SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Secretaría Académica, de Investigación e Innovación
Dirección de Posgrado, Investigación e Innovación

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

# Centro Nacional de Investigación y Desarrollo Tecnológico

## Subdirección Académica

## Departamento de Ciencias de la Computación

## TESIS DE MAESTRÍA EN CIENCIAS

### Detección del estado emocional mediante la voz en español de México

presentada por
**Ing. Roberto Hernández Tamayo**

como requisito para la obtención del grado de
**Maestro en Ciencias de la Computación**

Director de tesis
**Dr. Máximo López Sánchez**

Codirector de tesis
**Dr. Noé Alejandro Castro Sánchez**

**Cuernavaca, Morelos, México. Junio de 2016.**

Cuernavaca, Morelos a 14 de junio del 2016
OFICIO No. DCC/153/2016
**Asunto:** Aceptación de documento de tesis

**C. DR. GERARDO V. GUERRERO RAMÍREZ**
**SUBDIRECTOR ACADÉMICO**
**PRESENTE**

Por este conducto, los integrantes de Comité Tutorial del **Ing. Roberto Hernández Tamayo**, con número de control M14CE011, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado "**Detección del estado emocional mediante la voz en español de México"** y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTOR DE TESIS

Dr. Máximo López Sánchez
Doctor en Ciencias de la Computación
7498547

CO-DIRECTOR DE TESIS

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la Computación
08701806

REVISOR 1

Dr. Juan Gabriel González Serna
Doctor en Ciencias de la Computación
7820329

REVISOR 2

Dr. Dante Mújica Vargas
Doctor en Comunicaciones y Electrónica
09131756

C.p. Lic. Guadalupe Garrido Rivera - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

AMR/lmz

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 17 de junio de 2016
OFICIO No. SAC/214/2016

**Asunto:** Autorización de impresión de tesis

**ING. ROBERTO HERNÁNDEZ TAMAYO**
**CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS DE LA COMPUTACIÓN**
**P R E S E N T E**

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Detección de estado emocional mediante la voz en español de México"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

**ATENTAMENTE**
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

SEP TecNM
CENTRO NACIONAL
DE INVESTIGACIÓN
Y DESARROLLO
TECNOLÓGICO
SUBDIRECCIÓN
ACADÉMICA

**DR. GERARDO VICENTE GUERRERO RAMÍREZ**
**SUBDIRECTOR ACADÉMICO**

C.p. Lic. Guadalupe Garrido Rivera.- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

# Dedicatoria

Dedico la realización de la presente tesis a:

- A mi mama, por ser pieza clave en mi formación, por las incontables ocasiones en que se esforzó junto conmigo para iniciar un nuevo día escolar y porque siempre será un ejemplo de gran voluntad.
- A mi papa, que siempre ha estado presente en mis logros y fracasos, por todos los consejos y ejemplos ofrecidos que siempre fueron los indicados.
- A mis hermanos Paola y Juan Pablo, que son mis amigos de viaje y con su compañía la vida es más amena.
- A Miriam, por ese gran apoyo ofrecido y porque siempre me has dado ánimos de seguir adelante pese a todo.

# Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico que me brindó para realizar mis estudios de maestría y la estancia en la ciudad de Tepic, Nayarit.
Al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) por brindarme la oportunidad de superarme profesionalmente al formar parte del programa de Maestría en Ciencias de la Computación.
A mi director de tesis, el Dr. Máximo López Sánchez, por sus enseñanzas, la paciencia que me tuvo, sus consejos, por todo el tiempo invertido, por su dedicación, motivación y ánimo que me transmitió. Pero sobretodo, por la confianza que depositó en mí en todo momento.
A mi codirector de tesis, el Dr. Noé Castro Sánchez que gracias a sus indicaciones me ayudaron a realizar un mejor trabajo.
A mis revisores, la Dra. Graciela Vazquez Alvarez, al Dr. Juan Gabriel González Serna y al Dr. Marco Antonio Oliver Salazar, por sus enseñanzas, su tiempo y dedicación. Y por sus correcciones y observaciones durante el desarrollo de la presente tesis.
Al CICESE Unidad de Transferencia Tecnológica de Tepic (CICESE UT3) por permitirme realizar una estancia y al Dr. Humberto Pérez Espinosa por brindarme la oportunidad de estar trabajando con él, por sus asesorías y amistad.
A mis profesores, el Dr. Andrés Blanco, el Dr. René Santaolaya, el M.C. José Luis Alcántara, a la Dra. Alicia Martínez, al Dr. Guillermo Rodríguez por sus enseñanzas y asesorías brindadas.
A mis padres Concepción y Pablo, a mis hermanos Paola y Juan, y mi novia Miriam, porque sin ellos ningún logro vale lo suficiente.
A mis nuevos amigos que conocí durante este periodo, Vania, Salvador, Bismark, Félix, Jorge Anaya, Jorge Lara, Alyda, Rita, Domingo, Pedro, Sandro, Yahir, Julia Arana, Rodrigo, Alondra, Hector, Hugo, Kenia, Zayani, que han estado conmigo en las reuniones, fiestas así como en los momentos más difíciles. Muchas gracias por su amistad y apoyo.

# Resumen

Durante los últimos años el interés por el reconocimiento de emociones en la voz se ha incrementado de manera notable. Diversos investigadores estudian los problemas y desafíos que implica el reconocimiento de emociones en la voz. Estos desafíos se dividen en tres etapas principales 1) generación y adquisición de datos; 2) extracción y selección de características acústicas; 3) clasificación y métodos de regresión. Estos tres elementos pueden cambiar de acuerdo con varios factores, por ejemplo, el lenguaje, el ruido en la grabación, las emociones para reconocerse, diversidad de oradores, entre otros. Así mismo, se proponen métodos para acercarse a la solución de estos desafíos en diversos contextos.

Hoy en día algunos estudios proponen la creación de bases de datos utilizando oradores en diferentes idiomas; la extracción de conjuntos de características de la voz; y la experimentación con clasificación automática. En la investigación que aquí se reporta se estudian las particularidades del reconocimiento de emociones en la voz en el español hablado en México.

La metodología que se propone en esta investigación se divide en dos etapas principales: extracción y clasificación. En el módulo de extracción se realiza el proceso de extracción de características. Se toman muestras de dos bases de datos emocionales: Emo_voz.mx1 y EmoWisconsin: se realiza el proceso de ventaneo a cada una de las muestras de los corpus. Este proceso lleva a cabo tres tareas: primero se realizan cortes a la señal de audio; se aplica la función de ventaneo y se realizan saltos del 50% de la longitud de cada corte. Posteriormente se extraen 45 características a cada corte. A un conjunto de características se les aplica la función Delta y Delta-Delta y a otro conjunto no. A estos conjuntos se les aplican 24 funciones estadísticas. Por último, se utilizan algoritmos para realizar el proceso de selección de mejores características en el cual se obtienen subconjuntos.

En la etapa de clasificación se utilizan cuatro algoritmos de clasificación para realizar el proceso de entrenamiento y posteriormente su evaluación con cada uno de los subconjuntos.

Los resultados muestran que las características con mejor rendimiento en ambos corpus son la energía, volumen y los MFCCs. En conjunto con la función de ventaneo Blackman y el algoritmo clasificador SMO el cual está basado en las Máquinas Vectores de Soporte. El resultado utilizando la medida F como referencia para Emo_voz.mx1 fue de 0.758 y para EmoWiscosin de 0.451

# Abstract

The interest over the past few years by the recognition of emotions in the voice has increased significantly. Various researchers studying the problems and challenges that involves the recognition of emotions in the voice. These challenges are divided into three main stages 1) generation and acquisition of data; 2) extraction and selection of acoustic features; 3) classification and regression methods. These three elements can change according to several factors, for example, the language, the noise in the recording, emotions to be recognized, diversity of speakers, among others. Furthermore, proposes methods to approach the solution of these challenges in various contexts.

Today some studies suggest the creation of databases using speakers in different languages; the extraction of sets of features of the voice; and experimentation with automatic classification. In the research that is reported here is studying the particularities of the recognition of emotions in the voice in the Spanish spoken in Mexico.

The methodology proposed in this research is divided in two main stages: extraction and classification. In the extraction module performs the process of feature extraction. Samples are taken of two emotional databases: Emo_voice.mx1 and EmoWisconsin: performs the process of windowing to each one of the samples of the corpus. This process performs three tasks: first cuts are carried out to the audio signal; applies the function of window and performs jumps of the 50% of the length of each cut. Subsequently extracted 45 features to each cut. A set of features are applied to them the Delta and Delta-Delta function and another set no. To these sets are applied 24 statistical functions. Finally, algorithms are used to perform the process of selection of best features in which are obtained subsets.

In the classification stage are used four classification algorithms to perform the training process and subsequently its assessment with each of the subsets.

The results show that the features with better performance in both corpus are energy, volume and the MFCCs. In conjunction with the window function Blackman and the SMO classifier algorithm which is based on the Support Vector Machines. The result to using the Measure F as reference for Emo_voice.mx1 was of 0.758 and 0.451 for EmoWiscosin.

# Índice

# Índice de figuras

# Índice de tablas

# Capítulo 1 Introducción

## 1.1 Motivación

En la actualidad los sistemas de Interacción Humano-Computadora (IHC) están ganando mucho interés ya que juegan un papel importante en la vida diaria. Son sistemas de habla y visión, ya que estos medios de percepción son los canales más naturales en la comunicación humana (Peréz Espinosa & Reyes García, 2010). Uno de los grandes objetivos que tienen los sistemas IHC es que la computadora escuche el mensaje del usuario y responda de manera natural. En el campo del habla de los sistemas IHC existen muchas aplicaciones que al día de hoy resultan de gran ayuda para las personas. Algunas de estas aplicaciones son los IVR (*Interactive Voice Response*), sistemas inteligentes de automóviles, videojuegos, sistemas de transcripción de texto por medio del dictado (Ortego Resa, 2009).

La voz provee de gran información que permite saber si una persona está diciendo la verdad o una mentira. También refleja el estado emocional de la persona que habla (Solís Villarreal, 2011) y es en este punto donde se centra esta tesis.

Existen numerosas investigaciones que trabajan en la detección de estados emocionales por medio de la voz. Para lograr esto se necesita llevar a cabo una serie de pasos (Rao Krothapalli & G. Koolagudi, 2013): primero tener un corpus con muestras de datos etiquetados para cada una de las emociones. Actualmente, existe un gran número de corpus para diferentes idiomas que se observan en la investigación realizada por (Ververidis & Kotropoulos, 2006). Posteriormente, se extraen características de las muestras de audio que se clasifican en acústicas y lingüísticas (Chiriacescu, 2009). Por último, se entrena un modelo de clasificación que se encargue de clasificar las características y decidir a qué emoción pertenecen esas características. Existen muchos algoritmos de aprendizaje máquina para entrenar modelos como se observa en la investigación de (Ververidis & Kotropoulos, 2006) y dependerá de las características del problema que algoritmo aplicar.

Tomando en cuenta lo citado en los párrafos anteriores; la detección del estado emocional a través de la voz es de suma importancia ya que podría servir de base para crear aplicaciones que permitan monitorear estos cambios emocionales y actuar con base en ello. En esta tesis se pretende desarrollar un módulo de clasificación que detecte el estado emocional a partir de la voz en idioma español de México. Dado que actualmente la mayoría de las investigaciones relacionadas detectan las emociones en idioma distinto al que se estudia en esta tesis.

## 1.2 Planteamiento del problema

Actualmente, existen investigaciones que han estado trabajando en la identificación de estados emocionales mediante la voz, como los realizados por (Solís Villarreal, Yáñez Márquez, & Suárez Guerra, 2011), (Echeverry Correa & Morales Pérez, 2008) y (Bojanić & Delić, 2013). Estas investigaciones comparan diferentes métodos de clasificación para encontrar los diversos niveles de estados emocionales. Permitiendo definir las mejores características acústicas de voz.

Para realizar lo anterior y verificar que funcione correctamente el clasificador. Se realizan pruebas en bases de datos emocionales, la mayoría en idioma alemán (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005), dado que su uso es de libre acceso. También, se han realizado trabajos de esta índole con bases de datos emocionales en el idioma Inglés (Martin, Kotsia, Macq, & Pitas, 2006) y en el idioma Español de España (Montero, Gutiérrez, Palazuelos, Enríquez, Aguilera, & Pardo, 1998). Las bases de datos mencionadas se utilizan mucho en este tipo de investigaciones aunque otras bases de datos se encuentran en la investigación realizada por (Ververidis & Kotropoulos, 2006).

Los trabajos actuales que detectan la emoción en base a muestras de audios de datos emocionales regularmente se hacen en idiomas distintos al español y solo unas pocas las dedican al idioma español nativo de España. A la fecha solo existen cuatro bases de datos emocionales de idioma Español nativo de México las cuales son investigaciones realizadas por: (Reyes Vargas, y otros, 2013), (Caballero Morales, 2013), (Pérez Espinosa, Reyes García, & Villaseñor Pineda, 2011) y (Macías Kempe, 2008).

El problema que se observa entre el idioma nativo de España y México es que la eficiencia entre ambos es significativa. Se analizará la manera en que afectan cambios en el procesamiento de características acústicas en la clasificación de emociones así como el aporte de diferentes tipos de características acústicas a la clasificación.

## 1.3 Justificación

Uno de los usos que se le puede dar a un sistema de reconocimiento de emociones es en la inteligencia ambiental. La inteligencia ambiental aplica los sistemas IHC en los entornos en donde se encuentra el usuario. Son capaces de interactuar con las personas de un modo natural, sensible al usuario y al contexto, para actuar de forma proactiva. Cuando los sistemas son sensibles al contexto procesan la información y emiten una respuesta en base a la característica que presenta el usuario ya sea situacional, emocional o temporal.

Los *Ambient Assited Living* (ALL) son aplicaciones en el paradigma de la inteligencia ambiental que promueve y prolonga la vida independiente de las personas mayores y personas con alguna discapacidad. En México hay un total de 119, 530, 753 personas, esto en base a la encuesta intercensal 2015 que realizó (INEGI, 2015). En esta encuesta se observa que del total el 7.2% representa a la población en edad avanzada. En el año 2000 esta parte de la población representó el 5.0% y en 2010 representó el 6.2%. El proceso de envejecimiento en el país es notoria por lo que la implementación de este tipo de aplicaciones se hará más evidente. Ayudar a las aplicaciones ALL a ser sensibles al contexto, específicamente en detectar las emociones, ayudará a cumplir uno de sus objetivos: Apoyar el mantenimiento de la salud y capacidad funcional de las personas mayores de una manera natural.

Otra de las aplicaciones de los sistemas IHC y que su utilización está siendo cada vez mayor son los asistentes virtuales. Un asistente virtual es un software, que en la mayoría de los casos se encuentran en *smartphones*, que reconoce el lenguaje natural para procesar la información y ofrecer un servicio mediante la voz. Las tareas que realizan los asistentes en *smartphones* son: control de agenda, solicitud de información (clima, tráfico, noticias), enviar notificaciones, realizar reservaciones (restaurantes, cine, espectáculos), manejador personal de salud, entre otras funciones. Algunos ejemplos de estas aplicaciones son: Siri de Apple, Google Now de Google y Cortana de Microsoft.

Estas aplicaciones solo procesan la información de lo que el usuario solicita pero no toman en cuenta otro tipo de información complementaria. La voz es el medio por el cual procesa la solicitud. Si además de las instrucciones tomara parámetros para detectar el estado emocional del usuario, ofrecería un mejor servicio al usuario. Por ejemplo, si el usuario solicita una tarea pero el asistente detecta que se encuentra en estado emocional negativo, dependiendo de la solicitud respondería de una manera diferente que le permitiera cambiar a una emoción positiva o en dado caso, rechazar la solicitud.

En México la contratación de dispositivos móviles va a la alza. De acuerdo a un informe estadístico del Instituto Federal de Comunicaciones (IFT) (ift, 2016) al terminar el año 2015 hay un total de 107.7 millones de suscripciones a telefonía móvil. De esto el 71.6% son teléfonos inteligentes, según datos de *The Competitive Intelligence Unit* (the-ciu, 2016), lo que hace que un gran porcentaje de la sociedad mexicana cuente con un dispositivo para

implementar aplicaciones de este tipo. Al ser un producto novedoso su utilización será mayor. Implementar un módulo de reconocimiento de emociones en aplicaciones IHC ayudaría a mejorar la experiencia de usuario.

## 1.4 Objetivos

### 1.4.1 Objetivo general

Implementar un modelo de clasificación que en conjunto con características acústicas permita reconocer los estados emocionales por medio de la voz en idioma español mexicano.

### 1.4.2 Objetivos específicos

- Obtener datos de corpus en idioma español mexicano.
- Analizar grupos de características de la voz y definir la mejor de ellas.
- Definir un clasificador de los que se han trabajado en el reconocimiento de emociones por medio de voz.
- Desarrollar una aplicación que relacione el clasificador y el grupo de características definidos para clasificar correctamente las muestras que aparecen en los corpus que se tomaron.

## 1.5 Alcances y limitaciones del proyecto

De acuerdo al trabajo que se pretende realizar, se presentan a continuación los alcances y limitaciones.

### 1.5.1 Alcances:

- La detección de emociones se realizará solo para el idioma español de México.
- Las muestras de audio de los corpus se usarán para la extracción de características.
- Se utilizarán dos bases de datos emocionales en idioma Español de México.

