emociones, dentro de los cuales se consideran los más importantes: pitch, duración, calidad de voz y forma del pulso glotal y tracto vocal. Estudios previos muestran que es difícil encontrar características de voz especificas que puedan usarse como indicadores confiables de la emoción presente en el habla [Álvarez et al., 2006]. La voz no es otra cosa que un sonido y como tal, se caracteriza por una serie de elementos.

3.3.1. Tono

El tono (pitch en inglés), se podría definir como la impresión perceptiva que nos produce la frecuencia fundamental *(F_0)* de la onda sonora. Es, por tanto, una cualidad subjetiva dependiente de una propiedad física[Monzo Sánchez et al., 2010]. Está relacionado con la cantidad de vibraciones que posee una onda de sonido. A mayor número más aguda será la voz. Estas vibraciones se producen en el ser humano en la laringe y se miden en Hertzios o Hertz (Hz). Las voces masculinas oscilan entre los 75 Hz y los 200 Hz. Las femeninas entre los 150 Hz y los 300 Hz. El pitch también conocido como melodía [Garrido Almiñana, 1991] tiene las siguientes propiedades:

- *Frecuencia fundamental (F_0)* : Se define como el ciclo periódico de la señal de voz, siendo el resultado de la vibración de los pliegues vocales. Su medida habitual es el hercio (Hz), que da una medida de los ciclos por segundo.

- *Curva de F_0 o melódica* : Se trata de la secuencia de valores de F_0 para una elocución, y se relaciona con la percepción de la entonación del habla.

- *Jitter*: Parámetro que caracteriza la perturbación de F_0 debida a fluctuaciones en los tiempos de apertura y de cierre de los pliegues vocales de un ciclo al siguiente.

Figura 3.6: Relación tonos-emociones[Duque and Morales, 2007].

3.3.2. Volumen

El volumen o intensidad es aire que al salir de los pulmones golpea la glotis y produce vibraciones.Se mide en decibelios (dB) y para tener una referencia, una conversación normal ronda entre los 50 dB. Tiene efectos en el oyente porque transmite emociones. Un volumen de voz alto se asocia a la agresividad, nerviosismo, tensión y lejanía. Al contrario, un volumen bajo puede sugerir depresión, cansancio y proximidad. Las propiedades relacionadas con el volumen son los siguientes:

- *Intensidad* : Medida de la energía de la onda acústica. Habitualmente se utiliza una transformación logarítmica de la amplitud de la señal, llamada decibelio (dB), que representa mejor la percepción humana del sonido.

- *Shimmer*: Parámetro que caracteriza la perturbación en la intensidad debida a fluctuaciones en la amplitud de un ciclo al siguiente.

3.3.3. Duración

La duración es la componente de la prosodia descrita por la velocidad del habla y la situación de los acentos, cuyos efectos son el ritmo y la velocidad.

El ritmo en el habla deriva de la situación de los acentos y de la combinación de las duraciones de las pausas y de los fonemas. Las propiedades relacionadas con los aspectos temporales del habla son:

- *Velocidad del habla*: Se mide a partir de la duración de los segmentos del habla o como el número de unidades lingüísticas por unidad temporal (p.ej. sílabas por segundo).
- *Pausas*: El número y la duración de los silencios en la señal de voz es un parámetro del que habitualmente se realiza su medida.

3.3.4. Comparativo de Características del Habla

La tabla 3.1 se presenta un resumen de las relaciones entre las emociones y los parámetros del discurso. Como se puede observar, únicamente aparecen cinco emociones. Estas corresponden con las emociones primarias o básicas.

	Felicidad	**Ira**	**Disgusto**	**Miedo**	**Tristeza**
Velocidad del habla	Ligeramente acelerada, con incremento	Ligeramente acelerada	Lenta	Muy Acelerada	Pausada
F0	Incremento de la media, variabilidad	Incremento de la media mediana, variabilidad	———	Incremento en la F0 media, perturbación, variabilidad del movimiento de F0	Debajo de la F0 media normal
Articulación	Normal	Tensa	Normal	Precisa	Arrastrada
Intensidad	Alta (en Incremento)	Alta	Baja	Normal	Baja
F0 promedia	Alta	Alta	Baja	Alta	Baja
Espectro	Incremento de la energía de alta frecuencia	Elevado en el punto medio	———	Aumento de la energía de alta frecuencia	Disminución de la energía de alta frecuencia
Tono Medio	Incremento	Bajo	Alto	Alto	Muy bajo
Otros	Distribución irregular de acentos	Habla cortada	———	Irregularidad en la sonorización	Ritmo con pausas irregulares

Tabla 3.1: Comparativo de emociones [Ortego Resa et al., 2009][Cowie et al., 2001].

Es conocido que existe una relación entre la información prosódica y la expresión de emociones en el habla; rasgos como la intensidad, la curvatura de frecuencia fundamental y la velocidad de locución son características importantes den la discriminación de emociones en la voz [Nwe et al., 2003] [Montero Martínez, 2003].

Hasrul (2012), agrupa su trabajo en 13 características que han sido utilizadas para la detección de emociones en la voz. Estos parámetros se muestran en la tabla 3.2.

Características Utilizadas	**Descripción**
Ancho de banda	Este rango se mide en Hercios (Hz)
Áreas del tracto vocal	Numero de armónicos ocasionados por el flujo de aire no lineal en el tracto vocal que produce la señal de voz.
Características espectrales	Contenido energético de bandas de frecuencia divididas por la longitud de muestra
Detección de la Actividad del Habla	Esta propiedad se define como el perfil rítmico del habla
Duración	Diferencia entre el instante de inicio y final de una secuencia hablada obteniendo una taza de duración de sentencias de tipo emocional y neutras
Energía	Es el valor de la magnitud física que expresa la mayor o menor amplitud de las ondas sonoras.
Formantes	Son frecuencias reforzadas por la resonancia
Intensidad	Se mide en Decibelios (dB)
LPCs (Linear Prediction Coefficients)	Conjunto de formulaciones esenciales equivalentes para modelar una forma de onda dada
MFCCs (Mel Frequency Cepstrum Coefficients)	Técnica de fraccionar la señal inicial en un conjunto discreto de bandas espectrales que contiene información analoga
Pitch	Se representa como F_0 (Frecuencia Fundamental)
Tasa de cruce por ceros	Representa cuantas veces la señal cambia de signo pasando por el eje de las abscisas
Velocidad del habla (speaking rate)	La proporción de unidades segméntales, silabas y pausas por unidad de tiempo producidas por un locutor

Tabla 3.2: Características usadas en el reconocimiento de emociones en el Habla [Hasrul et al., 2012].

3.4. Dimensiones Emocionales

Las dimensiones emocionales son una representación simplificada de las propiedades esenciales de las emociones. Evaluación (positiva / negativa) y activación (activa / pasiva) son las dimensiones más importantes, en algunas ocasiones se complementan con la dimensión poder (dominante / sumiso)[Wundt, 1896].

A continuación, se presentan las tres dimensiones [Monzo Sánchez et al., 2010] más utilizadas junto con diferentes términos para referirse a ellas:

- *Evaluación / agrado / valoración*: Corresponde al eje "Positivo-Negativo", clasificando las emociones según lo placentero o desagradable de estas (p. ej. Desde la alegría hasta el enfado).
- *Activación / actividad*: Corresponde a la escala ¨ Activo - Pasivo ¨, indicando la presencia o ausencia de energía o tensión (p. ej. desde estar furioso a estar aburrido).
- *Potencia / fuerza*: Corresponde a la escala "Dominante-Sumiso", distinguiendo emociones iniciadas por el sujeto de aquellas causadas por el entorno (p. ej. desde el desprecio al temor o a la sorpresa).

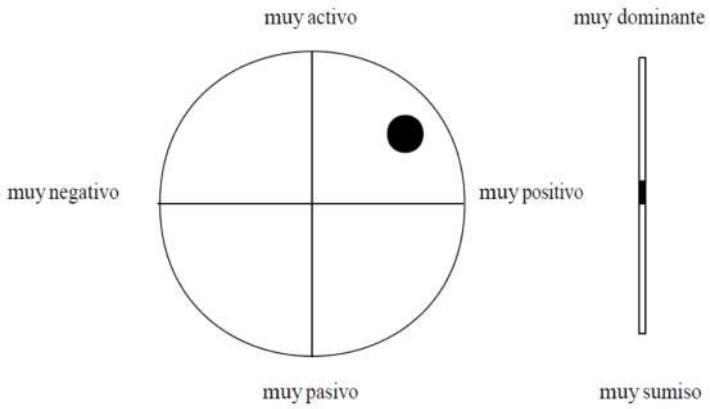

Figura 3.7: Representación de las emociones en el espacio semántico [Duque and Morales, 2007].

Las emociones son descritas en términos de valencia y activación [Steidl, 2009]. La valencia, también llamada placer describe qué tan negativa o positiva es una emoción específica.

La activación, también llamada intensidad, describe la excitación interna de un individuo y va desde estar muy tranquila hasta estar muy activa.

La energía o dominación son las que describen el grado de control del individuo sobre la situación, en otras palabras, qué tan fuerte o débil se muestra el individuo. La dominación ayuda a distinguir entre emociones como miedo y enojo ya que ambas tienen valencia y activación similar.

Aquellas emociones con una actividad similar, como puede ser el caso de la alegría o del enfado, se confunden más entre sí que emociones con valoración o fuerza parecida [Monzo Sánchez et al., 2010].

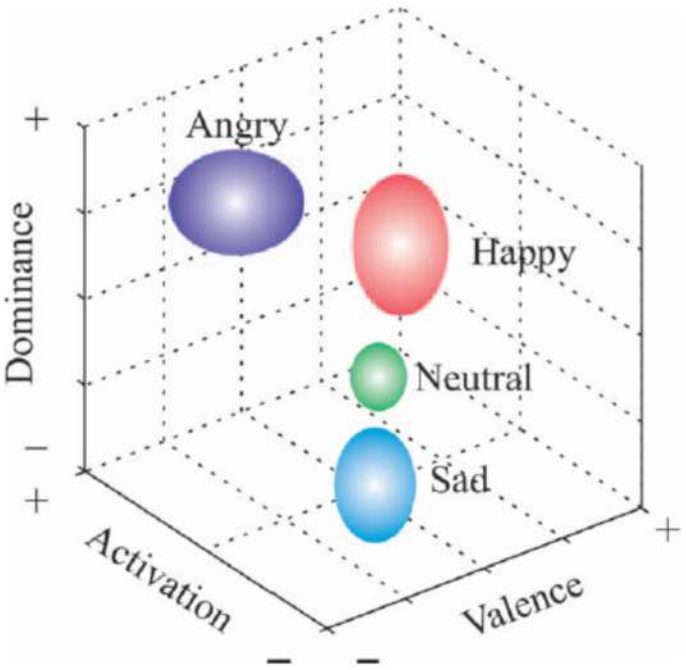

Figura 3.8: Modelo Tridimensional Continuo de las Emociones. [Espinosa et al., 2010].

3.5. El Lenguaje Matlab

MATLAB es el nombre abreviado de "MATrix LABoratory". Es un lenguaje de alto nivel y de ambiente interactivo que permite realizar tareas intensas y con una mayor velocidad que los lenguajes de programación comúnmente usados.

MATLAB es un lenguaje de programación técnico-científico que básicamente trabaja con variables vectoriales y matriciales. Es fácil de utilizar debido a que contiene varias cajas de herramientas con funciones incorporadas (toolbox de procesamiento de señales, teoría de control, wavelets y matemática simbólica) [Tintaya, 2005].

MATLAB se especializa en cálculos numéricos con vectores y matrices, como casos particulares puede trabajar también con otras estructuras de información. Aunque cada objeto es considerado como un arreglo. El lenguaje está construido por código llamado M-code que puede ser fácilmente ejecutado en la ventana de comandos. Con lo cual se pueden crear funciones, etc. Pero la razón principal para la elección de este lenguaje de programación son las herramientas que proporciona para el procesamiento de señales, y el conjunto de funciones para el procesamiento digital. Además, para crear entornos gráficos se puede utilizar el GUIDE de MATLAB, que provee herramientas para crear GUIs, "Graphical User Interface", con lo cual se puede crear la forma del entorno gráfico, así como asociar funciones a los elementos del GUI. MATLAB también incluye funciones para manipular archivos.

Es apropiado para el caso de muchas señales de interés, donde la frecuencia de muestreo sea menor que 44.1 KHz.

Capítulo 4

Metodología de Desarrollo

En este capitulo se presenta la metodología del desarrollo del proyecto, la cual se divide en 4 componentes: *el estudio de parámetros acústicos y lingüísticos* que contienen características de los estados emocionales,*el diseño del modulo de grabación*, *la captura de frases para tener el corpus emocional* y las *pruebas con el algoritmo de reconocimiento de estados emocionales primarios* (véase figura 4.1).

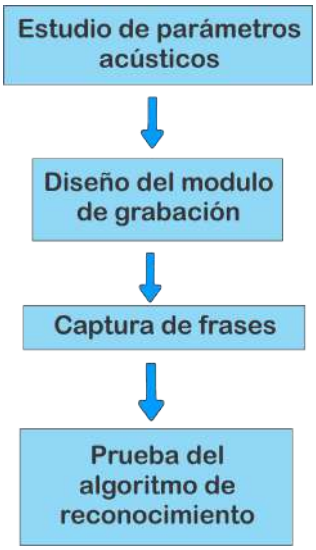

Figura 4.1: Etapas del proyecto.

4.1. Estudio de parámetros acústicos

Los parámetros acústicos son medidas que se emplean para el análisis acústico de la voz que deben observarse en toda exploración acústica, e incluyen la frecuencia fundamental (F_0), la intensidad, las perturbaciones de amplitud (shimmer), perturbaciones de frecuencia (jitter) y la expresión del ruido espectral (calculada mediante la relación armónico/ ruido), de modo que es posible evaluar hasta los más pequeños cambios en la masa y tensión, así como el carácter bioquímico de las cuerdas vocales [Adrián Torres and Casado Morente, 2002].