### 1.5.2 Limitaciones:

- Hasta la fecha solo existen cuatro bases de datos emocionales en idioma Español de México para realizar las pruebas. Las cuales son investigaciones de: (Reyes Vargas, y otros, 2013), (Caballero Morales, 2013), (Pérez Espinosa, Reyes García, & Villaseñor Pineda, 2011) y (Macías Kempe, 2008).

- Las características con las que cuentan estas bases de datos como: el acento de las personas de diferentes partes del país, el número de participantes, las emociones con las que cuenta cada corpus, el contexto de generación de emociones, el ambiente de grabación y herramientas utilizadas para llevarlas a cabo. Todo esto puede influir en el resultado de predicción.

# Capítulo 2 Fundamento Teórico

A continuación se describen las definiciones de los conceptos más relevantes utilizados en esta tesis.

## 2.1 Base de Datos Emocional

Para evaluar el sistema de reconocimiento de emociones es necesario contar con muestras de audio que se encuentren clasificadas para cada una de las emociones. Estas muestras se encuentran en repositorios conocidas como bases de datos emocionales. Los resultados obtenidos en el sistema de reconocimiento de emociones son muy dependientes de estas muestras utilizadas para la evaluación. Estos repositorios proporcionan información del procedimiento utilizado para grabar las voces: el idioma, las etiquetas que contiene cada muestra (alegre, triste, enojado, etc.), tipo de discurso (natural, simulada o provocada) y otras señales fisiológicas presentes en las grabaciones que son de utilidad para la detección de una emoción (Ververidis & Kotropoulos, 2006).

## 2.2 Características en el habla

En la voz existen dos tipos de características que se extraen para ser analizadas: características acústicas y características lingüísticas (Chiriacescu, 2009). Actualmente las investigaciones que trabajan en la detección de emociones tratan de encontrar cual es el conjunto de características o la combinación de estas más adecuada que den mejores resultados para el proceso de clasificación (Echeverry Correa & Morales Pérez, 2008). A continuación se realiza una explicación de las características acústicas y posteriormente las características lingüísticas.

### 2.2.1 Características acústicas

Dentro de este tipo de características se encuentran tres categorías que se relaciona con la detección de emociones (Chiriacescu, 2009): prosódicas, espectrales y calidad de voz.

- Prosódicas (Peréz Espinosa & Reyes García, 2010). Esta cumple una función clave del discurso. Trasmite información emotiva, sociolingüística y dialectal. Es el conjunto de fenómenos fónicos que abarcan más de un fonema o segmento. Estudia el tono, volumen, velocidad del habla, duración, pausa y el ritmo de los segmentos más grandes de la oración como silabas, palabras o fonemas. A continuación se explican brevemente las características que se utilizaron en esta investigación:
    - Volumen: Es el volumen o la fuerza de una señal de audio.

  - Formantes de frecuencia (*formant frequency*): Son picos de frecuencia que tienen en el espectro un alto grado de energía.
  - Formantes de ancho de banda (*formant bandwidth*): Se calculan a 3 dB por debajo de la frecuencia central.
  - Velocidad del habla (*Speechrate*): Es la medida de la velocidad del habla que calcula palabras por minutos o silabas por minuto.
- Espectrales (Chiriacescu, 2009). También llamada análisis en el dominio de la frecuencia. Localiza parámetros de la señal de voz atendiendo a la información que provee su espectro. A continuación se explican brevemente las características que se utilizaron en esta investigación (Giannakopoulos & Pikrakis, 2014):
  - Centroide espectral (*Spectral centroid*): Es el "centro de gravedad" del espectro.
  - Spread espectral (*Spectral spread*): Es el segundo momento central del espectro.
  - Entropía espectral (*Spectral entropy*): Se calcula como la entropía de la energía pero esta vez en el dominio de la frecuencia.
  - Flujo espectral (*Spectral flux*): Mide que tan rápido el espectro de una señal está cambiando el cual se mide entre dos tramas sucesivas. Es utilizado para determinar el timbre de una señal de audio.
  - Rollof espectral (*Spectral Rollof*): Representa la frecuencia en la cual cierto porcentaje (usualmente 80-90 %) de la magnitud del espectro está concentrada. Se utiliza para discriminar entre sonidos sordos y sonoros.
  - Coeficientes Cepstrales en la Frecuencia de Mel (MFCC por sus siglas en inglés): Son una representación definida como el cepstrum de una señal ventaneada en el tiempo que ha sido derivada de la aplicación de una Transformada Rápida de Fourier, pero en una escala de frecuencias no lineal, las cuales se aproximan al comportamiento del sistema auditivo. En muchas aplicaciones se selecciona las primeras 13 porque consideran que contienen suficiente información discriminativa.
  - Relación armónica (*Harmonic Ratio*): Es la relación de potencia de la frecuencia fundamental en una trama de audio. Es una medida para el grado de armonicidad contenida en una señal.
  - Frecuencia fundamental (*Fundamental frequency*): Es la frecuencia a la que vibran las cuerdas vocales. Se le conoce también como F0. Las características de la frecuencia fundamental incluyen contorno, media, variabilidad, y distribución.
  - Vector croma (*Chroma vector*): También se le conoce como cronograma. Es un espectrograma que representa la energía espectral de las 12 clases de tono.

- Calidad de voz (Ortego Resa, 2009). Permite distinguir los sonidos que tienen el mismo tono y volumen. Estas características marcan las diferencias entre emociones. Algunas características de este grupo son: irregularidades de voz o timbre, ruido, intensidad, cociente de la abertura de las cuerdas vocales y altura que es la que produce la sensación de agudo o grave.

También están las características de la voz en donde la señal no se lleva a ningún tipo de representación y se toman los valores de acuerdo al dominio del tiempo. A continuación se explican las siguientes características que se aplicaron en esta investigación (Giannakopoulos & Pikrakis, 2014):

- Tasa de cruces por cero (*Zero Crossing Rate*): Representa la cantidad de veces que la señal cambia de signo, de positivo a negativo y viceversa. Se emplea para medir el ruido en una señal.
- Energía (*Energy*): También llamado fuerza de una señal. Cuando un audio tiene un volumen alto tendrá mucha energía, esta característica también es usada para identificar silencios.
- Entropía de energía (*Entropy of energy*): Es una medida que mide los cambios bruscos en el nivel de la energía de una señal de audio. Se utiliza en la detección de disparos, explosiones y varios sonidos ambientales.

### 2.2.2 Características lingüísticas

Estas se basan en que las señales lingüísticas se representan como una colección desordenada de palabras (bolsa de palabras) (Chiriacescu, 2009) sin tener en cuenta la gramática, pero se hace el seguimiento de frecuencia de las palabras. Otro tipo de caracterización se basa en la estimación de la probabilidad de una emoción dando una cierta secuencia de palabras muy similar a los modelos de lenguaje que se manejan en el reconocimiento de voz. Los que más se utilizan son los unigramas y bigramas.

Otro enfoque de las características lingüísticas son los estados afectivos asociados con las palabras. Se pueden encontrar utilizando el diccionario afectivos en lenguaje de (Whissell, 1989) o el léxico afectivo de Ortony (Ortony, Clore, & Collins, 1988)

## 2.3 Delta y Delta-Delta (Young, y otros, 2006)

Otra de las características que se obtienen en esta investigación es aplicar funciones Deltas a cada una de las características que se extrajeron. La función Delta también se le conoce como coeficientes de velocidad. Mide la variación en un tiempo de las características a las que se le aplica. A la función Delta-Delta también se le conoce como Doble Delta o coeficientes de

aceleración dado que mide la variación en un tiempo de los coeficientes de velocidad (Delta). Estas funciones miden si la señal no mantiene cambios durante un determinado tiempo.

## 2.4 Técnicas de clasificación

Actualmente existen varias técnicas de clasificación para el reconocimiento de estados emocionales en el habla. En esta sección se describen brevemente las más populares.

### 2.4.1 Análisis Discriminante Lineal (Chiriacescu, 2009)

Esta técnica de clasificación es la más simple y la más rápida. Normalmente sus resultados son comparables con los resultados de clasificadores complejos. Esta técnica trabaja de la siguiente forma: se tiene un conjunto de objetos que se clasifican en una serie de grupos. Esta técnica equivale a un análisis de regresión donde la variable de regresión dependiente es categórica y tiene como categorías la etiqueta de cada uno de los grupos y, las variables independientes son continuas y determinan a que grupos pertenecen los objetos. Tratan de encontrar relaciones lineales entre las variables continuas que mejor discriminen en los grupos dados a los objetos. En la Figura 2.1 se observa la relación lineal que mejor describe que datos pertenecen a una clase.

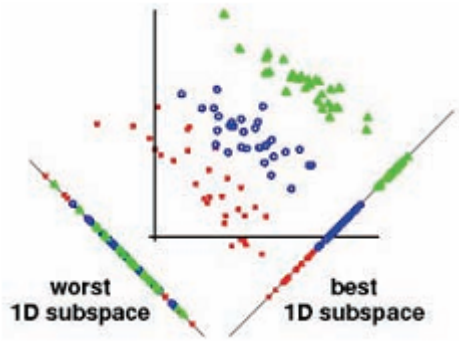


*Figura 2.1 Análisis de Discriminante Lineal*

### 2.4.2 K Vecinos más Cercanos - KNN (Moujahid, Inza, & Larrañaga, 2008)

El algoritmo K vecinos más cercanos también es un clasificador lineal. Clasifican un nuevo elemento por mayoría de votos de acuerdo a sus k vecinos más cercanos. Primero se calculan las distancias de todas las muestras de entrenamiento en base al nuevo caso que se quiere clasificar, después, se seleccionan los k elementos más cercanos al nuevo caso y dependiendo del número más frecuente entre los k elementos es donde se clasificará el nuevo caso. Esto se observa en la Figura 2.2, donde el nuevo dato se clasificará en la clase de círculos ya que la cantidad de vecinos respecto al nuevo dato es mayor en comparación con la clase de cruces.

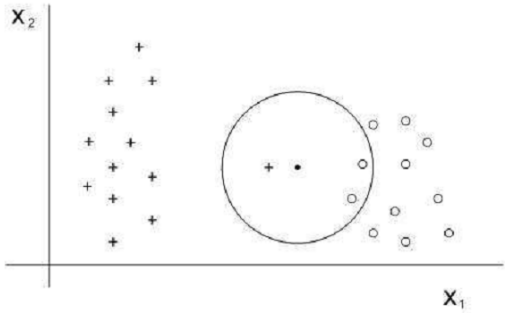


*Figura 2.2 K Vecinos más cercanos*

### 2.4.3 Teorema de Bayes (Mamani Laqui, 2014)

Este tipo de clasificador se basa en el teorema de Bayes y es de los más usados en el campo de reconocimiento de patrones. Es un método de clasificación probabilístico. Este clasificador hace uso del conocimiento de las distribuciones de probabilidad de las características para clasificar objetos, dado que no es posible encontrar características que permitan separar linealmente las clases. En la Ecuación 1 se observa la fórmula donde se expresa la probabilidad condicional de que ocurra un evento h dado D. Donde:

- P(h|D): Probabilidad de que h sea cierta después de observar D.
- P(D|h): Es la probabilidad de observar el conjunto de entrenamiento D en un universo donde se verifica la hipótesis h.
- P(h): Probabilidad de h sin ninguna observación.
- P(D): Probabilidad de observar D, sin saber que hipótesis se verifica.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

*Ecuación 1. Teorema de Bayes*

### 2.4.4 Modelo Oculto de Markov – HMM (Maldonado, 2012)

Describe un proceso de probabilidad el cual produce una secuencia de eventos simbólicos observables. Son autómatas abstractos finitos que modelan procesos no deterministas. La ocurrencia de los estados está asociada con una distribución de probabilidad y las transiciones entre los estados son un conjunto de probabilidades llamadas probabilidades de transición de estados. La observación en un estado particular se genera de acuerdo a una distribución de probabilidad. Los estados no son visibles y su ocurrencia depende del estado anterior. Este modelo es muy utilizado en el área de reconocimiento de voz dado que representa

estadísticamente la característica acústica de voz, dado que dichas características no son estacionarias. Por ejemplo modelan como suenan los distintos fonemas y palabras.

En la Figura 2.3 se observa una pequeña presentación de la transición de estados de un HMM, donde "x" representa los estados ocultos, "y" las salidas observables, "a" las probabilidades de transición y "b" las probabilidades de salida.

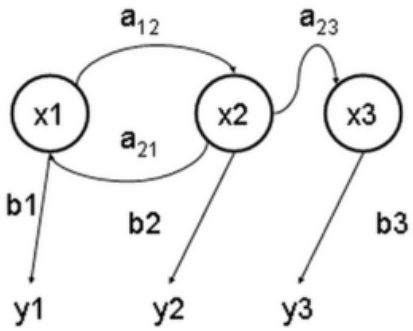


*Figura 2.3. Transición de estados en un HMM*

## 2.4.5 Modelo de Mezcla Gaussiana – GMM (Ortego Resa, 2009)

Aplicado en el reconocimiento automático de emociones identifica que las emociones tienen diferentes sonidos y que la frecuencia de los sonidos es diferente una de otra. Este modela la distribución de probabilidad de los parámetros de un fragmento de audio. El modelado de la distribución de probabilidad de los parámetros se realiza a partir de una suma de M funciones de densidad Gaussiana, cada una parametrizada por un vector de medias y una matriz de covarianza. En la Figura 2.4 se observa un ejemplo de la probabilidad que tiene cada componente.

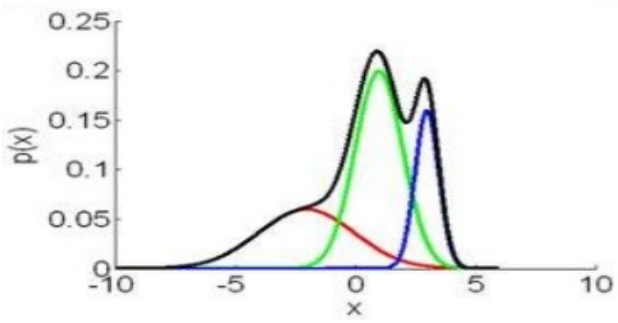


*Figura 2.4. Probabilidad de cada componente en GMM*

### 2.4.6 Máquinas Vectores de Soporte – SVM (Rao Krothapalli & G. Koolagudi, 2013)

Este es de los clasificadores más populares en el reconocimiento de emociones. Es un clasificador binario que asigna cada patrón a una clase. Este clasificador trata de determinar un hiperplano que maximice el margen entre los dos conjuntos de datos. Separa entre dos clases la cual una es verdadera y la otra falsa. Las muestras que se encuentran en el margen son llamados vectores de soporte. Cuando se define el hiperplano de separación entre las dos clases, se debe encontrar una función que clasifique las muestras en su clase correspondiente. En la Figura 2.5 se observa un ejemplo de separación de dos clases de datos creando un gran margen entre ellas.

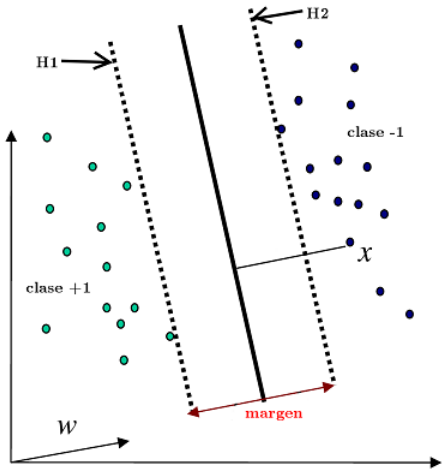


*Figura 2.5. Separación de dos clases con SVM*

### 2.4.7 Redes Neuronales Artificiales (Navarrete García, 2003)

Las Redes Neuronales Artificiales (ANN por sus siglas en inglés) son un modelo que trata de imitar el comportamiento de las neuronas biológicas. Se compone de un conjunto de neuronas artificiales interconectadas. Una ANN es una estructura compuesta de un número de neuronas artificiales. Cada una de las neuronas artificiales posee una característica de entrada/salida e implementa una función. La salida de cada neurona está determinada por su característica de entrada, la interconexión con otras neuronas y (opcionalmente) de sus entradas externas. En la Figura 2.6 se observa la representación de funcionamiento del cálculo de una neurona artificial. Se aplica un conjunto de entradas, donde cada una representa la salida de otra neurona, posteriormente se realiza una suma ponderada con esos valores y se obtiene el valor ejecutando una función.