Dentro de las actividades a realizar en esta sección son las siguientes:

1. Búsqueda de Información: Identificar grupos de características usadas hasta el momento mediante la revisión del estado del arte.

   a) Hacer una recopilación de las características extraídas de la señal de voz que hayan sido propuestas en los trabajos en esta área publicados hasta el momento.

   b) Buscar una relación de los métodos de clasificación empleados con cada conjunto de características.

   c) Realizar una lista de las bases de datos utilizadas en trabajos relacionados al proyecto poniendo especial atención en bases de datos de emociones primarias.

2. Estudiar métricas de calidad de voz y articulación usadas en diferentes áreas y comprobar la viabilidad de aplicación.

   a) Realizar un estudio sobre estándares y metodologías de medición de calidad y otros aspectos en la de voz en áreas distintas (educativas, medicas e inteligencia artificial).

   b) Adoptar características acústicas para la clasificación de emociones basadas en los diferentes casos de estudio.

3. Estudiar las características específicas para reconocer los estados afectivos.

a) Estudiar las características espectrales que describen las propiedades de una señal en el dominio de la frecuencia mediante armónicos y formantes.

b) Estudiar las características de calidad de voz que definen estilos al hablar como neutral, susurrante, jadeante, estrepitoso resonante, sonoro y ruidoso.

c) Estudiar las características prosódicas que describen fenómenos suprasegmentales como entonación, volumen, velocidad, duración, pausas y ritmo.

4.2. Modulo de grabación

Un sistema de adquisición de datos mediante una tarjeta de sonido de una PC, es un conversor análogo digital.

Es necesario llevar acabo un proceso de acondicionamiento para el aprovechamiento total de la señal capturada y la calidad establecida.

Para el diseño de la grabadora de audio se realizan los siguientes pasos:

1. Crear el esqueleto para una nueva aplicación que utilice un formulario de tipo Form como ventana principal.
2. Añadir los componentes necesarios al formulario.
3. Definir propiedades de los componentes.
4. Escribir el código para cada uno de los objetos.
5. Guardar la aplicación.
6. Crear un fichero ejecutable.

4.2.1. Requerimientos del sistema de grabación

La interfaz gráfica de usuario (GUI) para el sistema de grabación tiene las siguientes características:

1. Captura de señal de audio: El sistema debe permitir la captura de audio a una frecuencia de 44100 Hz, con una tasa de bits de 16 kbps (kilobits por segundo),un canal mono y en formato WAV.

2. Capacidad para guardar archivos de audio: El sistema permite guardar la voz del locutor en tiempo real en una carpeta llamada corpus.

3. Capacidad detener la grabación de audio: El sistema debe tener la opción de detener una grabación de voz en tiempo real.

4. Capacidad de eliminar: El sistema debe eliminar archivos de audio.

5. Capacidad de reproducir: El sistema debe permitir la reproducción de formatos de audio.

6. Capacidad de detener reproducción: El sistema debe permitir detener una reproducción en curso.

7. Capacidad de cambiar la ruta: El sistema debe permitir cambiar la ruta para guardar el formato de audio en otra dirección.

4.2.2. Objetos que forman la interfaz

La grabadora de audio incluye los siguientes objetos:

- Un formulario que permita implementar nuestra interfaz.

- 3 etiquetas para el nombre de archivo el formato y la ruta.

- 3 cajas de texto distribuidos de la forma siguiente:
  - 1 caja de texto, una para dar información al usuario.
  - 2 cajas de texto para escribir el nombre del archivo y la ruta.
- 5 botones de órdenes distribuidos de la forma siguiente:
  - Seleccionar la ruta.
  - Borrar archivo de audio.
  - Grabar audio.
  - Detener grabación de audio.
  - Detener reproducción de audio.
- Una caja de lista para visualizar los archivos de audio.
- Una etiqueta para el fondo de pantalla.
- Una barra de progreso que se va completando para indicar el progreso de una operación.

4.2.3. Eventos

Haciendo clic sobre las botones visualizaremos procesos en conjunto con la barra de progreso, así como mensajes de alerta. La entrada de voz se maneja con los botones mediante un clic sobre él. A dicha acción se le denomina Evento Clic.

4.2.4. Descripciones de controles

Una vez que se conocen los objetos y los eventos, se procede a diseñar la interfaz para la aplicación denominada Interface G. La tabla 4.1 muestra la descripción de los controles a utilizar.

Objeto	Propiedad	Valor
Formulario Frame	(Name) resizable preferredSize	Frame False [810, 550]
Etiqueta Fodo	(Name) icon	TemaFondo Fondo.JPG
Etiqueta Archivo	(Name) Font Text foreground	Enombre Arial 18 Bold Archivo: [255,255,255]
Etiqueta Ruta	(Name) Font Text foreground	Eruta Arial 18 Bold Ruta: [255,255,255]
Etiqueta WAV	(Name) Font Text foreground	Eruta Arial 18 Bold WAV: [255,255,255]
Barra de proceso bpProgreso	(Name) Value foreground	barProgreso 0 [102,153,255]
Boton Ruta	(Name) Font Text foreground	btRuta Arial 18 Bold Ruta [255,255,255]
Botón Borrar	(Name) Font Text foreground	btBorrar Arial 18 Bold Borrar [255,255,255]
Botón Detener	Name Font Text foreground enabled	btDetener Arial 18 Bold Detener [255,255,255] False
Botón Grabar	Name Font Text foreground	btGrabar Arial 18 Bold Grabar [102,204,0]
Botón DetenerG	Name Font Text foreground	btDetenerG Arial 18 Bold DetenrG [255,0,0]
Cuadro de texto Nombre	Name Font Text foreground	txtNombre Dialog 12 Plain Vacio [187,187,187]
txtRuta Ruta	Name Font Text foreground editable	txtRuta Dialog 12 Plain Vacio [187,187,187] False
Cuadro de texto Información	Name Font Text foreground editable	txtInforma Dialog 12 Plain Vacio [187,187,187] False
Jlist Lista	Name Font toolTipText foreground selectionMode	txtLista Dialog 12 Plain Vacio [187,187,187] MULTIPLE_INTERVAL

Tabla 4.1: Descripción de controles.

4.3. Corpus emocional

Con la intención de determinar si los parámetros acústicos y la velocidad de habla funcionan como elementos caracterizadores de los distintos tipos de emociones, se creó el corpus emocional recogido por alumnos del ITSM. Este corpus esta constituido por una serie de grabaciones en las cuales se recogen emociones simuladas por los estudiantes.

Las emociones que fueron consideradas para el desarrollo de la investigación son: ira, felicidad, neutral, miedo y disgusto. Los textos de estímulo para las frases fueron concebidos en el contexto de situaciones de la vida cotidiana. Se diseñaron 16 enunciados para cada emoción. Estos enunciados fueron producidos por once hablantes: 2 hombres y 9 mujeres. Cada uno de los participantes grabarán el enunciado con todas emociones indicadas.

Una vez que se han seleccionado los participantes del corpus, lo más recomendable es disponer de una área sin ruidos para realizar la toma de datos. La cual es para crear un espacio que provoque un clima de confianza y procurar que no se produzcan interrupciones. Es importante que antes de que se comience con el proceso de adquisición de datos se explique a los participantes que no se puede comer durante la grabación, que no se puede levantar, por ejemplo, para ir a saludar a alguien que ha visto pasar, y que apague el teléfono móvil, pues creará interferencias en la señal en caso de que le llamen. Por tal motivo, es aconsejable informar del tiempo que va a durar la grabación para que el informante disponga del tiempo suficiente para las 16 frases, aproximadamente se puede durar entre 10 a 30 minutos; no es recomendable que duren más de una hora para evitar agotar al locutor. Es recomendable que se eviten las horas de las comidas o las horas de preparación de las mismas. En todo caso no está demás tener un vaso de agua cerca para que los participantes beban un poco si se sienten cansados [Cano, 2018].

4.4. Etapas del reconocimiento de emociones en la voz

El proceso inicia cuando el sonido de un vocablo es capturado por la computadora realizando una grabación o lectura en formato WAV, acto seguido la señal es procesada y se normaliza, para después obtener de ella la Transformada Rápida de Fourier (por sus siglas en inglés: FFT), finalmente, mediante los coeficientes de error, el resultado es comparado en la base de datos que contiene los patrones a reconocer. Se abordará el uso de la técnica de correlación muestral como un método de reconocimiento de emociones en la voz (REV).

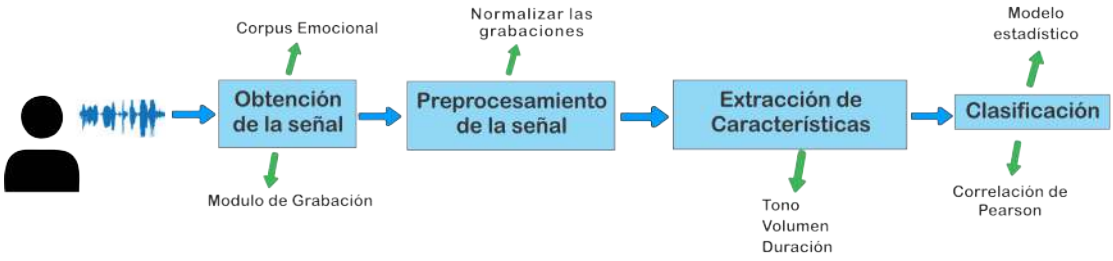

Figura 4.2: Diagrama a bloques de la estructura general del sistema propuesto.

4.4.1. Obtención de la señal

Hay dos factores importantes durante este proceso. Primero está la taza de muestreo, es decir, que tan seguido los valores de voltaje son grabados (Fs = 44100 Hz). Segundo, son los bits por segundo, es decir, que tan exactamente los valores son grabados (Tasa de bits = 16). Otro factor es el número de canales (mono o estéreo), pero para las aplicaciones de reconocimiento de voz un canal mono es suficiente. La mayoría de aplicaciones vienen con valores predeterminados, durante la codificación se deben de cambiar los parámetros para ver lo que mejor funciona en el algoritmo. Haciendo uso de un programa de escritorio, se graban audios con una frecuencia de muestreo de 44100 Hz y una tasa de audio de 16 bits. La grabación da como resultado un vector de miles de datos, de los que se discriminarán los datos mas significativos mediante un umbral de 0.1.

4.4.2. Preprocesamiento de la señal

El preprocesamiento consiste dar un tratamiento a la señal acústica para encontrar el conjunto óptimo de características que permitan realizar la clasificación optima de emociones. El proceso para la etapa de preprocesamiento es el siguiente:

- Guardar los dos audios en variables para su tratamiento.
- Obtener los parámetros acústicos como el pitch o la altura.
- Normalizar las grabaciones.
- Se cortan los primeros 60000 primeros valores de los audios almacenados en la base de datos con la misma frase a evaluar.

El pitch es la frecuencia a la que las cuerdas vocales vibran, también llamada frecuencia fundamental o F_0. Se considera que las características son una de las principales portadoras de la información sobre las emociones.

La normalización de la señal en amplitud se hace de forma manual, ubicando primeramente el máximo valor obtenido correspondiente a la amplitud y se obtiene el recíproco para poder multiplicar ese valor por toda la señal obtenida. El proceso consiste en ajustar todos los parámetros a una sola escala para que al momento de ser utilizados en el módulo de extracción de características no causen problemas de estabilidad.

La normalización es realizada mediante la Ecuación 4.1, donde X representa los datos a normalizar y μ, σ, su media y desviación típica respectivamente. Durante el entrenamiento del sistema, la media y desviación típica (μtrain y σtrain) son calculadas en el dominio de cada grupo de rasgos y para cada clase.

$$\hat{x}(n) = x_a(nT_s, -\infty < n < \infty) \tag{4.1}$$

En general se entiende que la normalización es la operación mediante el cual un conjunto de valores de una determinada magnitud es transformado en otros de tal manera que estos últimos pertenezcan a una escala predeterminada.

Es posible normalizar un conjunto de valores en el intervalo [0,1] aplicando para cada valor la operación que se muestra en la ecuación 4.2.

$$\nu_i = \frac{a_i - min}{max - min} \tag{4.2}$$

Donde a_i es el valor a transformar, min y max son el mínimo y el máximo del conjunto de valores y ν_i es el valor normalizado.

El algoritmo de función que normaliza los datos de un vector numérico que recibe como parámetro es el siguiente:

- Devuelve el valor absoluto máximo del vector a transformar.
- Devuelve el número de elementos del vector a transformar (Tamaño del vector = n)
- Devuelve un vector de ceros de n filas y 1 columna.
- Se hace un bucle donde el valor inicial de i es 1 y se va incrementando en 1 hasta que llegue a ser el valor de n.
- Se divide el vector en la posición i entre su valor máximo absoluto.

4.4.3. Extracción de Características

Este módulo consiste en agrupar las características acústicas espectrales, estas describen las propiedades de una señal en dominio de la frecuencia mediante armónicos y formantes, también se extrae información prosódica (volumen, velocidad, duración).El algoritmo para extraer características es la transformada rápida de furrier FFT el cual realiza el siguiente proceso:

- Se obtiene el valor absoluto de la transformada de Fourier de la grabación.
- Se multiplica el resultado por el conjugado del vector original.
- Se establece que solo se acepten las frecuencias arriba de 150 Hz.
- Se normaliza el vector utilizando la norma euclidiana.

La transformada rápida de Fourier tiene gran importancia en una gran variedad de aplicaciones, como ejemplo el procesamiento digital de señales. FFT es la abreviatura usual (de sus siglas en inglés Fast Fourier Transform), y es un eficiente algoritmo que permite calcular la transformada discreta de Fourier y su inversa dados vectores de longitud N. Sean X_0,.....,X_{n-1} números complejos la transformada se define:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\Pi k \frac{n}{N}} \tag{4.3}$$

$$k=0,1...,N\text{-}1 \text{ , } n=0,1...,N$$

La ecuación 4.3 es la fórmula para la transformada discreta de Fourier, misma que convierte las señales (como una grabación de sonido digital) muestreadas a el dominio de la frecuencia. Siendo este el motor matemático detrás de una gran parte de la tecnología que utiliza hoy en día.