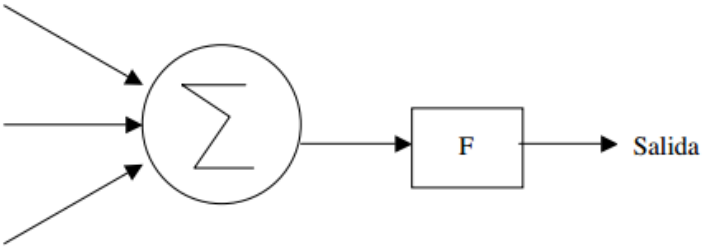


*Figura 2.6. Calculo de salida de una neurona artificial*

## 2.4.8 Árboles de decisión (Chiriacescu, 2009)

Este tipo de clasificador se representa gráficamente por un árbol de decisión. El objetivo es tener buenos clasificadores a partir de árboles sencillos. A partir de estos se exploran árboles más complejos hasta llegar a un compromiso entre exactitud y complejidad. Los árboles de decisión también denominados *Top Down Induction of Decision Trees* (TDIDT), se caracterizan por utilizar una estrategia de divide y vencerás descendente, es decir, partiendo de los descriptores hacia los ejemplos, dividen el conjunto de datos en subconjuntos siguiendo un determinado criterio de división. A medida que el algoritmo avanza, el árbol crece y los subconjuntos de ejemplos son menos numerosos. Cada nodo hoja tiene una clase asociada y cada nodo interno tiene un predicado (o, de manera más general, una función) asociado. Para clasificar un nuevo caso se empieza por la raíz y se recorre el árbol hasta alcanzar una hoja; en los nodos internos se evalúa el predicado (o la función) para el ejemplo de datos, para hallar a qué nodo hijo hay que ir. El proceso continúa hasta que se llega a un nodo hoja. En la Figura 2.7 se observa una representación de un árbol decisión tomando el ejemplo del artículo de (Chambi, 2013). Se desea tomar una decisión con respecto a otorgar un crédito o no a un cliente en una institución de micro finanzas.

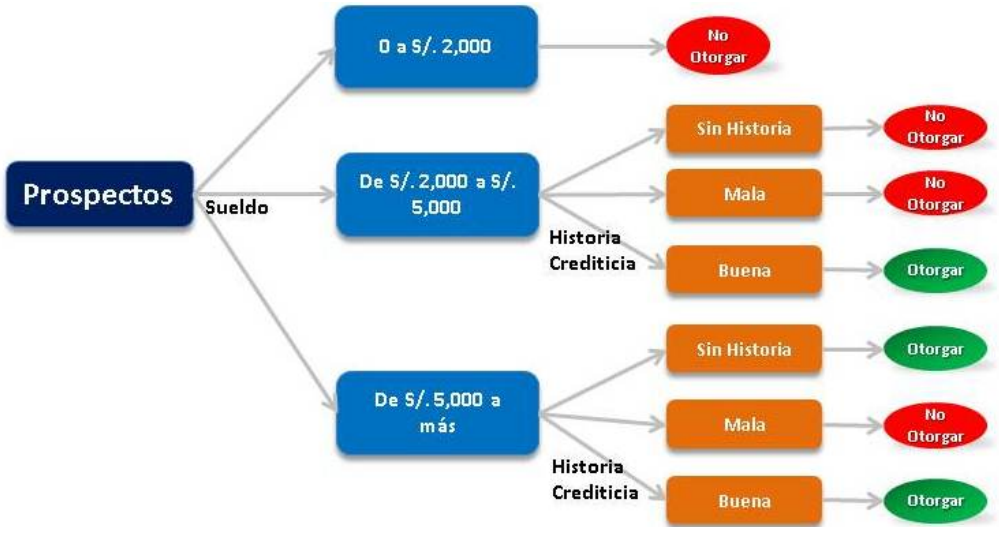


*Figura 2.7. Ejemplo de árbol de decisión*

## 2.5 Waikato Environment for Knowledge Analysis (WEKA)
(Witten, Frank, & Hall, 2011)

WEKA es conjunto de librerías Java que permite analizar, evaluar y aplicar tareas de minería de datos como: pre-procesamiento, clasificación, agrupación, asociación, selección de atributos y visualización. Es una herramienta que ha sido desarrollada en la universidad de Kakato (Nueva Zelanda) bajo la licencia GNU *Public License* (GPL). En esta investigación se hace uso de dos tareas de este software: selección de tributos y clasificación. A continuación se dará una breve explicación.

### 2.5.1 Selección de atributos

La selección de atributos en WEKA son un conjunto de métodos que permiten identificar, mediante un conjunto de datos que poseen ciertos atributos, aquellos atributos que tienen más peso a la hora de terminar si los datos son de una clase u otra. Se selecciona el subconjunto más pequeño que cumpla con este objetivo tal que no se afecte significativamente el porcentaje de clasificación y que la distribución resultante sea lo más parecida a la original. La intención es eliminar los datos redundantes, irrelevantes o ruidosos y evitar procesos lentos debido a un exceso de información poco significativa.

Para la selección de atributos se realizan dos tareas en conjunto:

- Método de evaluación (*Attribute Evaluator*): Este método es el encargado de evaluar cada uno de los casos a los que se le enfrente y dotar a cada atributo de un peso específico. Determina la calidad del conjunto de atributos para discriminar la clase.
    - *CfsSubsetEval*. Calcula la correlación que tiene cada clase con cada uno de los atributos. Se prefieren los subconjuntos de características que están altamente correlacionados con la clase.
    - *InfoGainAttributeEval*. Evalúa los atributos midiendo la ganancia de información con respecto a la clase. Evalúa un atributo a la vez.
    - *WrapperSubsetEval*. Evalúa subconjuntos de atributos utilizando un clasificador. La validación cruzada es utilizada para medir la calidad del subconjunto de atributos a través de la tasa de error del clasificador. El método envuelve al clasificador para explorar la mejor selección de atributos que optimiza sus prestaciones.
- Método de búsqueda (*Search Method*): Atraviesa el espacio de atributos para encontrar un buen subconjunto. La calidad de estos subconjuntos se miden por los métodos de avaluación que se explicaron.
    - *BestFirst*. Busca en el espacio de los subconjuntos de atributos utilizando la estrategia *Greedy Hillclimbing* con *Backtracking*. La búsqueda puede comenzar

con el conjunto de atributos vacío y hacia adelante, o con el conjunto de atributos lleno y buscar atrás, o empezar en cualquier punto y realizar la búsqueda en ambas direcciones.

- *GeneticSearch*. Utiliza un algoritmo genético simple (Goldberg, 1989).
- *LinearForwardSelection*. Es un método de búsqueda sub-óptima en escalada. El procedimiento es el siguiente: se elige primero el mejor atributo, después, se añade el siguiente atributo que más aporta y continúa así hasta llegar a la situación en la que añadir un nuevo atributo empeora la situación.
- *Ranker*. Ordena los atributos de mayor a menor de acuerdo a su ganancia de información.

### 2.5.2 Clasificación

La clasificación es una técnica que permite identificar la categoría a la que pertenece una nueva instancia, tomando como base un conjunto previo de instancias categorizadas en clases (entrenamiento). El algoritmo clasificador realiza una generalización de las instancias conocidas para predecir la categoría de la nueva.

WEKA dispone de varios algoritmos de clasificación. Los que se utilizan en esta investigación de describen a continuación:

- *NaiveBayes*. Es un algoritmo clasificador fundamentado en el teorema de Bayes. Construye modelos que predicen la probabilidad de posibles resultados.
- *MultilayerPerceptron*. Es un algoritmo clasificador basado en una red neuronal artificial multicapa. Utiliza *Backpropagation* para clasificar las instancias.
- *RandomForest*. Es un algoritmo de clasificación desarrollada por Leo Breiman (Breiman, 2001) que mejora la precisión en la clasificación mediante la incorporación de aleatoriedad en la construcción de cada clasificador individual. Esta aleatorización puede introducirse en la partición del espacio (construcción de árboles de decisiones), así como en la muestra de entrenamiento.
- *SMO*. Esta técnica implementa el algoritmo Optimización Mínimo Secuencial de Jhon Platt (Platt, 1998) para el entrenamiento de un clasificador de Máquina Vector de Soporte. Trabaja con problemas cuadráticos grandes los mismos que los divide en pequeños. Para problemas multiclase se resuelve utilizando la clasificación por parejas.

WEKA permite la aplicación de varios tipos de pruebas. La que se aplicó en esta investigación es la siguiente:

- *Cross-validation*. Divide el total de las instancias en carpetas (*folds*). Para esta investigación se le especificó un total de 10 carpetas. Para la construcción del modelo, se consideran las instancias de una carpeta como datos de prueba y el resto como datos

de entrenamiento. Esto se repite con cada una de las carpetas. Los errores calculados serán el promedio de todas las ejecuciones.

Al terminar las pruebas, WEKA proporciona diversas métricas e información que son de relevancia para la investigación. Las que se consideraron son las siguientes:

- Porcentaje de instancias clasificadas correctamente e incorrectamente.
- Precisión. Es la proporción del número de instancias de la clase x entre el número de instancias que fueron clasificadas como de la clase x. El valor de 1 significa mayor precisión.
- Exhaustividad (*Recall*). Es la proporción del número de instancias que fueron clasificados como clase x de entre el número de instancias de la clase x. El valor de 1 significa mayor exhaustividad.
- Medida F (*F-Measure*). Es una medida combinada de precisión y exhaustividad $\frac{2*Precisión*Recall}{Precisión+Recall}$. El valor de 1 significa un óptimo resultado.

# Capítulo 3 Estado del arte

Los trabajos relacionados presentados en esta tesis se tomaron en cuenta porque algunos detectan la emoción en idioma Español de México, el cual es el tema de investigación de esta tesis, y los demás en distintos idiomas. También nos permite conocer la metodología que cada uno utilizó. Esto nos permite conocer y comparar los resultados de predicción en diversos idiomas. Los siguientes trabajos se dividen en tres categorías: creación de una base de datos emocional, extracción de características y modelos de clasificación. El primero se refiere a investigaciones que se enfocan en la creación de una base de datos emocional donde se detalla el procedimiento y herramientas que utilizaron. El segundo, son investigaciones que se enfocan en la extracción de características de la voz observando mediante pruebas cual es el mejor conjunto de característica. El tercer paso se refiere a investigaciones que comparan cual modelo de clasificación es el mejor para la detección de emoción.

Los criterios de evaluación para cada una de las investigaciones que aquí se presentan son: objetivo general, metodología y resultados. A continuación se detalla cada criterio:

- Objetivo general: Se detalla el objetivo general de la investigación. Especifica las herramientas y técnicas que utilizaron.
- Metodología: Se especifica el procedimiento que utilizaron para el reconocimiento de emociones. Se detalla la base de datos emocional, los conjunto de características acústicas y los clasificadores que utilizaron.
- Resultados: Se presenta los porcentajes de precisión que obtuvieron en cada una de las pruebas que realizaron y el conjunto de características con el clasificador que mejor dio resultado.

## 3.1 Creación de base de datos emocional

A continuación se enlistan las investigaciones que crean bases de datos emocionales, las cuales permitirán evaluar sistemas de reconocimiento de emociones.

### 3.1.1 EmoWisconsin: Una base de datos de habla en niños emocional en el español de México (Pérez Espinosa, Reyes García, & Villaseñor Pineda, 2011)

#### Objetivo general

Esta investigación presenta la creación de una base de datos emocional de habla inducida en Español de México el cual lleva por nombre “EmoWisconsin”. Fue grabada con niños de entre 7 y 13 años de edad mientras realizaban un juego de cartas con un examinador adulto. El audio fue segmentado y anotado con 6 categorías emocionales y 3 emociones primitivas continuas

por 11 evaluadores humanos. Posteriormente se realizaron pruebas de clasificación y regresión usando un conjunto de 6,552 características acústicas.

## Metodología

Para la realización de la base de datos EmoWisconsin los autores adoptaron la prueba neurológica *Wisconsin Card Sorting Test* (WCST) (Grant & Berg, 1948) para inducir diferentes estados emocionales a 28 niños de entre 7 y 13 años de edad de los cuales 11 fueron niñas y 17 niños. Se grabó en dos sesiones en un ambiente controlado, moderadamente ruidoso y bajo la supervisión de dos examinadores adultos. En la primera sesión un examinador controlaba un ambiente motivante para los niños y les daba consejos. Esto daba como consecuencia que se evocaran emociones positivas como alegría y seguridad. La segunda sesión el otro examinador controlaba un ambiente estresante y sin ayuda para los niños lo cual provocaba emociones negativas tales como frustración y nerviosismo.

Con los audios grabados realizan una segmentación manualmente. Obtienen un total de 3,098 segmentos donde 1,421 se obtuvieron de la primera sesión (positiva) y 1,674 de la segunda sesión (negativa). En el proceso de etiquetado utilizan dos tipos de esquema: el primero es un enfoque categórico el cual está basado en el concepto de emociones básicas tales como enojo, alegría, tristeza, etc. Y el segundo representa un enfoque continuo donde los estados emocionales se representan usando un espacio multidimensional continuo donde cada dimensión representa una emoción primitiva. En este enfoque utilizan 3 emociones primitivas: valencia, activación y dominación. Utilizaron la prueba piloto y determinaron 6 estados emocionales que fueron los más recurrentes en el habla de los niños: dudoso, molesto, motivado, nervioso, neutro y seguro. 11 evaluadores etiquetaron los datos utilizando la plataforma de avaluación *TRUE* (Planet, Iriondo, Martínez, & Montero, 2008).

Para el proceso de extracción de características extrajeron 6,552 características por instancia usando la herramienta OpenEAR las cuales son: energía, probabilidad de sonoridad, MFCC, energía espectral en bandas, flujo espectral, máximo y mínimo espectral, tasa de cruce por ceros, contorno F0, espectro MEL, punto rollof espectral, centroide espectral. Realizaron un proceso de selección de características. Para el enfoque categórico utilizaron un proceso basado en *Cfs Subset* usando un algoritmo genético como método de búsqueda. Para el enfoque continuo utilizaron un proceso basado en un algoritmo *Wrapper* que evalúa subconjuntos de características utilizando *Regression Support Vector Machine* y como método de búsqueda el *Linear Forward*. En la clasificación para el enfoque categórico utilizaron el algoritmo de *Support Vector Machine* y fue evaluado con *10-fold cross validation*. Para el enfoque continuo usaron el algoritmo *Support Vector Machine for Regression*. Para los procesos de selección de características y clasificación utilizaron la herramienta WEKA (Witten, Frank, & Hall, 2011).

## Resultados

De acuerdo a las pruebas que realizaron, utilizaron 7 emociones de la base de datos creada: indeterminado, dudoso, molesto, motivado, nervioso, neutro y seguro. La Tabla 3.1 muestra el número total de muestras de la base de datos, el total de características que se obtuvo después de realizar el proceso de selección de características para los dos enfoques y los resultados obtenidos en el proceso de clasificación para ambos enfoques. Siendo la unidad el 100% de datos clasificados correctamente.

| Instancias | Selec. Caract. | Categórica | V | A | D |
|---|---|---|---|---|---|
| 2,040 | 2,622/13/39/29 | 0.407 | 0.6146 | 0.7349 | 0.6744 |

*Tabla 3.1 Resultados de clasificación categórica y continua*

La tabla muestra que no se obtuvieron buenos resultados en el enfoque categórico, pero en el continuo los resultados fueron mejores, siendo la Valencia la más difícil de estimar y Activación la más fácil.

### 3.1.2 Agrupación jerárquica y clasificación de las emociones en el habla humana utilizando matrices de confusión (Reyes Vargas, y otros, 2013)

## Objetivo general

Esta investigación crea una base de datos emocional en idioma Español de México que contiene 7 emociones: disgusto, enojo, felicidad, miedo, neutral, sorpresa y tristeza. Hacen uso de las estrategias jerárquicas para la tarea de clasificación. Proponen un método para la elección de niveles jerárquicos que les permitan agruparlas en una matriz de confusión.

## Metodología

En la obtención de muestras de voz esta investigación no recurre a las diferentes bases de datos que existen. Crearon una base de datos emotiva de español mexicano, dado que no encontraron ninguna base de datos de habla español mexicano. Para crear la base de datos un actor mexicano profesional grabó tres conjuntos de datos de voz. Cada conjunto tiene 40 palabras seleccionadas de la lista Swadesh para español (Swadesh lists for Spanish), y 40 frases que incluían cada palabra. Swadesh, originalmente ideado por el lingüista Morris, contiene palabras que están presentes en casi todos los idiomas y forman la base para la comunicación entre los seres humanos. Los juegos de palabras seleccionadas incluyen nombres (números, colores, animales, partes del cuerpo, etc.), pronombres y verbos. Esta base de datos está etiquetada con siete emociones: enojo, disgusto, miedo, felicidad, tristeza,

sorpresa y neutral. Los registros de estas muestras fueron realizadas por un orador profesional masculino mexicano.

En la siguiente etapa, extracción de características, se extraen 14 características de las muestras de audio de cada una de las emociones: 12 MFCC, frecuencia fundamental F0 y los coeficientes de la energía.

En la última etapa de la metodología aplicada en esta investigación, se construyen y realizan pruebas a un clasificador binario jerárquico para las siete clases de emociones registradas en su base de datos creada. Para realizar lo mencionado anteriormente primero realizan un primer clasificador: j48 que es una versión del algoritmo C4.5 de Quinlan. Utilizan las siete clases de emociones guardando los resultados en una matriz de confusión. Posteriormente aplican el método *Ward* a la matriz resultante. El resultado de este proceso es un dendrograma (diagrama de datos en forma de árbol que organiza los datos en subcategorías) que proporciona los niveles jerárquicos. A continuación utilizan la medida de la distancia euclidiana para determinar la similitud entre las clases. Por último utilizan las Máquinas Vectores de Soporte (SVM por sus siglas en inglés) como clasificadores binarios en cada conjunto.

### Resultados

Las pruebas se realizaron en la misma base de datos que desarrolló esta investigación sobre las siete emociones dando un margen de error del 33.59%.

## 3.2 Extracción de características

A continuación se enlistan trabajos que se enfocan en encontrar el mejor grupo de características de voz para la detección de emociones.

### 3.2.1 Reconocimiento de emociones en el habla (Echeverry Correa & Morales Pérez, 2008)

### Objetivo general

Esta investigación utiliza una base de datos emocional en idioma Español que contiene 5 emociones: alegría, enojo, neutral, sorpresa y tristeza. Realizan un análisis en el dominio temporal y un análisis acústico empleando los Coeficientes Cepstrales en la Frecuencia de Mel (MFCC por sus siglas en inglés). Utilizan un clasificador basado en la teoría de Bayes para evaluar el sistema.

## Metodología

Primero toman las muestras de la base de datos *Spanish Emotional Speech* (SES) (Montero, Gutiérrez, Palazuelos, Enríquez, Aguilera, & Pardo, 1998) el cual tiene etiquetados cinco estados emocionales: felicidad, enojo, tristeza, sorpresa y neutral. Fue grabada por un actor profesional que simulaba las emociones. Posteriormente realizan un ventaneo en donde dividen la señal de voz en intervalos de 30 ms. Luego en la caracterización, que es la parte más significativa de este trabajo, emplearon dos técnicas para extracción de características en la señal de voz: datos crudos (*raw data*) y MFCC.

Con los datos crudos (*raw data*) obtienen características de la voz como la intensidad, duración, acentos, pausas y calidad de voz. Con esas características calculan funciones estadísticas como: media, mediana, máximo, mínimo, desviación estándar, varianza, asimetría y curtosis. También obtienen parámetros como son: perturbación de la amplitud, perturbación de la amplitud máxima y coeficiente de amplitud. Con MFCC trabajan únicamente con la frecuencia y esto les permite obtener los Coeficientes Cepstrales en la Frecuencia de Mel la cual calculan con la función mfcc.m del Auditorytoolbox de Matlab®. Utilizan la primera y segunda derivada de los coeficientes para obtener parámetros estadísticos.

En la etapa de clasificación utilizan un clasificador basado en la teoría de Bayes para entrenar un sistema que es capaz de realizar el reconocimiento automático de emoción. Utilizan el método de validación cruzada *leave-one-out,* el cual toma una señal de voz que es utilizada de prueba y las restantes de entrenamiento.

## Resultados

Realizan tres tipos de pruebas utilizando tres subconjutnos de características y las 5 emociones del corpus SES. Los subconjuntos son: características *raw data*, los MFCC y una combinación de ambas. En todos los conjuntos se empleó el clasificador basado en la teoría de Bayes. La Tabla 3.2 muestra que realizando una combinación de características de *raw data* y MFCC arroja una buena tasa de reconocimiento.

| **Características** | **Eficiencia (%)** |
|---|---|
| Raw data | 80.66 |
| MFCC | 88 |
| Raw data y MFCC | 94 |

*Tabla 3.2. Resultado empleando diferentes conjuntos de características*

### 3.2.2 Reconocimiento automático de emociones en el habla en lenguaje Serbio (Bojanić & Delić, 2013)

#### Objetivo general

Esta investigación utiliza una base de datos emocional en idioma Serbio que contiene 5 emociones: alegría, enojo, miedo, neutral y tristeza. Comparan grupos de características de la voz para conocer cuáles son las mejores y formar un vector de características eficientes para el reconocimiento de emociones. Para validar estas características utilizan 3 modelos de clasificación: *Linear Discriminant Classifier* (LDC), *k-Nearest Neighbors* (kNN) y *Neural Network*. Los resultados de los 3 modelos se comparan y se observa cual es el mejor para realizar esta tarea.

#### Metodología

Para la primera etapa obtienen muestras de audio de la base de datos "*Govorna Ekspresija Emocija i Stavova*" (GEES) (Jovičić, Kašić, Djordjević, & Rajković, 2004). Las muestras se encuentran grabadas en idioma Serbio y etiquetadas en 5 estados emocionales: felicidad, enojo, miedo, tristeza y neutral.

En la etapa de caracterización se obtienen las características de cada una de las muestras que se van ingresando al sistema. Las dividen en 3 grupos: en el primer grupo incluye características prosódicas entre las cuales hay 12 funciones estadísticas aplicadas en valores del tono y energía. El segundo grupo incluye características espectrales de las cuales se toman 12 MFCC, sus primeras derivadas y 12 funciones estadísticas aplicadas en cada una de ellas. Para el tercer grupo incluyen características espectrales: 12 MFCC, tono, sonoridad y energía, de estas características se aplican 12 funciones estadísticas.

Para la última etapa esta investigación utiliza 3 modelos de clasificación: LDC, k-NN y *Neuronal Network*. Para el LDC se consideraron dos casos: *Linear Bayes* y *Perceptron Rule*. En el caso del modelo de kNN, k=9 es el dato utilizado ya que reflejó el mejor resultado al momento de realizar las pruebas. Y para las redes neuronales se hace uso de dos algoritmos: El algoritmo de *BackPropagation* (BP) y el algoritmo de *Levenberg-Marquardt* (LMBP).

#### Resultados

En la fase de pruebas se toman todos los datos y las 5 emociones del corpus. La investigación únicamente muestra los resultados del tercer grupo de características el cual contiene: 12 MFCC, tono, sonoridad y energía. Dado que fue el mejor resultado que se obtuvo de los tres grupos. La Tabla 3.3 muestra los resultados de aplicar los clasificadores: k-NN, *Linear Bayes* y *Perceptron Rule* con el tercer grupo de características, además, se realiza otro caso de prueba aplicando reducción de dimensionalidad al grupo de características y se observa que los resultados son mejores.

| Clasificador | Eficiencia (%) | Eficiencia (%) aplicando reducción de dimensionalidad |
|---|---|---|
| k-NN | 36.96 | 91.26 |
| Linear Bayes | 91.5 | 91.52 |
| Perceptron rule | 89.6 | 90.5 |

*Tabla 3.3. Aplicando 3 tipos de clasificadores al tercer grupo de características*

Para el caso de los algoritmo de *Neuronal Network* utilizando el tercer grupo de características se obtiene para el BP un 90.4% de acierto y para LMBP un 82%. Se observa entonces que con el tercer grupo de características aplicando reducción de dimensionalidad y utilizando el clasificador *Linear Bayes* se obtiene el mejor de resultado.

### 3.2.3 Reconocimiento automático de voz emotiva con memorias asociativas Alfa-Beta – SVM (Solís Villarreal, Yáñez Márquez, & Suárez Guerra, 2011)

#### Objetivo general

Esta investigación utiliza una base de datos emocional en idioma Alemán que contiene siete emociones: aburrimiento, alegría, disgusto, enojo, miedo, neutral y tristeza. Proponen un nuevo modelo asociativo para la tarea de clasificación. El modelo está basado en las máquinas asociativas *Alfa-Beta Support Vector Machine* (SVM). El algoritmo *Alfa-Beta SVM* utiliza dos fases: la fase de aprendizaje que utiliza el operador α y la fase de recuperación que usa el operador β.

#### Metodología

Primero toman muestras de la base de datos de Berlín en idioma Alemán (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005). Fue grabado por 10 actores profesionales donde cinco son mujeres y cinco hombres. Tiene etiquetadas siete emociones: felicidad, enojo, miedo, aburrimiento, tristeza, disgusto y neutral. Contiene 535 instancias.

Esta investigación obtiene dos conjuntos de características. En la primera usan los parámetros clásicos de voz: 95 parámetros de energía, frecuencia, duración de los silencios, las 4 primeras formantes y 13 MFCC's. Posteriormente este modelo utiliza el método Encadenamiento hacia adelante para la selección de características, el cual ayuda a mejorar el desempeño del clasificador. Después de realizar estos procesos se utiliza el software WEKA el cual selecciona 14 parámetros y estos datos son los que pasan a la siguiente fase que es la del clasificador. El segundo conjunto solo utiliza el parámetro de energía, porque se observa que este parámetro es el que más información afectiva extrae de la voz. Se representa de forma

bidimensional la energía. Esta representación se realiza en forma de matriz para que sirva de entrada para la fase de clasificación.

### Resultados

En la fase de pruebas se utilizan las siete emociones que proporcionan los datos. Con el primer conjunto de características se compara el desempeño de tres modelos de clasificación que presenta WEKA con el de las Alfa-Beta SVM. La Tabla 3.4 muestra los resultados donde se observa que el modelo que proponen supera a los otros tres modelos.

| Modelo | Instancias correctas | Instancias incorrectas | Eficiencia (%) |
|---|---|---|---|
| Naive Bayes | 327 | 208 | 61.12 |
| Simple Logistic | 427 | 108 | 79.81 |
| Perceptron Multicapa | 463 | 72 | 86.54 |
| Alfa-Beta SVM | 508 | 27 | 94.95 |

*Tabla 3.4. Desempeño de los modelos en WEKA y por el Alfa-Beta SVM*

Con el segundo conjunto de características solo se aplicó al modelo *Alfa-Beta SVM*, logró clasificar 506 instancias y obtener un desempeño del 94.5%. El resultado es muy cercano al obtenido en la primera prueba pero en este caso solo utilizaron una característica: la energía.

## 3.2.4 Un estudio en la búsqueda de las características del habla más discriminatorias en el orador dependiente del reconocimiento de emociones en el habla (Long Pao, Hsiang Wang, & Ji Li, 2012)

### Objetivo general

Esta investigación hace uso de una base de datos emocional China que contiene 6 emociones: aburrimiento, alegría, enojo, miedo, neutral y tristeza. Realizan un estudio para determinar cuáles son las características más discriminatorias a partir de un conjunto de 78 características. En el proceso de clasificación utilizan el *Gaussian Mixture Model* (GMM) y *Weighted Discrete – K-Nearest Neighbor* (WD-KNN).

### Metodología

Hacen uso del corpus MCEC2010 (Tang, Chu, Hasegawa-Johnson, & Huang, 2009) el cual es una base de datos emocional que fue grabada en idioma Chino mandarín por 34 actores profesionales. Del total de los actores 13 son hombres y 21 mujeres generando un total de

12,749 expresiones cortas. Esta base de datos contienen seis emociones básicas: aburrimiento, alegría, enojo, miedo, neutral y tristeza.

Primero realizan un pre-procesamiento el cual incluye: umbral de energía, pre-énfasis, enmarcado y ventaneo. Para el ventaneo utilizan la ventana de *Hamming* con un traslape del 50% para evitar perdida de información entre las ventanas. Posteriormente realizan la extracción de características donde obtienen 13 conjuntos de características que incluyen: formante, *Shimmer*, *Jitter*, Coeficiente de Predicción Lineal (LPC por siglas en inglés), Coeficiente Cepstral de Predicción Lineal (LPCC por siglas en inglés), Coeficiente Cepstral en la Frecuencia de Mel (MFCC por siglas en inglés), Delta MFCC (dMFCC), Delta-Delta MFCC (ddMFCC), Coeficientes de Potencia de Frecuencia Log (LFPC por siglas en inglés), Predicción Lineal Perceptual (PLP por siglas en inglés), Espectral Relativo PLP (RastaPLP por siglas en inglés), Log-Energía, y Tasa de Cruces por Ceros (ZCR por siglas en inglés). Todas las características se examinan para determinar cuál es el conjunto de características más discriminatorias para el reconocimiento de emociones en el habla. Dado que este proceso es de suma importancia para este tipo de sistemas. Y finalmente a los 13 conjuntos les aplican las siguientes funciones estadísticas: media, desviación estándar, máximo, mínimo y rango.

En el proceso de clasificación utilizan el GMM y el WD-KNN. Para evaluar el sistema adoptan la validación *Leave-one-out* y el conjunto de entrenamiento fijo.

## Resultados

Realizaron dos tipos de pruebas: dependiente del orador que sirve para encontrar las características más discriminatorias e independiente del orador. En la prueba de orador dependiente se aplicó el total de características para cada uno de los datos que contiene el corpus. La Tabla 3.5 muestra las características más discrimatorias en donde se usaron los dos clasificadores.

| Característica | GMM | WD-KNN |
|---|---|---|
| Media MFCC | 67% | 67% |
| Media RastaPLP | 67% | 59% |
| Media dMFCC | 67% | 62% |
| Media LPCC | 64% | 65% |
| Media ddMFCC | 61% | 54% |
| Media PLP | 61% | 59% |
| Media LPC | 58% | 53% |

*Tabla 3.5. Características más discriminatorias*

En la segunda prueba, independiente del orador, usaron el total de datos que contiene el corpus pero solo usando la característica Media MFCC. Dado que en la prueba anterior fue la

que dio mejor resultado. Utilizan los dos clasificadores para evaluar el sistema. La Tabla 3.6 muestra que clasificador otorgó mejor resultado.

| Clasificador | Tasa de reconocimiento |
|---|---|
| GMM | 48% |
| WD-KNN | 63% |

*Tabla 3.6. Resultado de clasificación usando la media de MFCC*

## 3.3 Modelos de clasificación

A continuación se listan los trabajos que se encuentran relacionados con la temática de Modelo de clasificación.

### 3.3.1 Un estudio sobre el reconocimiento de emociones en el habla basado en CCBC y Red Neuronal (Han, Lun, & Wang, 2012)

#### Objetivo general

Esta investigación hace uso de una base de datos emocional China que contiene 7 emociones: alegría, disgusto, enojo, miedo, neutral, sorpresa y tristeza. Primero realizan un pre-procesamiento a la señal de audio, después, realizan la extracción de características. Posteriormente incorporan el *Canonical Correlation Based on Compensation* (CCBC) para mejorar el desajuste entre el conjunto de entrenamiento y prueba. Finalmente realizan la clasificación usando *Back-Propagation Neural Networks* (BPNN).

#### Metodología

Hacen uso de una base de datos emocional que fue grabada en lenguaje Chino a 16 kHz debajo de los 10dB por siete actores. Contiene siete estados emocionales: alegría, disgusto, enojo, miedo, neutral, sorpresa y tristeza.

Tomando cada una de las muestras de audio con las que cuenta la base de datos China primero realizan un pre-procesamiento el cual realiza: pre-filtrado, cuantificación, pre-énfasis y detector de punto. Posteriormente se realiza el proceso de extracción de características donde se obtiene un total de 53. El cual se divide en dos tipos de características: prosódica y de calidad.

- Características prosódicas. Un conjunto de 37 características obtienen de este tipo los cuales son: modelo logarítmico F0 y energía. En conjunto con los siguientes datos estadísticos: máximo, mínimo, posición máximo y mínimo, media, desviación estándar, coeficientes de regresión, error de media cuadrática para coeficientes de

regresión, número de regiones sonoras y sordas, número de tramas sonoras y sordas, región más larga sonora y sorda, relación del número de tramas sonoras y sordas, relación del número de regiones sonoras y sordas, relación del número de tramas sonoras vs total de tramas, relación de número de regiones sonoras vs total de regiones. También se extraen las siguientes características: *Jitter*, tremor y derivadas estadísticas de *pitch*.

- Características de calidad. Un conjunto de 16 características obtienen para este tipo: los primeros tres formantes, sus anchos de banda, *harmonic to noise ratio*, distribución de energía espectral, relación de energía sonora a sorda y flujo glotal. Este conjunto de características se extrajo usando la herramienta PRAAT (Boersma & Weenink).

Después de realizar la extracción de características incorporan el CCBC. Dado que en ambientes naturales es difícil controlar los entornos acústicos lo cual provoca desajuste en los conjuntos de entrenamiento y pruebas. El desajuste de estos dos conjuntos se debe a: diferencias de los oradores, cambios de canal de grabación y ruido en el ambiente. CCBC reconstruye la correlación correcta entre los vectores de entrenamiento y pruebas por lo que compensa el desajuste entre ambos conjuntos.

En el proceso de clasificación utilizan el *Back-Propagation Neural Networks* que hace uso del *Multiple-Layer Perception* (MLP) como *framework* del sistema y el algoritmo *Back-Propagation* como regla de entrenamiento.

## Resultados

Las pruebas las realizaron utilizando las 7 emociones que contiene la base de datos emocional China donde 100 instancias por emoción se usaron para entrenamiento y 100 instancias por emoción para pruebas. La Tabla 3.7 muestra los resultados realizando dos tipos de pruebas: en la primera se realiza todo el proceso sin realizar el CCBC y la segunda se incluye el CCBC.

| Emoción | BPNN | CCBBC + BPNN |
|---|---|---|
| Alegría | 71 | 80 |
| Disgusto | 88 | 91 |
| Enojo | 71 | 78 |
| Miedo | 70 | 75 |
| Neutral | 85 | 94 |
| Sorpresa | 79 | 82 |
| Tristeza | 80 | 82 |

*Tabla 3.7. Resultados de clasificación usando CCBC*

Los resultados muestran que agregando el CCBC al sistema se obtienen un total de 83.14% de instancias clasificadas correctamente, mientras que descartándola se obtiene 77.71%

### 3.3.2 Reconocimiento de emociones en el habla (Albornoz, Crolla, & Milone, 2008)

#### Objetivo general

Esta investigación utiliza una base de datos emocional en idioma Alemán que contiene siete emociones: aburrimiento, alegría, disgusto, enojo, miedo, neutral y tristeza. Estudian dos modelos para el proceso de clasificación: *Gaussian Mixture Model* (GMM) y *Hidden Markov Model* (HMM).

#### Metodología

Obtienen muestras de audio de la base de datos de Berlín en idioma Alemán (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005). Fue grabado por 10 actores profesionales donde cinco son mujeres y cinco hombres. Tiene etiquetadas 7 emociones en las muestras: felicidad, enojo, miedo, aburrimiento, tristeza, disgusto y neutral. Contiene 535 instancias.