Se obtienen la FFT de cada tramo, teniendo 5 vectores por cada emoción con el objetivo de generar una superficie en la que se pueda observar las frecuencias y su variación en el tiempo. Se promedian las FFT de cada tramo, para obtener un patrón de la frase pronunciada.

El proceso obtener la FFT de cada tramo de las grabaciones tiene el objetivo de generar una superficie en la que se pueda observar las frecuencias y su variación en el tiempo. Se promedian las FFT de cada tramo, para obtener un patrón de la frase pronunciada.

La norma euclidiana (también llamada magnitud del vector, longitud euclidiana, o $2-Norm$) de un vector v con los elementos de N es definido por la ecuación 4.4.

$$\|\nu\| = \sqrt{\sum_{k=1}^{N} |\nu_k|^2} \tag{4.4}$$

4.4.4. Clasificación

Para la clasificación de emociones se utilizó el coeficiente de correlación de Pearson, pensado para variables cuantitativas (escala mínima de intervalo), es un índice que mide el grado de covariación entre distintas variables relacionadas linealmente.

Adviértase que decimos "variables relacionadas linealmente". Esto significa que puede haber variables fuertemente relacionadas al aplicarse la correlación de Pearson.

Se define el coeficiente de correlación de Pearson como un índice que puede utilizarse para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas y continuas. El coeficiente de correlación de Pearson es un índice de fácil ejecución.

En primera instancia, sus valores absolutos oscilan entre 0 y 1. Si tenemos dos variables X e Y, entonces se define coeficiente de correlación de Pearson entre estas dos variables como $r_{x,y}$. La ecuación 4.5 muestra la expresión que permite calcular el coeficiente de correlación de Pearson.

$$px, y = \frac{\sigma xy}{\sigma x \sigma y} = \frac{E\left[(X - \mu x)\,(Y - \mu y)\right]}{\sigma x \sigma y} \tag{4.5}$$

Donde:

- σxy Es la covarianza de (X, Y)
- σx Es la desviación típica de la variable X
- σy Es la desviación típica de la variable Y

Capítulo 5

Pruebas y Resultados

En este capítulo se muestran las pruebas realizadas con el modelo de clasificación seleccionado y se explica el resultado obtenido mediante el software Octave GNU.

Ademas se muestra la interfaz de usuario para el modulo de grabación así como los ensayos y proceso de grabación por los participantes. Una vez terminado el proceso de grabación, se trabajó con las etapas del sistema de reconocimiento de emociones en la voz. Las transformadas de tiempo y frecuencia de los espectros de grabación de todas las emociones básicas tienen un efecto significativo sobre la mayoría de los parámetros acústicos analizados en este estudio.

5.1. Interfaz Gráfica de Usuario

El modulo que se muestra en la Figura 5.1 contiene la grabadora digital de voz para una PC. El Usuario tiene la opción de grabar una palabra, una frase o un discurso completo, definiendo un tiempo de grabación en segundos. Se puede Realizar una nueva grabación después de finalizar alguna otra, también tienen la opción de eliminar, reproducir y guardar en un archivo de audio con formato WAV.

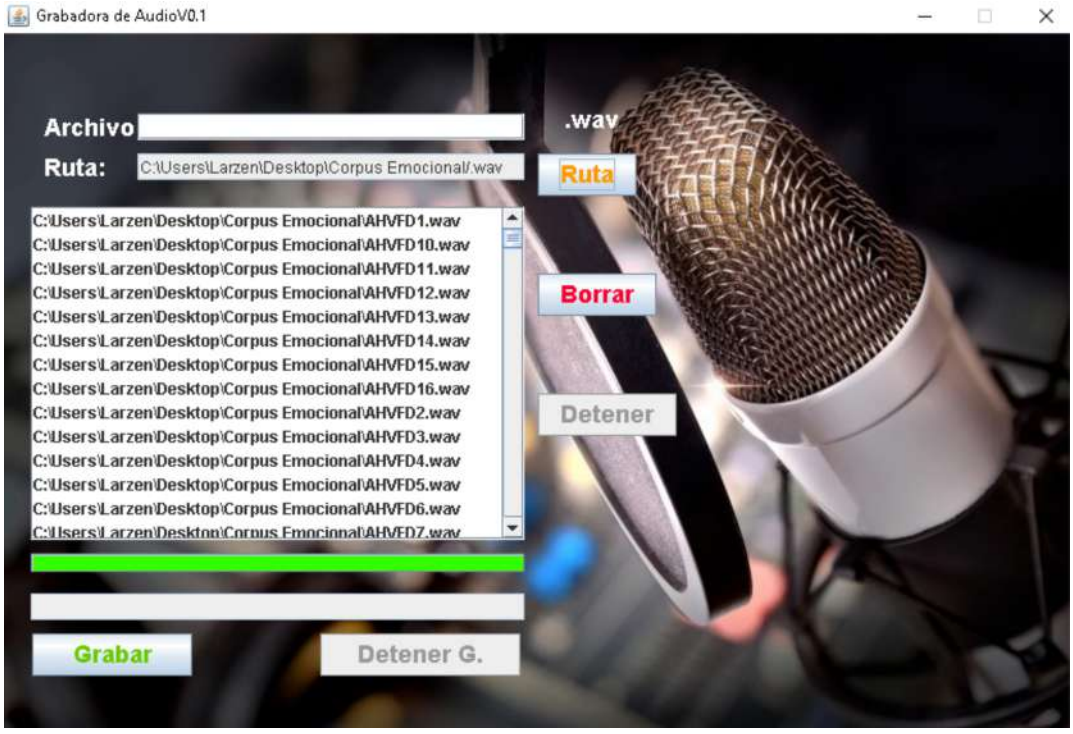

Figura 5.1: Modulo de Grabación.

El modulo de grabación lleva acabo el acondicionamiento de la señal de la amplitud y la frecuencia de la señal de voz.

Es necesario dicho acondicionamiento para el aprovechamiento total de la señal capturada, con una la calidad de voz establecida y con un ancho de banda de 4 Kh y así obtener un acondicionamiento optimo para un mensaje de voz claro.

Una vez realizada una grabación y guardarla en una ruta específica se genera un archivo en formato WAV en la carpeta Corpus emocional donde se encuentra nuestra base de datos con los estados emocionales. Haciendo uso del modulo de grabación montado en una PC de escritorio, se graban audios con una frecuencia de muestreo de 44100 Hz y una tasa de audio de 16 bits. Se usa un canal (Mono) que da como resultado un vector de miles de datos, de los cuales se discriminarán los datos significativos.