En el proceso de extracción de características obtienen características prosódicas. Parametrizan señales de voz usando los 12 MFCC junto con su primera y segunda derivada. Cada muestra es segmentada en ventanas de *Hamming* de 25 ms con un traslape de 10ms. En la siguiente etapa, la clasificación, utilizan dos modelos: GMM y HMM. Para el GMM utilizan 22 componentes, aumentando en dos componentes cuando son requeridas por el modelo y para el modelo HMM definen un modelo de dos estados. También utilizando 22 componentes gaussianos. Se aumentaron también dichos componentes cuando son requeridos por el modelo. Con la ejecución de 7 emociones, las GMM utilizan 32 componentes y los HMM utilizan el modelo de dos estados pero con 30 componentes gaussianos.

#### Resultados

Se realizan pruebas con los dos modelos. Primero con tres emociones: neutral, alegría y enojo. Para la segunda fase de prueba se utilizan sietes emociones: neutral, alegría, miedo, disgusto, tristeza, enojo y aburrimiento. La Tabla 3.8 muestra el desempeño para ambos clasificadores con diferente número de emociones.

| Clasificador | Eficiencia (%) con 3 emociones | Eficiencia (%) con 7 emociones |
|---|---|---|
| GMM | 93 | 67 |
| HMM | 97 | 76 |

*Tabla 3.8. Desempeño de GMM y HMM para 3 y 7 emociones*

Se observa que el HMM para ambos casos obtuvo un mejor desempeño que GMM y que los modelos se degradan a medida que se añaden más emociones para analizar.

### 3.3.3 Reconocimiento de emoción en el habla de la fusión de decisiones basado en la teoría de evidencia de DS (Kuang & Li, 2013)

#### Objetivo general

Esta investigación utiliza una base de datos emocional en idioma Alemán que contiene siete emociones: aburrimiento, alegría, disgusto, enojo, miedo, neutral y tristeza. Proponen la *Dempster-Shafer* (DS) *Evidence Theory* para ejecutar la fusión de decisiones de tres tipos de clasificadores: *Hidden Markov Model* (HMM), *Artificial Neural Network* (ANN) y un algoritmo hibrido de los dos clasificadores.

#### Metodología

En la primera etapa obtienen las muestras de audios. Hacen uso de la base de datos de Berlín (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005), la cual tiene etiquetadas siete emociones: felicidad, enojo, miedo, aburrimiento, tristeza, disgusto y neutral.

En la segunda etapa, caracterización, se extraen las características que dependen del clasificador que es utilizado para obtener un conjunto de características. En el caso de las redes neuronales se extrajeron diez tipos de características que incluyen: la media del *pitch*, la media de la primer frecuencia de formantes, la media de la energía a corto plazo y los diferentes alcances del *pitch*. Las características para el HMM se obtienen la primera y segunda derivada de la energía a corto plazo, la combinación de la primera y segunda derivada del pitch, 12 MFCC y 10 LPCC con lo cual formaron el vector de características. Para el algoritmo hibrido solo extraen el vector de características básicas.

En la etapa de clasificación, se desarrollan tres clasificadores: HMM, ANN y combinan los dos clasificadores anteriores para formar un sistema integrado. En la etapa de fusión que es la última aplicada en esta investigación, fusionan los 3 clasificadores que crearon en la etapa anterior, para ello, utilizan la Teoría de Evidencia de DS que es un método de fusión a nivel de decisión efectiva. Esta teoría permite combinar la evidencia de varias fuentes en este caso los resultados de los tres clasificadores y llegar a un grado de probabilidad que tiene en cuenta toda la evidencia posible. Los valores de probabilidad se asignan a conjuntos de posibilidades en lugar de eventos únicos.

#### Resultados

Las pruebas que se realizaron en esta investigación fue solo para cuatro tipos de emociones del corpus: disgusto, enojo, tristeza y sorpresa. La Tabla 3.9 muestra la eficiencia que contiene cada clasificador. Se observa que usando el método de evidencia DS se obtiene el mejor resultado que los demás clasificadores.

| Clasificador | Eficiencia (%) |
|---|---|
| HMM | 63.65 |
| ANN | 68.80 |
| HMM/ANN | 76.80 |
| DS | 83.66 |

*Tabla 3.9. Eficiencia de los clasificados utilizados*

## 3.4 Comparación de trabajos

En esta sección se describe una comparativa entre los trabajos relacionados. Los criterios que se utilizaron para realizar la comparación son:

- Características de voz: Este criterio determina cuales elementos de la voz obtienen en la etapa de caracterización. Nos muestra si usaron una característica o un conjunto de estas.
- Modelo de clasificación: Este criterio determina qué modelo de clasificador de patrón usaron en la etapa de clasificación.
- Idioma de las muestras: Este criterio muestra el idioma que tiene la base de datos emocional.
- Emociones de las muestras: Este criterio muestra el tipo de dato (natural, inducida o actuada) y las emociones que se encuentran etiquetadas en la base de datos que usaron para realizar las pruebas.
- Resultados obtenidos: Este criterio muestra el resultado obtenido de las pruebas. Dependerá el tipo de atributo que utilizan las investigaciones para mencionarlas.

A continuación en la Tabla 3.10 se presenta el resultado comparativo de los trabajos analizados.

| Trabajos | Características de voz | Modelo de clasificación | Idioma | Emociones | Resultado |
|---|---|---|---|---|---|
| **EmoWisconsin: Una base de datos de habla en niños emocional en el español de México** | Energía, probabilidad de sonoridad, MFCC, energía espectral en bandas, flujo espectral, máximo y mínimo espectral, tasa de cruce por ceros, contorno F0, espectro MEL, punto rollof espectral, centroide espectral | Maquina Vector de Soporte | Español (México) | Inducido: dudoso, indefinido, molesto, motivado, nervioso, neutro y seguro | Medida F: 0.407 |
| **Reconocimiento de emociones en el habla** | Combinando datos crudos, primera y segunda derivada del MFCC | Teorema de Bayes | Español (España) | Actuado: felicidad, enojo, tristeza, sorpresa y neutral | Exactitud: 94% |

| | | | | | |
|---|---|---|---|---|---|
| **Reconocimiento automático de emociones en el habla en lenguaje Serbio** | MFCC, tono, sonoridad y energía | Teorema de Bayes | Serbio | Actuado: felicidad, enojo, miedo, tristeza y neutral | Exactitud: 91.52% |
| **Un estudio sobre el reconocimiento de emociones en el habla basado en CCBC y Red Neuronal** | f0, energía, jitter, tremor, derivadas estadísticas del pitch y usando la herramienta PRAAT | Red Neuronal | Chino | Actuado: alegría, disgusto, enojo, miedo, neutral, sorpresa y tristeza | Exactitud: 83.14% |
| **Un estudio en la búsqueda de las características del habla más discriminatorias en el orador dependiente del reconocimiento de emociones en el habla** | Media MFCC | K vecino más cercano | Chino | Actuado: aburrimiento, alegría, enojo, miedo, neutral y tristeza | Exactitud: 63% |
| **Reconocimiento automático de voz emotiva con memorias asociativas Alfa-Beta – SVM** | Energía | Máquinas Vectores de Soporte | Alemán | Actuado: felicidad, enojo, miedo, aburrimiento, tristeza, disgusto y neutral | Exactitud: 94.5% |
| **Reconocimiento de emociones en el habla** | Primera y segunda derivada del MFCC | Modelo Oculto de Markov | Alemán | Actuado: felicidad, enojo, miedo, aburrimiento, tristeza, disgusto y neutral | Exactitud: 76% |
| **Agrupación jerárquica y clasificación de las emociones en el habla humana utilizando matrices de confusión** | MFCC, frecuencia fundamental y energía | Estrategia jerárquica | Español (México) | Actuado: felicidad, enojo, miedo, disgusto, tristeza, sorpresa y neutral | Tasa error: 33.59% |
| **Reconocimiento de emoción en el habla de la fusión de** | MFCC, LPCC, energía, frecuencia fundamental y | Teoría de evidencia de DS (HMM, ANN y | Alemán | Actuado: felicidad, enojo, tristeza y disgusto | Exactitud: 83.66% |

| | | | | | |
|---|---|---|---|---|---|
| **decisiones basado en la teoría de evidencia de DS** | frecuencia de formantes | HMM/ANN) | | | |

*Tabla 3.10. Tabla comparativa de los trabajos relacionados*

# Capítulo 4 Metodología de solución

En este capítulo se describe la metodología empleada para solucionar el problema del presente trabajo de tesis. El cual se divide en dos fases principales: Extracción y Clasificador.

## 4.1 Descripción general de la metodología de solución

En la Figura 4.1 se presenta la descripción general de la metodología de solución la cual se encuentra dividida en dos secciones.

Para realizar la extracción de características y entrenamiento de clasificadores probabilísticos fue necesario adquirir bases de datos emocionales. Las cuales contienen los archivos de audio necesarios para realizar las tareas posteriores. Una vez adquiridos los corpus, se inició el desarrollo del módulo de extracción. El cual está integrado por el módulo de ventaneo y extracción de cada una de las características que se implementa en esa tarea para obtener subconjuntos de características. Posteriormente se desarrolló el módulo del clasificador donde se realizaron pruebas a varios algoritmos de clasificación. Con base en los resultados se determinó el mejor subconjunto y clasificador.

Como resultado de las fases descritas, se realizó una aplicación de escritorio la cual por medio de un archivo de audio permite conocer el estado emocional al que pertenece dicho archivo. A continuación, se describen las actividades realizadas durante el desarrollo de este trabajo de investigación, basadas en la metodología de solución.

## 4.2 Muestras de audio

Para evaluar nuestro sistema de reconocimiento de emociones en el habla se necesitan bases de datos emocionales sobre las que testearlos. Cuanto mayor sea la diversidad de estas bases de datos más confiables serán los resultados obtenidos. Actualmente existe una variedad considerable de bases de datos emocionales con las que cuentan información de: idioma, locutores, emociones y el tipo de dato (natural, inducida, natural). En la investigación realizada por (Ververidis & Kotropoulos, 2006) muestran una colección de las principales bases de datos emocionales que existen para el reconocimiento de emociones.

Para las etapas de extracción de características, entrenamiento de los modelos de clasificación y su posterior evaluación, hacemos uso de dos bases de datos en idioma Español de México: Emo_voz.mx1 y EmoWisconsin. Dado que esta variedad del español es el objeto de estudio de esta investigación

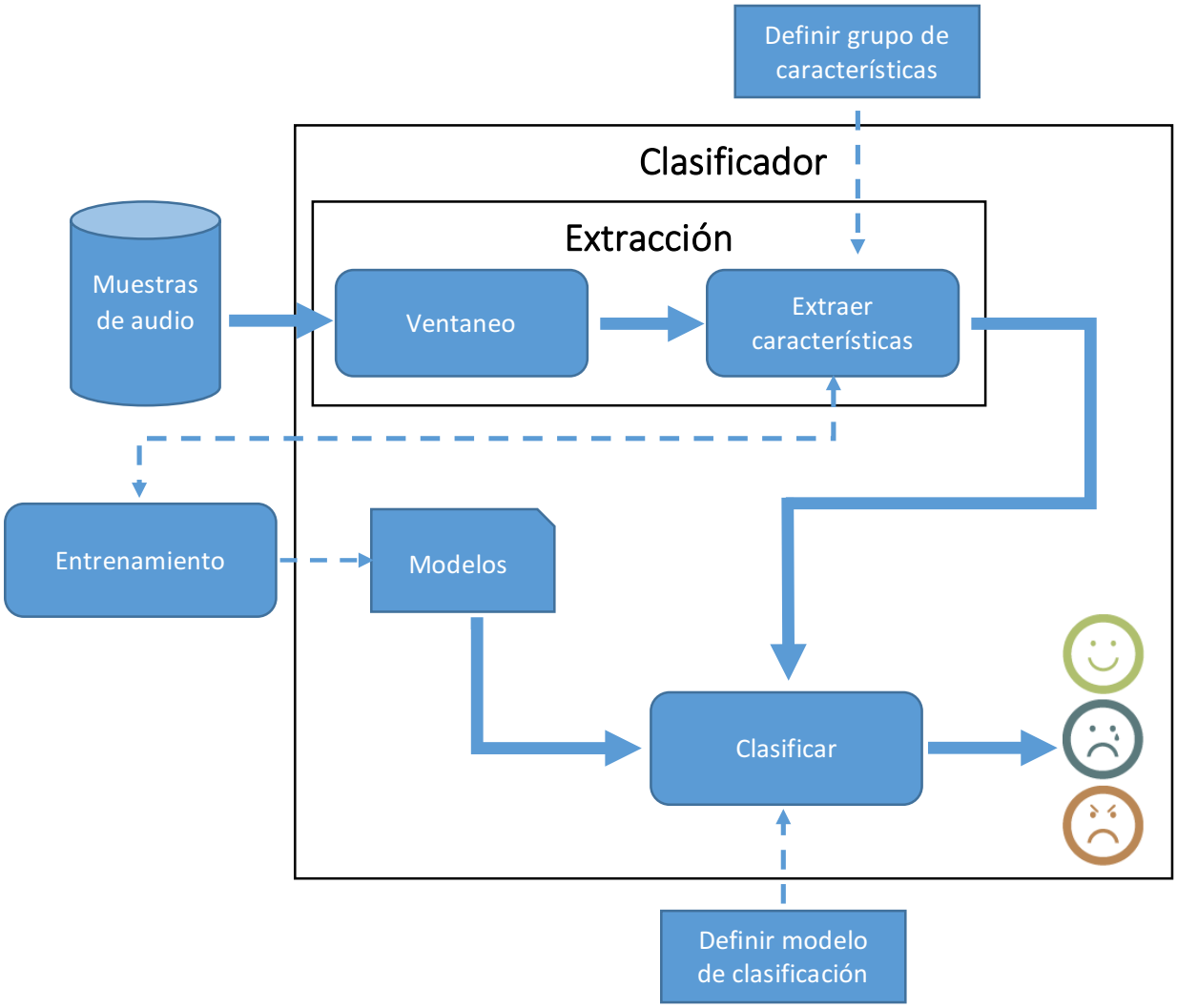


*Figura 4.1 Metodología de solución*

**Emo_voz.mx1** (Reyes Vargas, y otros, 2013)

Esta base de datos es propiedad del departamento de Ingeniería Eléctrica de la Universidad Autónoma Metropolitana (contacto: Fabiola Martínez Licona - fmml@xanum.uam.mx). En el subcapítulo 3.1.2 se habló de esta investigación pero en este apartado se explicará con más detalle lo que contiene este corpus.

Las oraciones se basaron en cada palabra de la lista y contienen la estructura completa: sujeto, verbo y predicado. El objetivo de utilizar estas frases era permitir que el actor se expresara mejor en términos de las emociones consideradas; esto fue acordado previamente con él. Los textos pertenecen a segmentos de la novela de Benito Pérez Galdos "Fortunata y

Jacinta", de la novela de Miguel de Cervantes Saavedra "La española inglesa" y el poema "Primer día" de Octavio Paz. Los textos tienen alrededor de 450 palabras, en promedio, y no contienen ninguna de diálogo; el poema tiene 94 palabras. Las grabaciones se llevaron a cabo en una PC de escritorio utilizando la herramienta Speech Filing System Versión 4.8 (Speech Filing System) con una frecuencia de muestreo de 16 KHz.

Esta base de datos contiene siete emociones: disgusto, felicidad, ira, miedo, neutro, sorpresa y tristeza. Tiene un total de 1541 instancias donde en la Tabla 4.1 se observa la distribución del total de instancias entre las siete emociones.

| **Emoción** | **Instancias** |
|---|---|
| Disgusto | 203 |
| Felicidad | 243 |
| Ira | 203 |
| Miedo | 203 |
| Neutro | 243 |
| Sorpresa | 203 |
| Tristeza | 243 |
| **Total** | **1541** |

*Tabla 4.1. Número de instancias para cada emoción del Corpus Emo_voz.mx1*

**EmoWisconsin** (Pérez Espinosa, Reyes García, & Villaseñor Pineda, 2011)

Esta base de datos es propiedad del Instituto Nacional de Astrofísica, Óptica y Electrónica (contacto: Carlos Alberto Reyes García – kargaxxi@inaoep.mx). En la subcapítulo 3.1.1 ya se dio una explicación de esta investigación en donde se menciona el proceso de grabación y herramientas utilizadas. Esta base de datos tiene un total de 2040 instancias donde en la Tabla 4.2 se observa la distribución del total de instancias entre las siete emociones que contiene.

| **Emoción** | **Instancias** |
|---|---|
| Indefinido | 237 |
| Inseguro | 515 |
| Molesto | 17 |
| Motivado | 72 |
| Nervioso | 267 |

| | |
|---|---|
| Neutro | 21 |
| Seguro | 911 |
| **Total** | **2040** |

*Tabla 4.2. Número de instancias para cada emoción en EmoWisconsin*

Nos contactamos con las personas responsables de las dos bases de datos para tener acceso a los datos. Se obtuvo respuesta positiva de ambas investigaciones y se realizaron los arreglos para la licencia y permitirnos el acceso a dichos archivos.