5.1.1. Pruebas de funcionamiento

Para el uso de la interfaz se tomaron al azar alumnos del ITSM a ellos se les proporciono una capacitación del uso de la interfaz y ensayos simulados de las frases emotivas a grabar (véase la figura 5.2).

En la figura 5.2 muestra la asesorías y uso adecuado del software a los alumnos previo a la grabación. La Figura 5.3 muestra el proceso de grabación de audio realizado en un aula cerrada, ubicada en el laboratorio de computo del ITSM, con el fin de reducir ruidos y distractores.

Figura 5.2: Asesorías para los discursos emotivos a los alumnos del ITSM.

El formato de archivo de audio WAV, es un formato sin pérdidas de estandarizado que permite llevar el archivo a distintos reproductores y tener la seguridad de que se está reproduciendo; de esta manera, se evitan problemas de compatibilidad o de la falta de algún *codec o plugin* para reproducirse.

Figura 5.3: Alumnos del ITSM utilizando el modulo de grabación.

El formato WAV es indicado para el corpus emocional creado. El corpus de voz consta de 880 frases en español grabadas por 11 alumnos del ITSM con edades entre 18 y 26 años. estas frases expresan 5 estados emocionales diferentes: *disgusto, ira, felicidad, miedo y neutral* con un total de 16 frases (véase tabla 5.1) por cada uno de ellos. Se han escogido frases cuyo contenido semántico no implique ninguna emoción en concreto de forma que la clasificación se pueda realizar con base a detalles prosódicos.

Frases
1.- Los Tiempos ya no son como antes
2.- De que estas hablando pues
3.- £Quieres un consejo?
4.- La tarea es para mañana
5.- Él es el jefe de grupo
6.- Si, es verdad
7.- No lo creo , no seas chismoso
8.- Siempre llegas tarde
9.- £Puedes guardar silencio por favor?
10.- Si no te gusta , hazlo tu
11.- La computadora de mi mama está descompuesta
12.- La escuela está pintada de rosa
13.- Vivirás conmigo
14.- Mi punto de Vista es otro
15.- Esa actividad no me corresponde
16.- Ahí está un loco

Tabla 5.1: Frases de estimulo diseñadas para cada emoción.

5.2. Algoritmo de reconocimiento de emociones en la voz

5.2.1. Resumen de resultados

En la etapa de procesamiento se logró obtener la señal de audio (véase la figura 5.4).

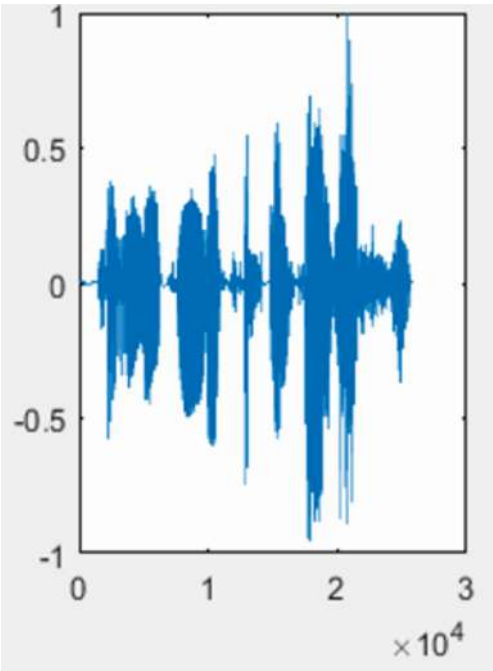

Figura 5.4: La Frase: "Vivirás conmigo"grabada por alumnos del ITSM.

Posteriormente se aplico la etapa de extracción de características obteniendo el espectro de la señal (véase la figura 5.5).

En la etapa extracción de características se logró obtener el espectro de frecuencia que contiene un vector con patrones necesarios para detectar las 5 emociones (véase la figura 5.6).

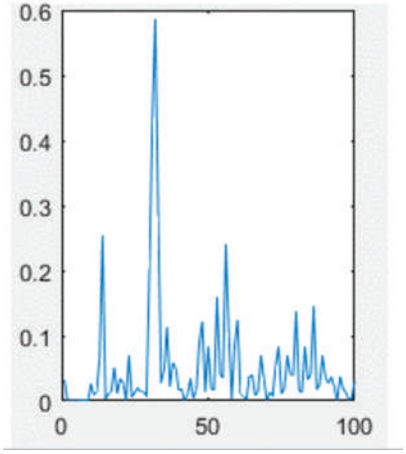

Figura 5.5: Espectro de grabación de la frase "Vivirás conmigo".

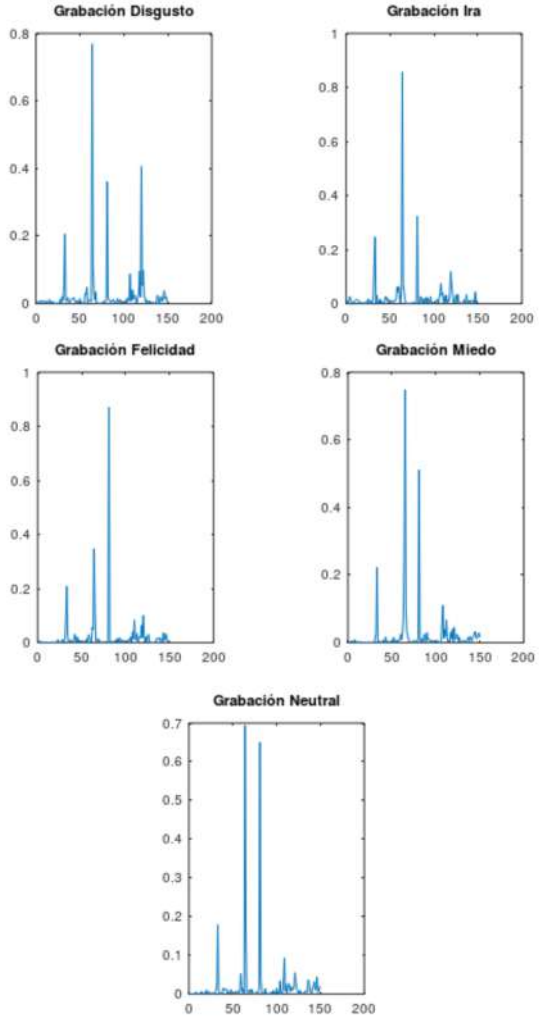

Figura 5.6: Patrón de la frase pronunciada en cada emoción.

En la etapa de clasificación se utilizaron métodos estadísticos que dieron como resultado las diferencias entre el vector a clasificar y los vectores de características almacenados en la base de datos mediante la correlación de Pearson detectando las diferencias por medio del coeficiente de error.

En la tabla 5.2 se muestra el éxito en la detección de la emoción "Disgusto"mediante el coeficiente de error que es el más cercano a 0 y así señalando la semejanza mas significativa en el vector de características con la emoción a reconocer.

Correlación de Pearson	**0.15327**
Coeficiente de Error DISGUSTO:	0.018317
Coeficiente de Error IRA:	0.021492
Coeficiente de Error FELICIDAD:	0.022185
Coeficiente de Error MIEDO:	0.020861
Coeficiente de Error NEUTRAL:	0.052955
Emoción Identificada:	*DISGUSTO*

Tabla 5.2: Reconocimiento del "Disgusto"mediante el método de correlación muestral.