## 4.3 Extracción

En este módulo se inicia el proceso de extracción de características. Es muy importante en esta investigación dado que la buena calidad de dichas características proporcionará buenos resultados. Para este módulo los scripts se realizaron utilizando el lenguaje de programación Matlab. En la Figura 4.2 se muestre el proceso que se realiza en este módulo: primero se toman los archivos de audio de las bases de datos emocionales del módulo anterior y posteriormente se realizan los siguientes submódulos que se explican a continuación:

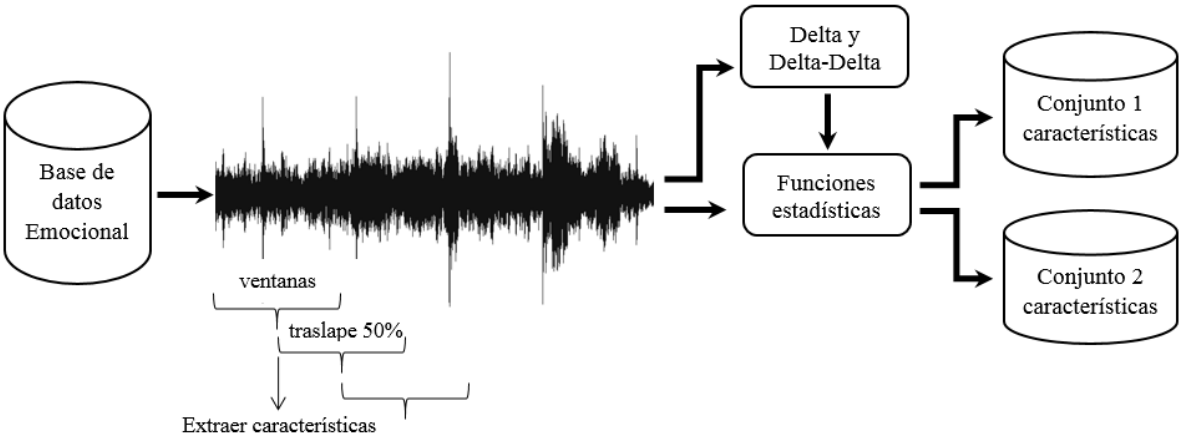


*Figura 4.2. Extracción de características*

### 4.3.1 Ventaneo

En este submódulo se realizan los cortes a cada uno de los archivos de audio que contienen las bases de datos emocionales. Se realizan tres etapas: primero se realizan cortes a la señal de audio por un rango determinado en milisegundos. Posteriormente se realiza el traslape, que mueve cada uno de los cortes por un rango también milisegundos. Por último, se aplican

funciones de ventaneo para suavizar los cambios que surgen al realizar el corte. El objetivo de este módulo es adecuar la señal para su posterior procesado.

#### 4.3.1.1 Ventanas

La señal de voz no es estacionaria, por lo que para procesar largos segmentos es necesario un método por el cual el audio se divida en una secuencia de segmentos cortos. A esto se le conoce como ventaneo de la señal de voz. Para realizarlo se asume un comportamiento estacionario en el periodo de duración de cada segmento. Se examina normalmente en intervalos cortos (entre 20 y 60 ms) donde las características permanecen invariantes (Ávila & Quintana, 1994). El número de segmentos depende de la longitud del archivo de audio. En la Figura 4.3, como ejemplo, se observa el ventaneo de un archivo de audio donde se realizan cortes de 30 ms.

Para esta investigación se tomaron los siguientes cinco rangos de ventanas: 20, 25, 30, 35 y 40 ms. El script de extracción de características para esta tarea obtiene los siguientes datos:

- Datos de muestreo y la frecuencia de los archivos de audio. Se hace uso de la función audioread de Matlab.
- Longitud del archivo de audio haciendo uso de la función length de Matlab en conjunto con los datos de muestreo.
- Longitud de la ventana multiplicando el rango de la ventana (por ejemplo 25 ms) por la frecuencia de muestreo del archivo de audio.

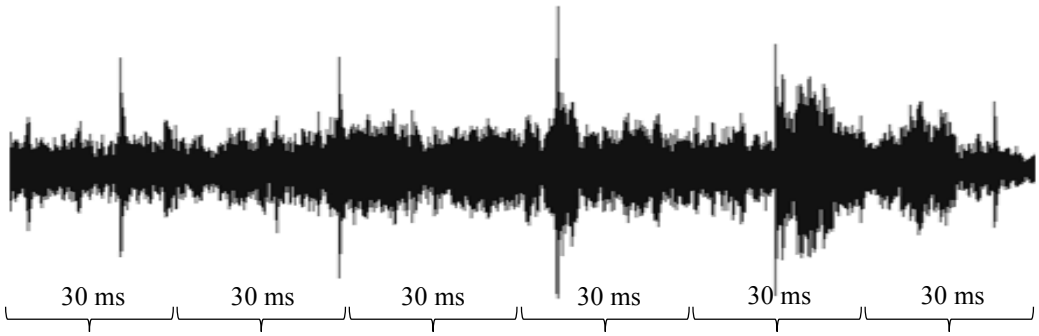


*Figura 4.3 Ventanas de 30 ms*

#### 4.3.1.2 Traslape

Cuando la señal de audio es cortada en segmentos es habitual tomar el siguiente bloque por una distancia adecuada, de esta manera, los bloques se interponen unos con otros. A esto se le conoce como traslape o solapamiento. El traslape tiene una gran importancia, ya que, garantiza

la correlación entre marcos adyacentes y minimiza la varianza espectral entre ellos. En la Figura 4.4 se observan tres ejemplos de solapamiento de 25%, 50% y 75%.

En este trabajo de investigación se tomó como medida de traslape el 50% de la longitud de la ventana. Como se mencionó los rangos de las ventanas son de 20, 25, 30, 35 y 40 ms por lo que las medidas de traslape son las siguientes: 10, 125, 15, 175 y 20 ms. Se obtienen los siguientes datos en el script de extracción de características:

- Salto de la ventana. Multiplicando la medida de traslape (por ejemplo 125 ms) por la frecuencia de los datos de muestreo del archivo de audio.
- El número total de segmentos a analizar. Esto se obtiene de la longitud del archivo de audio restando la longitud de ventana y posteriormente dividiendo el resultado con la medida del salto de la ventana.

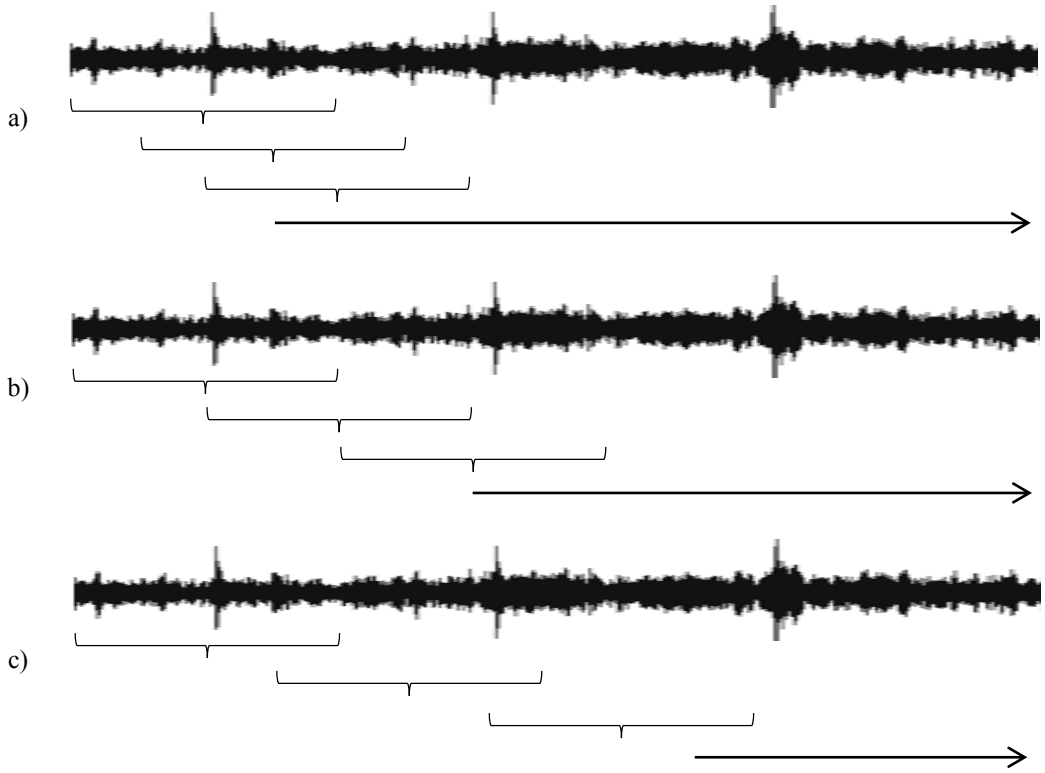


*Figura 4.4 Traslape a) 25 ms, b) 50 ms y c) 75 ms*

### 4.3.1.3 Función de ventaneo

Durante el proceso de ventaneo en bloques se producen desviaciones en el espectro de la señal. El efecto de discontinuidad motivado por el corte de la señal al inicio y al final de los tramos conlleva la presencia de componentes no deseados en el espectro. Para evitarlos es

habitual multiplicar cada bloque por una función que cambie suevamente desde valores cercanos a cero a un valor máximo y retroceda nuevamente a un valor cercano a cero. A este proceso se le conoce como "enventanado" (*windowing*).

El ventaneo Hamming es muy utilizado en investigaciones de procesamiento de audio cómo los realizados por (Long Pao, Hsiang Wang, & Ji Li, 2012) y (Albornoz, Crolla, & Milone, 2008). Pero existen otras funciones que realizan estas tareas. En la investigación de (Podder, Zaman Khan, Haque Khan, & Muktadir Rahman, 2014) realizan una comparación de tres funciones de ventaneo: Haming, Hanning y Blackman. Esa comparación la realizaron tomando la respuesta en fase, respuesta de magnitud, ancho de banda equivalente de ruido y respuesta en el dominio de tiempo y frecuencia. Para realizar una comparación de esas respuestas construyeron un filtro FIR (Finite Impulse Response) de paso bajo, paso alto, paso de banda y filtro eliminador de ventana. Con base en los resultados de ganancia del filtro FIR, mencionan que la función Blackman presenta un mejor desempeño que las otras dos funciones.

En las siguientes ecuaciones (Podder, Zaman Khan, Haque Khan, & Muktadir Rahman, 2014) se observa cada una de las funciones de las ventanas que se aplicaron en esta investigación, las cuales son: Blackman, Hamming y Hanning. Donde N representa la longitud de la ventana:

$$v(n) = a_0 - a_1 cos\left(\frac{2\pi n}{N-1}\right) + a_2 cos\left(\frac{4\pi n}{N-1}\right)$$

*Ecuación 2. Blackman donde a0=0.42, a1=0.5 y a2=0.08*

$$v(n) = a_0 - a_1 cos\left(\frac{2\pi n}{N-1}\right)$$

*Ecuación 3. Hamming donde a0=0.54 y a1=0.46*

$$v(n) = a_0 - a_1 cos\left(\frac{2\pi n}{N-1}\right)$$

*Ecuación 4. Hanning donde a0=0.5 y a1=0.5*

Como se mencionó en las tareas anteriores se tomaron cincos rangos de ventanas y traslapes, por lo que para cada uno de estas medidas se le aplicaron estas tres funciones de ventanas. En esta tarea se obtienen los siguientes datos:

- Ventana. Se obtiene ejecutando la función window de Matlab pasando como parámetros el nombre de la ventana (@blackman, @hamming y @hann) y la longitud de la ventana.
- Con el dato del total del número de segmentos se procede a realizar un análisis por medio de un ciclo donde se recorre cada uno de estos segmentos.

- Dentro del ciclo se obtiene el dato bloque (*frame*), que contiene la parte del audio a analizar. Esto se obtiene dependiendo del tamaño de la ventana. Para obtener el siguiente *frame* dentro del ciclo, se suma la posición actual con el dato de salto.
- El *frame* se multiplica con el dato ventana. Se procede a realizar la extracción de características de audio con el nuevo dato.

### 4.3.2 Extraer características

En este submódulo los bloques son procesados individualmente para obtener una nueva representación en forma de secuencia de vectores (matriz), uno por segmento. Estos valores se llaman coeficientes o parámetros. Cada vector contiene un número fijo de coeficientes que dependerá tanto de la frecuencia de muestreo como del tipo de parámetro utilizado. En la Figura 4.5 se muestra el esquema general del proceso de extracción de parámetros. El cuadro de la señal de audio se ira desplazando conforme se dé el salto de la ventana. N es el número total de segmentos a analizar y F es el total de características a extraer.

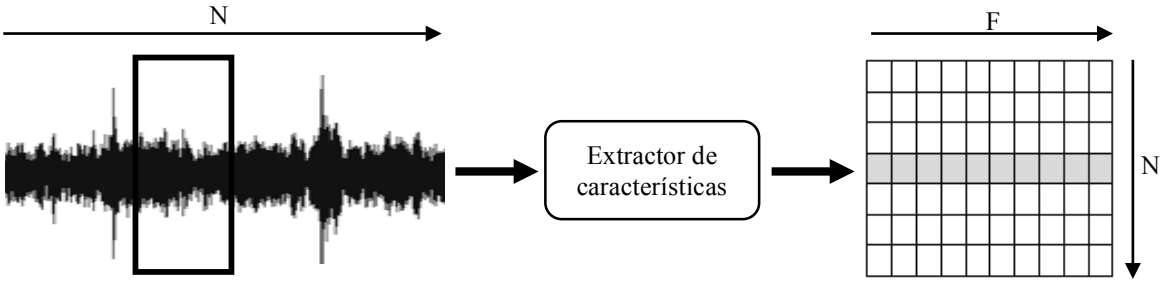


*Figura 4.5 Proceso de guardado de las características en una matriz*

En esta investigación, para cada bloque se calculan y se extraen 45 características. Las cuales se encuentran categorizadas en: dominio del tiempo, dominio de la frecuencia y prosódicas. Donde las características de dominio del tiempo y frecuencia se extrajeron tomando las funciones que implementa el Toolbox Audio Analysis Library (Giannakopoulos & Pikrakis, 2014). Para el cálculo de los formantes de frecuencia, ancho de banda y el speechrate se tomaron las funciones que implementa el SLT_Toolbox (Zhou, Yuan, Horta, Cai, Muraleedharan, & Kohl, 2013). A continuación, se enlistan las 45 características que se extraen en esta tesis:

a) Dominio-tiempo
    1. *Zero Crossing Rate*
    2. *Energy*
    3. *Entropy of energy*

b) Dominio-frecuencia
    4. *Spectral centroid*
    5. *Spectral Spread*
    6. *Spectral Entropy*
    7. *Spectral Flux*
    8. *Spectral Rollof*
    9. *MFCCS* (Primeros 13)
    10. *Chroma vector* (12 elementos)
    11. *Harmonic ratio*
    12. *Fundamental frequency*
c) Prosódicas
    13. *Volume*
    14. *Formants frequency* (primeros 4)
    15. *Formants bandwidth* (primeros 4)
    16. *SpeechRate*

Al terminar de realizar este submódulo se obtiene una matriz de 45xNúm_de_bloques. Las 45 columnas pertenecen al total de características que se extraen y el Núm_de_bloques depende de la longitud del archivo de audio. Dado que este dato es distinto de un archivo con duración de 50 segundos que uno de 1:35 min. Además del lenguaje Matlab también se utilizó el software de aprendizaje automático WEKA para realizar el proceso de selección de características. Para realizar dicha tarea se requiere crear archivos ARFF para procesar los datos. Para diferenciar una característica de otra en dicho archivo se nombró a cada característica con un nombre para que WEKA los tome como atributos. En la Tabla 4.3 se muestra el nombre de atributo que se le dio a cada característica en el archivo ARFF.

| Número | Características | Nombre en archivo ARFF |
|---|---|---|
| 1 | Zero Crossing Rate | ZCR |
| 2 | Energy | Energy |
| 3 | Entropy of Energy | EntropyOfEnergy |
| 4 | Spectral Centroid | SpectralCentroid |
| 5 | Spectral Spread | SpectralSpread |
| 6 | Spectral Entropy | SpectralEntropy |
| 7 | Spectral Flux | SpectralFlux |
| 8 | Spectral Rollof | SpectralRollof |
| 9 | MFCC | MFCC# |
| 10 | ChromaVector | ChromaVector# |

| | | |
|---|---|---|
| 11 | Harmonic Ratio | HarmonicRatio |
| 12 | Fundamental Frequency | Pitch |
| 13 | Volume | Volume |
| 14 | Formant Frequency | FormantFrequency# |
| 15 | Formant Bandwidth | FormantBandwidth# |
| 16 | SpeechRate | SpeechRate |

*Tabla 4.3. Identificadores de las características en el archivo ARFF.*

Para el nombre de algunas características como MFCC, *Chroma vector*, *Formant frequency* y *formant bandwidth* tienen el signo de numeral. Esto se debe a que estas características contienen más de un dato, por ejemplo, del MFCC se toman 13 datos, del *Chroma Vector* 12 y cada uno de estos datos se debe de identificar.