5.2.2. Evaluación del algoritmo

A continuación se muestran los porcentajes de detección de emociones en la tabla 5.3.

Emoción	**Aciertos**
Disgusto	81.25 %
Ira	76.14 %
Felicidad	65.91 %
Miedo	65.34 %
Neutral	45.45 %

Tabla 5.3: Porcentaje de detección de emociones.

La Tabla 5.4 muestra la Matriz confusión del algoritmo utilizado en este trabajo donde se pueden observar que la emoción neutral tiene mayor confusión a diferencia de las demás emociones, también cabe mencionar que el disgusto, la ira y felicidad son emociones claramente identificadas con mayor exactitud por este clasificador. La precisión general es del 80 % de efectividad.

	Disgusto	**Ira**	**Felicidad**	**Miedo**	**Neutral**	**Totales**
Disgusto	143	8	9	5	11	176
Ira	23	134	12	2	5	176
Felicidad	25	13	116	4	18	176
Miedo	18	14	9	115	20	176
Neutral	39	34	21	2	80	176
Totales	248	203	167	128	134	880

Tabla 5.4: Matriz de confusión para el algoritmo de clasificación.

Conclusiones y Trabajos Futuros

El reconocimiento de emociones humanas de manera automatizada es un campo activo de investigación debido a su amplia variedad de aplicaciones. El reconocimiento de emociones es un aspecto clave para obtener interacciones parecidas a las humanas, por eso ha recibido mucha atención por parte de la comunidad científica y, por ende, ha surgido una gran demanda en el desarrollo de aplicaciones que puedan predecir el estado anímico de un usuario. Por tal razón, surge el interés de parte de centros de investigación y diferentes empresas que les gusta estar siempre a la vanguardia tecnológica compartiendo el mismo objetivo de crear sistemas que incorporan la extracción de señales acústicas, hasta poder llegar a tener una aproximación muy significativa en los estados afectivos del hablante.

En este trabajo se mostró una metodología que extrajo los parámetros acústicos para el reconocimiento de estados emocionales en la voz en el área de sistemas inteligentes; el algoritmo matemático mediante Octave GNU incluye: la transformada rápida de Fourier y coeficientes de correlación de Pearson de esta manera se tiene un modelo estadístico capaz de reconocer un 80 % de las frases con emoción actuada por los alumnos del Instituto Tecnológico San Marcos.

Los porcentajes individuales de detección de emociones fueron los siguientes: disgusto 81 %, ira 76 %, felicidad 66 %, miedo 65 % y neutral 45 %. Los resultados demuestran la necesidad de más parámetros en la etapa de extracción de características.

Fue necesario crear una base de datos de emociones por medio de frases por la falta de estandarización en la obtención de emociones y la inexistencia de normas que den garantía en la reproductibilidad. Es indispensable utilizar más métodos de clasificación y técnicas de aprendizaje artificial para tener una mayor eficiencia en la clasificación.

Como trabajo futuro se tiene previsto evaluar el desempeño en otros contextos tales como: llevar a cabo evaluaciones sobre diferentes bases de datos tanto de emociones, reales, como actuadas con el fin de evaluar el alcance del sistema, hacer una evaluación subjetiva con personas no especializadas o no entrenadas y finalmente integrar el sistema de reconocimiento de emociones a un STI.

Bibliografía

[Adrián Torres and Casado Morente, 2002] Adrián Torres, J. A. and Casado Morente, J. C. (2002). *La evaluación clínica de la voz: fundamentos médicos y logopédicos*. Ediciones Aljibe.

[Álvarez et al., 2006] Álvarez, A., Cearreta, I., López, J. M., Arruti, A., Lazkano, E., Sierra, B., and Garay, N. (2006). Feature subset selection based on evolutionary algorithms for automatic emotion recognition in spoken spanish and standard basque language. In *International Conference on Text, Speech and Dialogue*, pages 565–572. Springer.

[Arias Mejía et al., 2015] Arias Mejía, J. M., Bolaños, B., Alexander, E., Orozco Arroyave, J. R., Arias Londoño, J. D., and Vargas Bonilla, J. F. (2015). Evaluacion de metodos de fourier y maxima entropía para la detección automática de la enfermedad de parkinson. *Journal of Research of the University of Quindio*, 27(1).

[Bachorowski, 1999] Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current directions in psychological science*, 8(2):53–57.

[Batliner et al., 2003] Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2003). How to find trouble in communication. *Speech communication*, 40(1):117–143.

[Beceiro et al., 2018] Beceiro, D. I. E., Macias, F. S., Ortiz, S. D. C., and Reyes, E. J. M. (2018). Analizador de voz, llanto infantil y habla usando matlab.

[Brown and Bradshaw, 1985] Brown, B. L. and Bradshaw, J. M. (1985). Towards a social psychology of voice variations. *Recent advances in language, communication, and social psychology*, pages 144–181.

[Burkhardt et al., 2005] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*.

[Bustamante et al., 2015] Bustamante, P., Celani, N. L., Perez, M., and Montoya, O. Q. (2015). Recognition and regionalization of emotions in the arousal-valence plane. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 6042–6045. IEEE.

[Cano, 2018] Cano, N. P. (2018). Recomendaciones para la confección de un corpus oral válido para el análisis fonético. *e-Scripta Romanica*, 5:71–79.

[Carrera and Fernández, 1988] Carrera, M. J. M. and Fernández, A. J. (1988). El reconocimiento de emociones a través de la voz. *Estudios de Psicología*, 9(33-34):31–52.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

[Cowie et al., 2001] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

[Darwin, 1872] Darwin, C. (1872). 1965. the expression of the emotions in man and animals. *London, UK: John Marry*.

[Duque and Morales, 2007] Duque, S. C. and Morales, P. M. (2007). Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones. B.S. thesis, Pereira: Universidad Tecnológica de Pereira.

[Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

[Ekman and Friesen, 1971] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

[Escobedo et al., 2008] Escobedo, D., Sanabria, F., Cano, S., and Marañón, E. (2008). Manual de usuario anavoz 1.0 (registro: 1846-2008). *Universidad de Oriente, Santiago de Cuba*.

[Espinosa et al., 2010] Espinosa, H. P., García, C. A. R., and Pineda, L. V. (2010). Features selection for primitives estimation on emotional speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5138–5141. IEEE.

[Fechner, 1978] Fechner, E. H. (1978). *Children's understanding of the nonverbal communication of emotion in the visual, vocal and gestural.* PhD thesis, ProQuest Information & Learning.

[Garrido Almiñana, 1991] Garrido Almiñana, J. M. (1991). Estilización de patrones melódicos del español para sistemas de conversión texto-habla. *Procesamiento del lenguaje natural. N. 11 (diciembre 1991); pp. 209-219.*

[Hasrul et al., 2012] Hasrul, M., Hariharan, M., and Yaacob, S. (2012). Human affective (emotion) behaviour analysis using speech signals: A review. In *Biomedical Engineering (ICoBE), 2012 International Conference on*, pages 217–222. IEEE.

[Hernández, 2016] Hernández, R. (2016). Sistema de control activado por voz para uso en domótica.

[Hernández et al., 2008] Hernández, Y., Sucar, E., and Conati, C. (2008). An affective behavior model for intelligent tutors. In *International Conference on Intelligent Tutoring Systems*, pages 819–821. Springer.

[Herrera, 2006] Herrera, A. L. R. (2006). Identificacion automatica del lenguaje hablado sin reconocimiento fonético de la señal de voz.

[Kenny et al., 2007] Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447.

[Kim et al., 2007] Kim, E. H., Hyun, K. H., Kim, S. H., and Kwak, Y. K. (2007). Speech emotion recognition using eigen-fft in clean and noisy environments. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pages 689–694. IEEE.

[Larcher et al., 2013] Larcher, A., Bonastre, J.-F., Fauve, B. G., Lee, K.-A., Lévy, C., Li, H., Mason, J. S., and Parfait, J.-Y. (2013). Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *Interspeech*, pages 2768–2772.

[Li and Zhao, 1998] Li, Y. and Zhao, Y. (1998). Recognizing emotions in speech using short-term and long-term features. In *Fifth International Conference on Spoken Language Processing*.

[Marco, 2017] Marco, Giménez, L. (2017). Evaluación y uso del estado emocional en entornos educativos interactivos.

[Martínez-Sánchez, 2010] Martínez-Sánchez, F. (2010). Trastornos del habla y la voz en la enfermedad de parkinson. *revista de Neurología*, 51(9):542–550.

[Miyara, 1999] Miyara, F. (1999). La voz humana. *Laboratorio de Acústica y Electroacústica, Escuela de ingeniería, Electrónica, Facultad de Ciencias Exactas, Ingeniería y Agrimensura, Universidad Nacional de Rosario, Rosario, Santa Fe, Argentina. Obtenido de http://www. fceia. unr. edu. ar/prodivoz/fonatorio. pdf.*

[Montero Martínez, 2003] Montero Martínez, J. M. (2003). *Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano*. PhD thesis, Telecomunicacion.

[Monzo Sánchez et al., 2010] Monzo Sánchez, C. M. et al. (2010). Modelado de la cualidad de la voz para la síntesis del habla expresiva.

[Noroozi et al., 2017] Noroozi, F., Kaminska, D., Sapinski, T., and Anbarjafari, G. (2017). Supervised vocal-based emotion recognition using multiclass support vector machine, random forests, and adaboost. *Journal of the Audio Engineering Society*, 65(7/8):562–572.

[Nwe et al., 2003] Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623.

[Ortego Resa et al., 2009] Ortego Resa, C. et al. (2009). Detección de emociones en voz espontánea. B.S. thesis.

[Pacheco et al., 2015] Pacheco, O. R. A., Beceiro, D. I. E., Macias, F. S., and Lahera, I. N. (2015). Alteración de parámetros acústicos de la voz y el habla en la enfermedad de parkinson. *Simposio Internacional de Comunicación Social. Comunicación Social: Retos y Perspectivas*, 2:679–684.

[Parent, 2005] Parent, A. (2005). Duchenne de boulogne: a pioneer in neurology and medical photography. *Canadian journal of neurological sciences*, 32(3):369–377.

[Pérez Badillo et al., 2013] Pérez Badillo, E. O., Poceros Martínez, F., and Villalobos Ponce, J. A. (2013). Sistema de seguridad por reconocimiento de voz. *Instituto Politécnico Nacional, México*.

[Pérez-Gaspar et al., 2015] Pérez-Gaspar, L.-A., Morales, S. O. C., and Trujillo-Romero, F. (2015). Integración de optimización evolutiva para el reconocimiento de emociones en voz. *Research in Computing Science*, 93:9–21.

[Pérez Pascual, 2017] Pérez Pascual, F. (2017). Speech emotion recognition: Un sistema de reconocimiento de emociones por voz basado en ivectors. B.S. thesis, Universitat Politècnica de Catalunya.

[Plutchik, 1980] Plutchik, R. (1980). Emotion: A psychoevolutionary analysis. *Nueva York: Harper and Row*.

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B. (1993). Fundamentals of speech recognition (prentice hall ptr. *Upper Saddle River, New Jersey*.

[Rowden, 1992] Rowden, C. (1992). *Speech Processing (Essex Series in Telecommunication and Information Systems)*. Mcgraw-Hill (Tx).

[Sánchez et al., 2007] Sánchez, C. D., Pérez, M. M., et al. (2007). *Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones*. PhD thesis, Universidad Tecnológica de Pereira. Facultad de Ingenierías Eléctrica, Electrónica, Física y Ciencias de la Computación. Ingeniería Eléctrica.

[Scherer, 1982] Scherer, K. (1982). Parameters of research on vocal communication: Paradigms and parameters. *Handbook of research methods in nonverbal communication research*.

[Scherer, 1981] Scherer, K. R. (1981). Speech and emotional states. *Speech evaluation in psychiatry*, pages 189–220.

[Scherer, 2000] Scherer, K. R. (2000). Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162.

[Sjölander and Beskow, 2018] Sjölander, K. and Beskow, J. (2018). Wavesurfer.

[Solís, 2011] Solís, V. J. F. (2011). *Modelo de procesamiento de voz para la clasificación de estados*. PhD thesis, Instituto Politécnico Nacional. Centro de Investigación en Computación.

[Steidl, 2009] Steidl, S. (2009). *Automatic classification of emotion related user states in spontaneous children's speech.* University of Erlangen-Nuremberg Erlangen, Germany.

[Sundberg et al., 2011] Sundberg, J., Patel, S., Bjorkner, E., and Scherer, K. R. (2011). Interdependencies among voice source parameters in emotional speech. *IEEE Transactions on Affective Computing*, 2(3):162–174.

[Tintaya, 2005] Tintaya, C. O. J. (2005). Software en matlab para un sistema de adquisición de datos utilizando la tarjeta de sonido de una pc. *Revista de Investigación de Física*, 8(02).

[Vargas et al., 2017] Vargas, E. Y. M., Alencastro, L. T., Balleteros, E. Y. B., and Perdomo, G. R. Á. (2017). El impacto de la afectividad docente en el desempeño académico del estudiante universitario. *Revista Didasc@ lia: Didáctica y Educación. ISSN 2224-2643*, 8(2).

[Ververidis and Kotropoulos, 2006] Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181.

[Vrinda and Shekhar, 2013] Vrinda, M. and Shekhar, M. C. (2013). Speech recognition system for english language. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(1):919–922.

[Wundt, 1896] Wundt, W. M. (1896). *Grundriss der psychologie.* W. Engelmann.

[Zatarain-Cabada et al., 2016] Zatarain-Cabada, R., Barrón-Estrada, M. L., and Muñoz-Sandoval, G. (2016). Premoc: Plataforma de reconocimiento multimodal de emociones. *Research in Computing Science*, 111:97–110.