### 4.3.3 Delta y Delta-Delta

En este submódulo se determinó realizar dos conjuntos de características generales: datos con funciones Deltas y sin ellas. Estas funciones se aplican a las 45 características y muestra si la señal mantiene cambios durante un determinado tiempo. Solamente al segundo conjunto se le aplicaron las funciones Delta y Delta-Delta. El primer conjunto se salta este submódulo. Para el cálculo de las funciones Delta se aplicó la Ecuación 5 que se tomó del libro (Young, y otros, 2006). Donde w es la longitud de la ventana, por defecto tiene el valor de $w=2$ y $x^t$ es un contorno de datos (ventana de característica):

$$d^t = \frac{\sum_{i=1}^{w} i * (x^{t+i} - x^{t-i})}{2\sum_{i=1}^{w} i^2}$$

*Ecuación 5. Ecuación para el cálculo de funciones Delta*

Para mencionar este nuevo conjunto en el archivo ARFF solo se coloca el prefijo después del nombre de la característica. En la Tabla 4.4 se muestra un ejemplo mencionando a dos características.

| **Número** | **Delta** | **Delta-Delta** |
|---|---|---|
| 1 | ZCR_Delta | ZCR_Delta_Delta |
| 2 | Energy_Delta | Energy_Delta_Delta |

*Tabla 4.4. Identificadores de las características utilizando funciones Deltas en el archivo ARFF*

### 4.3.4 Funciones estadísticas

A partir de este submódulo todas las tareas se aplican a los dos conjuntos de características (con funciones Deltas y sin ellas). Para realizar un mejor análisis y llevar los datos a una mejor representación se aplicaron 24 funciones estadísticas para cada una de las características a excepción del Speechrate. Dado que este dato no se obtiene un vector de datos al ser un valor único. A continuación se listan las funciones estadísticas aplicadas:

1. Máximo
2. Mínimo
3. Mediana
4. Media aritmética
5. Media cuadrática
6. Media geométrica
7. Varianza
8. Desviación estándar
9. Rango
10. Simetría
11. Curtosis
12. Moda
13. Percentil 95%
14. Percentil 98%
15. 3 cuartiles
16. Inter-cuartil 2-1
17. Inter-cuartil 3-2
18. Inter-cuartil 3-1
19. Número de picos
20. Media aritmética de picos
21. Media aritmética de distancia entre picos
22. Media aritmética de distancia entre picos – media aritmética de picos

Continuando con la creación de nombre de atributos que tienen los características en el archivo ARFF para que WEKA procese los datos. En la Tabla 4.5 se muestra el nombre de atributo tomando como ejemplo la característica Zero Crosing Rate en conjunto con las 24 funciones estadísticas. Para el caso de las características Delta y Delta-Delta solo se incluye el nombre de dichas funciones antes de mencionar a las funciones estadísticas.

| Número | Estadística | Nombre en archivo ARFF |
|---|---|---|
| 1 | Máximo | ZCR_maximo |
| 2 | Mínimo | ZCR_minimo |
| 3 | Mediana | ZCR_median |
| 4 | Media Aritmética | ZCR_mean |
| 5 | Media Cuadrática | ZCR_mean_qua |
| 6 | Media Geométrica | ZCR_mean_geo |
| 7 | Varianza | ZCR_variance |
| 8 | Desviación Estándar | ZCR_std |
| 9 | Rango | ZCR_range |
| 10 | Simetría | ZCR_skewness |
| 11 | Curtosis | ZCR_kurtosis |
| 12 | Moda | ZCR_mode |
| 13 | Percentil 95% | ZCR_percentil95 |
| 14 | Percentil 98% | ZCR_percentil98 |
| 15 | Cuartil | ZCR_quartil# |
| 16 | Inte-cuartil 2-1 | ZCR_iqr12 |
| 17 | Inte-cuartil 3-2 | ZCR_iqr23 |
| 18 | Inte-cuartil 3-1 | ZCR_iqr13 |
| 19 | Número de Picos | ZCR_peakNum |
| 20 | Media Aritmética de Picos | ZCR_peakMean |
| 21 | Media Aritmética de distancia entre picos | ZCR_peakMeanDist |
| 22 | Media Aritmética de distancia entre picos – media aritmética de picos | ZCR_peakMeanMeanDist |

*Tabla 4.5. Identificadores de las características utilizando funciones estadísticas en el archivo ARFF*

### 4.3.5 Conjunto uno de características

Para este conjunto solo se toman las 45 características aplicando las 24 funciones estadísticas. Esto da un total de 1057 características por archivo de audio y un total de 15 conjuntos utilizando las tres funciones de ventaneo en conjunto con los cinco rangos de ventanas. La

matriz de características resultante para el corpus Emo_voz.mx1 tiene un rango de 1541x1057 y para el corpus EmoWisconsin 2040x1057. Estas matrices servirán para el proceso de selección de características. En la Figura 4.6 se muestra un ejemplo de un archivo ARFF extraído de una de las bases de datos emocionales. En dicha figura no se muestra la totalidad de los atributos dado por la cantidad de los mismos.

```
@attribute FormantBandwidth02_peakMeanMeanDist numeric
@attribute FormantBandwidth03_maximo numeric
@attribute FormantBandwidth03_minimo numeric
@attribute FormantBandwidth03_median numeric
@attribute FormantBandwidth03_mean numeric
@attribute FormantBandwidth03_variance numeric
@attribute FormantBandwidth03_std numeric
@attribute FormantBandwidth03_range numeric
@attribute FormantBandwidth03_skewness numeric
@attribute FormantBandwidth03_kurtosis numeric
@attribute FormantBandwidth03_mode numeric
@attribute FormantBandwidth03_mean_qua numeric
@attribute FormantBandwidth03_mean_geo numeric
@attribute FormantBandwidth03_percentil95 numeric
@attribute FormantBandwidth03_percentil98 numeric
@attribute FormantBandwidth03_quartil1 numeric
@attribute FormantBandwidth03_quartil2 numeric
@attribute FormantBandwidth03_quartil3 numeric
@attribute FormantBandwidth03_iqr12 numeric
@attribute FormantBandwidth03_iqr23 numeric
@attribute FormantBandwidth03_iqr13 numeric
@attribute FormantBandwidth03_peakNum numeric
@attribute FormantBandwidth03_peakMean numeric
@attribute FormantBandwidth03_peakMeanDist numeric
@attribute FormantBandwidth03_peakMeanMeanDist numeric
@attribute FormantBandwidth04_maximo numeric
@attribute FormantBandwidth04_minimo numeric
@attribute FormantBandwidth04_median numeric
@attribute FormantBandwidth04_mean numeric
@attribute FormantBandwidth04_variance numeric
@attribute FormantBandwidth04_std numeric
@attribute FormantBandwidth04_range numeric
@attribute FormantBandwidth04_skewness numeric
@attribute FormantBandwidth04_kurtosis numeric
@attribute FormantBandwidth04_mode numeric
@attribute FormantBandwidth04_mean_qua numeric
@attribute FormantBandwidth04_mean_geo numeric
@attribute FormantBandwidth04_percentil95 numeric
@attribute FormantBandwidth04_percentil98 numeric
@attribute FormantBandwidth04_quartil1 numeric
@attribute FormantBandwidth04_quartil2 numeric
@attribute FormantBandwidth04_quartil3 numeric
@attribute FormantBandwidth04_iqr12 numeric
@attribute FormantBandwidth04_iqr23 numeric
@attribute FormantBandwidth04_iqr13 numeric
@attribute FormantBandwidth04_peakNum numeric
@attribute FormantBandwidth04_peakMean numeric
@attribute FormantBandwidth04_peakMeanDist numeric
@attribute FormantBandwidth04_peakMeanMeanDist numeric
@attribute SpeechRate numeric
@attribute emotion {disgusto,felicidad,ira,miedo,neutro,sorpresa,tristeza}

@data
0.578125,0.003125,0.071875,0.107680,0.016726,0.129330,0.575000,2.043077,6.515278,0.003125,0.168173,0.047118,0.467500,0.515562,0.028125,0.071875,0.115625,0.043750,0.043750,0.087500,6
0.000000,0.158802,0.118750,0.040052,0.009767,0.000006,0.000134,0.000744,0.000002,0.001477,0.009762,3.103796,13.709779,0.000006,0.168173,0.000125,0.004210,0.006061,0.000014,0.000134,
0.000638,0.000120,0.000504,0.000623,65.000000,0.001468,0.001347,0.000121,2.592908,1.031958,2.238318,2.153256,0.062427,0.249854,1.560950,-1.244756,4.433574,1.031958,0.168173,2.136861
```

*Figura 4.6. Archivo ARFF*

## 4.3.6 Conjunto dos de características

Para el segundo conjunto se toman los mismos datos pero antes de aplicar las funciones estadísticas se aplican las funciones Delta y Delta-Delta. Por cada archivo de audio se obtienen 3169 características y un total de 15 conjuntos utilizando las tres funciones de ventaneo y los cincos rangos. La matriz de características resultante para el corpus Emo_voz.mx1 tiene un rango de 1541x3169 y para el corpus EmoWisconsin 2040x3169. El obtener dos conjuntos se debe a que en el proceso de pruebas se determinará si las características que se extraen al aplicar las funciones Delta y Delta-Delta mejoran el resultado predictivo.

## 4.4 Definir grupo de características

Con las características extraídas de las dos bases de datos emocionales se procede a realizar este submódulo: la selección de atributos. Este proceso se realiza para obtener las características que aporten información relevante y desechar aquellas que sean redundantes e irrelevantes. En total se tienen 30 conjuntos de características: 15 del primer conjunto y 15 del segundo conjunto. Los métodos de selección de atributos disponibles en WEKA que se aplicaron en esta investigación son los siguientes:

- CfsSubsetEval
- WrapperSubsetEval. Este método requiere de un algoritmo de clasificación para realizar una mejor evaluación. Se realizaron pruebas para determinar que algoritmo implementar en las pruebas posteriores. Los algoritmos son los siguientes algoritmos: Multilayer Perceptron, Naive Bayes, RandomForest y SMO. Siendo este último el que mejor dio resultados.

Los métodos anteriores se aplicaron por el funcionamiento que utilizan para evaluar los subconjuntos, además, uno de ellos utiliza un algoritmo clasificador para realizar una mejor evaluación ya que toma como la tasa de error del clasificador para medir la calidad de los subconjuntos. Cada uno de estos métodos se aplica en conjunto con métodos de búsqueda para encontrar el mejor subconjunto en base a la calidad de estos que proporcionaron los métodos de evaluación. Los siguientes métodos de búsqueda se seleccionaron dado que las búsquedas que implementan son diferentes y esto nos permite tener resultados diferentes para realzar una mejor evaluación:

- BestFirst
- GeneticSearch
- LinearForwardSelection

En la Figura 4.7 se observa la interfaz gráfica del software WEKA que se utilizó en esta investigación para los procesos de selección de características y de clasificación. Se muestra un ejemplo de selección de características o atributos cargando un archivo ARFF del conjunto uno utilizando un rango de ventaneo de 20 ms sin utilizar las funciones Delta. Se mencionó que para este conjunto se obtuvo un total de 1057 características, pero al momento de realizar el archivo ARFF al final se agrega un nuevo atributo: las clases a las que pertenecen los datos. Este dato es clave dado que servirá para diferenciar un dato de una clase con otra y es de suma importancia para el proceso de entrenamiento.

Se selecciona el método de selección y el método de búsqueda para iniciar el proceso de selección de atributos. Al terminar este proceso se obtiene un número de características menor que depende de los métodos de selección.

Se obtienen seis subconjuntos por cada conjunto que aplica un rango y una función de ventaneo. Entonces se tiene 30 subconjuntos utilizando los cinco rangos (20, 25, 30, 35 y 40 ms) con una función de ventaneo y ejecutando con las tres funciones se obtiene un total de 90 subconjuntos. Pero solo para las características que no se aplicaron las funciones Delta. Se obtienen otros 90 subconjuntos utilizando las características con funciones Deltas. Al término de este módulo se tiene un total de 180 subconjuntos de características cada una con un número de diferentes de características.

La utilización de conjuntos con funciones Delta se debe a que en el proceso de pruebas se observará si ayuda a mejorar el proceso de predicción o no. Dado que en algunas investigaciones que se explicaron algunos casos esas características arrojan buenos resultados y en otras lo contrario. En la investigación de (Echeverry Correa & Morales Pérez, 2008) utilizando la función delta y doble delta solo con la característica MFCC arroja un buen resultado. En las investigaciones de (Long Pao, Hsiang Wang, & Ji Li, 2012) y (Albornoz, Crolla, & Milone, 2008) utilizan las funciones delta y doble delta en algunas características en conjunto con otras y arrojan un buen resultado. Pero en las investigaciones de (Kuang & Li, 2013) los resultados utilizando estas funciones no son las mejoras y en la de (Bojanić & Delić, 2013) no utilizan el conjunto de característica que las contiene al no dar un buen resultado en el proceso de pruebas.

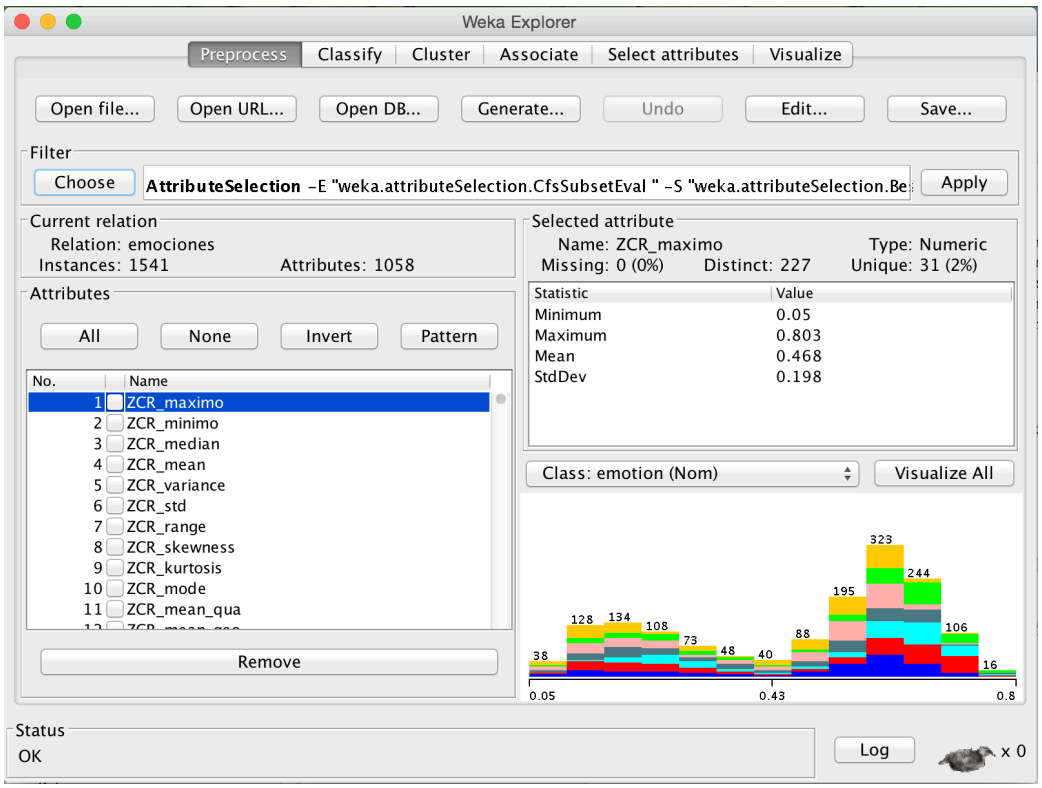


*Figura 4.7. Interfaz de WEKA para el proceso de selección de atributos*

## 4.5 Entrenamiento

Con los subconjuntos de características listos se realizó la segunda fase de la propuesta de solución. En este primer módulo de esta segunda fase se realiza el entrenamiento del clasificador. La utilización de herramientas de aprendizaje automático comprende dos pasos:

- Entrenamiento: Se parte de un conjunto de datos donde se conoce la clase y se utilizan para construir un modelo.
- Clasificación: una vez creado el modelo con los datos conocidos, se utiliza el modelo para clasificar nuevos datos para los que no se conoce la clase.

Para realizar la tarea de entrenamiento el conjunto total de datos se divide en dos subconjuntos disjuntos, uno de mayor tamaño que el otro. Este proceso se desarrolla en dos fases:

- Entrenamiento: construcción de un modelo de predicción utilizando el conjunto que tiene el mayor número de datos. Para este caso se utilizan los archivos ARFF que se crearon en la fase anterior.
- Prueba: con el modelo creado se comprueba su validez probándolo con el conjunto que resta. Los resultados se obtienen comparando las predicciones realizadas por el modelo con el valor real del atributo de etiqueta.

Con la etiqueta nos referimos a la clase que representa los valores. El modelo resultante del proceso de entrenamiento se denomina modelo de clasificación. El proceso de entrenamiento consiste en crear automáticamente un modelo a partir de un conjunto de entrenamiento y de un inductor. En la Figura 4.8 se observa dicho proceso.

- Conjunto de entrenamiento: datos utilizados para realizar el entrenamiento. En el conjunto total los datos se encuentran etiquetados con la clase a la que pertenece, pero solo una parte se toma para realizar esta tarea. Dependiendo de la implementación serán los datos a tomar para la clasificación.
- Inductor: algoritmo de clasificación que construye automáticamente un modelo de clasificación a partir del conjunto de entrenamiento.

El modelo resultante consiste en una serie de patrones o reglas que se utilizan para distinguir las clases.

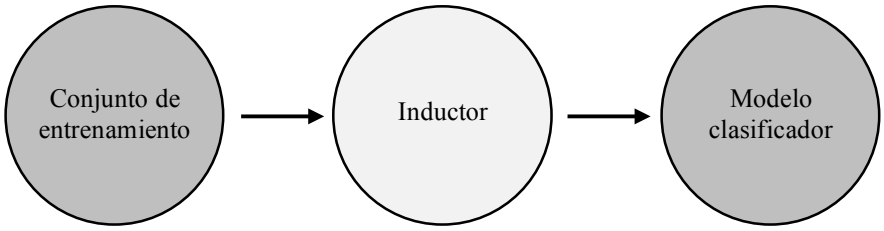


*Figura 4.8 Creación de un modelo clasificador*

Para la evaluación de las pruebas que realizó el entrenamiento se utilizó la función *Cross fold Validation* de WEKA. En la Figura 4.9 se observa un ejemplo de cómo funciona método de validación de resultados Hay un dato importante en esta función que se llama "fold" el cual divide el conjunto de características en subconjuntos. En el ejemplo de la figura se utilizan cuatro folds por lo que el conjunto principal se divide en cuatros subconjuntos: tres se utilizan para realizar el entrenamiento y uno se utiliza para realizar las pruebas.

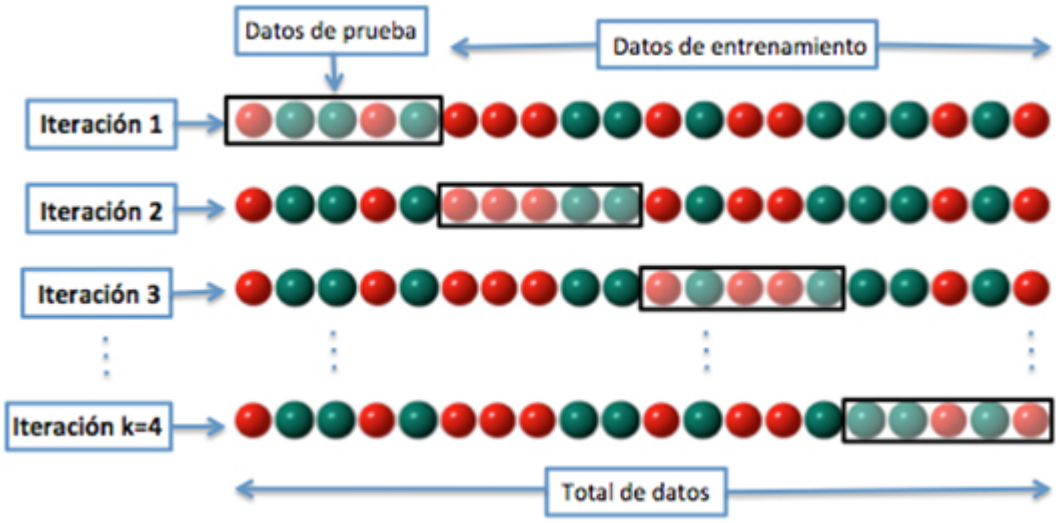


*Figura 4.9. Cross validation 4 fold*

Al término de ese entrenamiento se toma otro subconjunto para realizar las pruebas y los otros tres restantes para entrenamiento. Este proceso se repite hasta que los cuatro subconjuntos sean utilizados para entrenamiento y de igual forma para pruebas. Esto se realiza para tener un mejor resultado de predicción al tomar todos los datos para entrenamiento y pruebas.

En esta investigación se utiliza el dato 10 *folds*. Por lo cual nuestro conjunto de características se divide en 9 subconjuntos de entrenamiento y uno de prueba. En la Figura 4.10 se observa un ejemplo de entrenamiento utilizando un clasificador: primero se carga el archivo ARFF que se obtiene al término del módulo de extracción de características. Posteriormente como el proceso es para entrenar un clasificador se selecciona un clasificador en este caso la función "SMO". Se elige el tipo de prueba para el entrenamiento el cual es 10 *fold cross validation* y se inicia el proceso de entrenamiento de ese clasificador

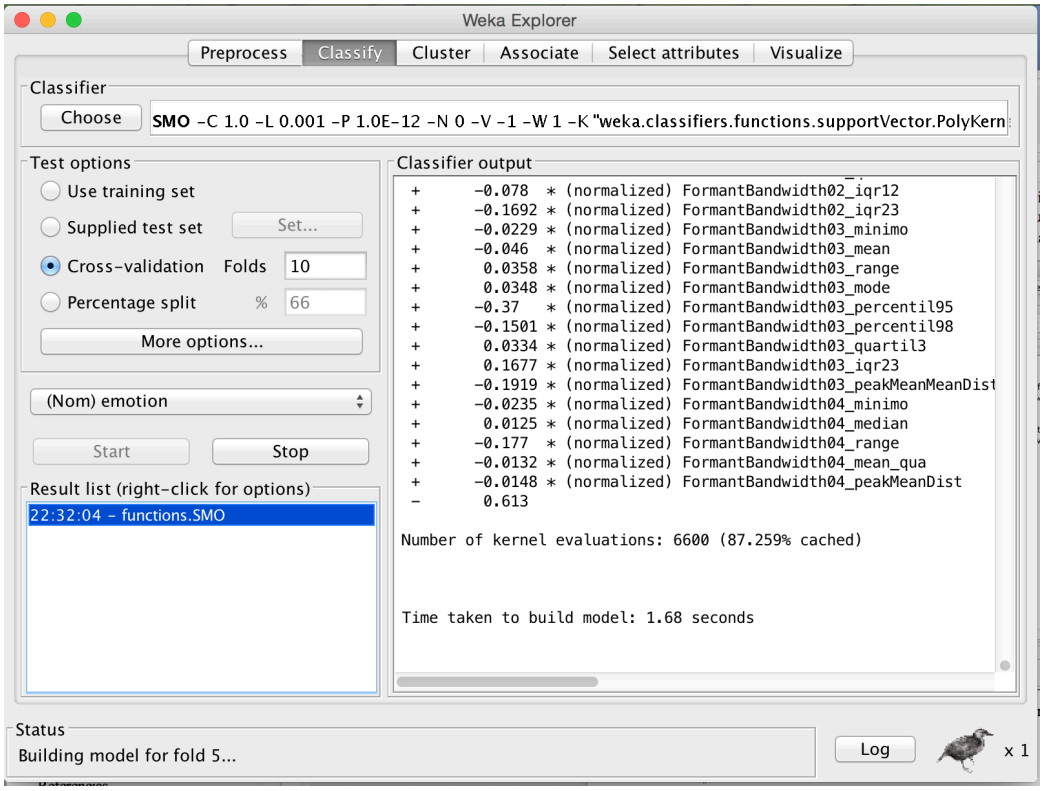


*Figura 4.10. Interfaz WEKA entrenando al clasificador SMO*

Al terminar el proceso, WEKA nos muestra los resultados. Nos indica que tan eficiente resultó el clasificador utilizado. Nos muestra la precisión, exhaustividad, la medida-F, entre otros datos para cada clase y de forma general. Posteriormente de manera gráfica nos muestra la cantidad de instancias clasificadas en cada clase para ver el margen de error. Esto se observa en la Figura 4.11.

## 4.6 Clasificador

En este módulo se realiza la clasificación una vez que se termina el entrenamiento. Se obtiene un modelo que se genera a partir de un algoritmo clasificador. Estos se encargan de clasificar nuevas instancias.

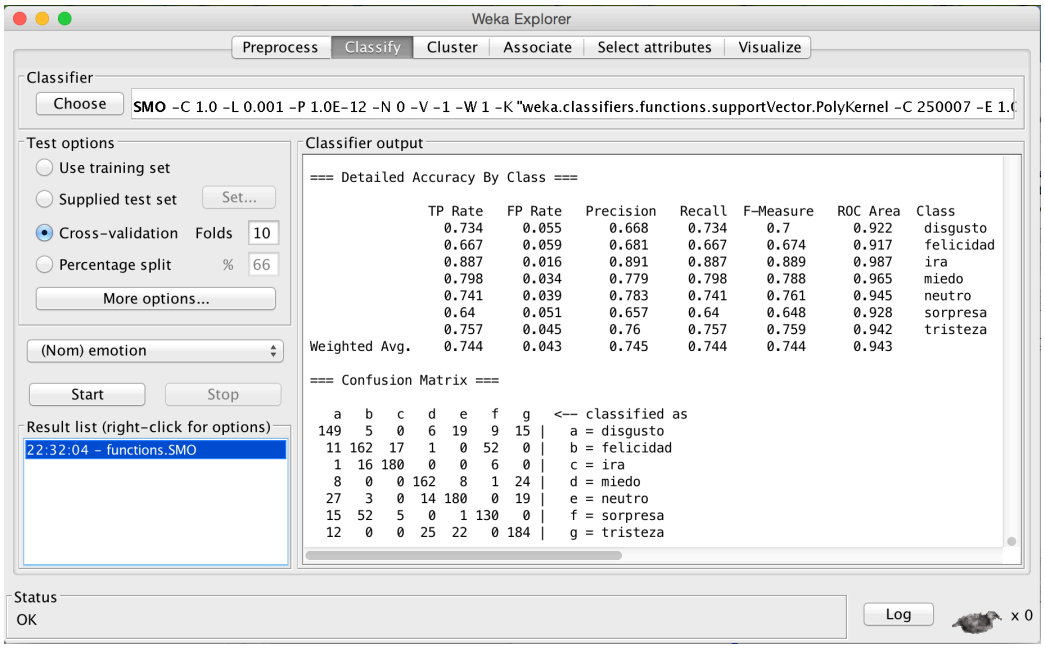


*Figura 4.11. Interfaz WEKA resultados de entrenamiento*

## 4.6.1 Modelos

Durante el proceso de entrenamiento se genera un modelo de clasificación el cual incluye reglas o patrones que se obtuvieron en base a las características en el proceso de extracción de características. Estas reglas le permiten al algoritmo clasificar nuevas instancias. WEKA permite guardar el modelo en un archivo con extensión .model. Al generar este archivo se guardan las reglas para que posteriormente se utilice sin necesidad de realizar un entrenamiento previo.

## 4.6.2 Clasificar

En este submódulo se realiza la clasificación de datos al que se desconoce su clase. Clasificar una serie de datos consiste en asignarlo a una de las clases disponibles que se encuentran en el modelo de clasificación. Para clasificar los datos es necesario definir una serie de fronteras entre las diferentes clases. Cada clasificador mediante sus reglas calcula de diferentes maneras estas fronteras. Para esta investigación se utilizaron los siguientes algoritmos de clasificación que se encuentran disponibles en WEKA:

- Multilayer Perceptron
- NaiveBayes
- RandomForest
- SMO

Cada uno con diferente tipo de implementación: el algoritmo MultilayerPerceptron utiliza las redes neuronales, NaiveBayes utiliza el teorema de Bayes, RandomForest utiliza los árboles de decisión y el SMO utiliza la implementación de la Máquinas Vectores de Soporte. La selección de estos algoritmos se debe a que la utilización de Máquinas Vector de Soporte arroja buenos resultados para la clasificación de emociones por medio de la voz. Esto se observa en la investigación de (Solís Villarreal, Yáñez Márquez, & Suárez Guerra, 2011), de igual forma para el teorema de Bayes como se ve en la investigación realizada por (Albornoz, Crolla, & Milone, 2008). Además, de tener variedad en las pruebas al utilizar diferentes modelos de clasificación.

Las fronteras se calculan mediante el proceso de entrenamiento en el que se usan las características de todos los archivos de voz. Es en este proceso cuando el clasificador infiere con sus reglas la especificación de las fronteras. Para clasificar un nuevo archivo de audio se toma el modelo de clasificación y mediante las fronteras de las características de este archivo se decide a que clase pertenece el nuevo conjunto de características del archivo de audio.

Como se aprecia en la Figura 4.12 se tiene un número de instancias o registros (características de archivos de audio) sin clasificar. Estos datos pasan por el modelo y éste, con base a los patrones del clasificador y al resultado de predicción decide a que clase (emoción) pertenece el registro.

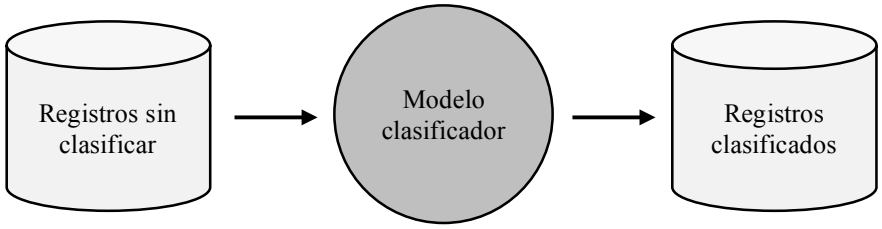


*Figura 4.12 Proceso de clasificación utilizando el modelo*

El proceso de clasificación en esta investigación es el siguiente:

- Cargar la matriz resultante del módulo Extracción.
- Cargar el archivo .model
- Utilizar la función "distributionForInstance" del archivo .model a la matriz de características.
- Se obtiene una matriz con la predicción de cada clase para ese conjunto de características. Además, la función también retorna la clase a la que pertenecen esas características.

## 4.7 Definir modelo de clasificación

Para determinar qué modelo de clasificación es el adecuado, se realizó el entrenamiento utilizando los cuatro algoritmos con cada uno de los 180 subconjuntos que se obtuvieron del módulo de extracción de características. El modelo que arroje el mejor resultado será el que se guarde el archivo .model para posteriormente usarlo en la aplicación que permitirá clasificar nuevos archivos de audio en una clase especifica.

## 4.8 Aplicación

En las actividades anteriores se determinarán qué características y modelo de clasificación son relevantes para determinar el estado emocional de la voz. En este módulo se realizó una aplicación para poder realizar predicciones. En la Figura 4.13 se observan las tareas que se aplican en esta aplicación. Se toma de base la metodología de solución pero realizando adecuaciones:

- Se cargan los archivos de audio a analizar.
- Se realiza una extracción de características de dichos archivos guardando el resultado en una matriz y sin realizar el archivo ARFF.
- Se toma el archivo .model para que el clasificador con las reglas y la matriz de las características del audio permita determinar a qué clase (emoción) pertenecen los archivos de audio.

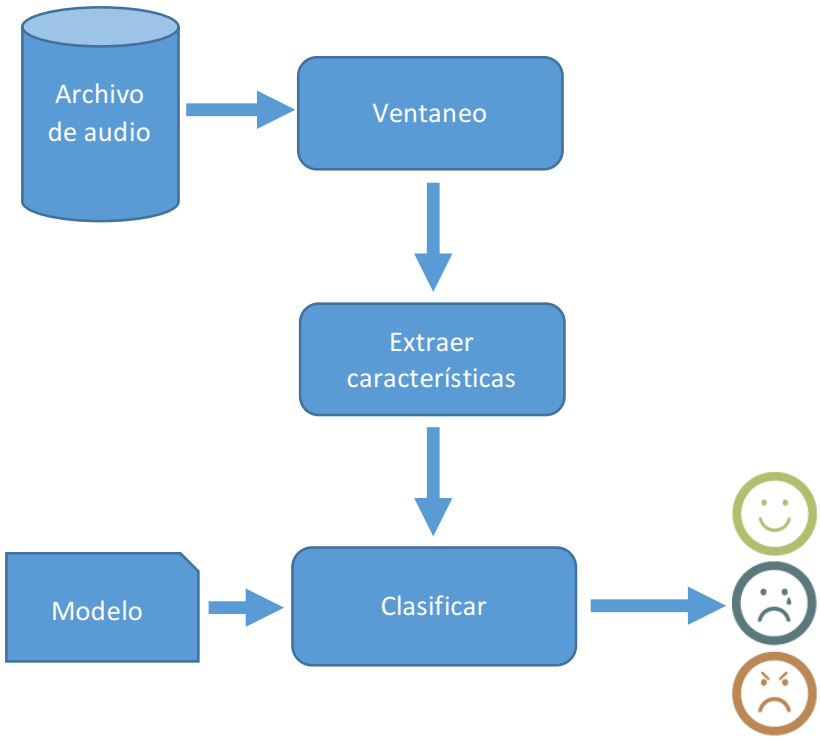


*Figura 4.13 Procesos de la aplicación para la detección de emociones*

La aplicación se desarrolló en Matlab. Se implementan todas las tareas que se enlistaron. En la Figura 4.14 se observa la interfaz de esta aplicación. Contiene un menú con una opción el cual permite cargar uno o más archivos de audio, una lista donde se muestran los nombres de los archivos de audio y una tabla donde por cada archivo de audio se mostrarán las predicciones de cada clase, esto, después de realizar la clasificación.

El funcionamiento de esta aplicación es la siguiente:

- Primero se seleccionan los archivos de audio por medio de la opción "Abrir" del menú "Archivo"
- Al cargarse los archivos, los nombres de dichos archivos se muestran en la lista
- Cuando se pulsa el botón "clasificar" la aplicación realiza el proceso de ventaneo en donde se realizan los cortes, el traslape y se aplica la función de ventaneo. Posteriormente se extraen las características para obtener la matriz de características. Luego esta matriz con ayuda del modelo de clasificación y con el método "distributionForInstance" de WEKA se realiza una predicción por cada clase de cada uno de los archivos de audio, mostrando el resultado en la tabla "Predicción de cada clase" e indicando a qué emoción pertenecen.

Tanto las características a extraer, la longitud de los cortes, la función de ventaneo y el algoritmo de clasificación utilizado para construir el modelo se determinaran cuando se tengan los resultados finales. Esto se observa en el Capítulo 5.

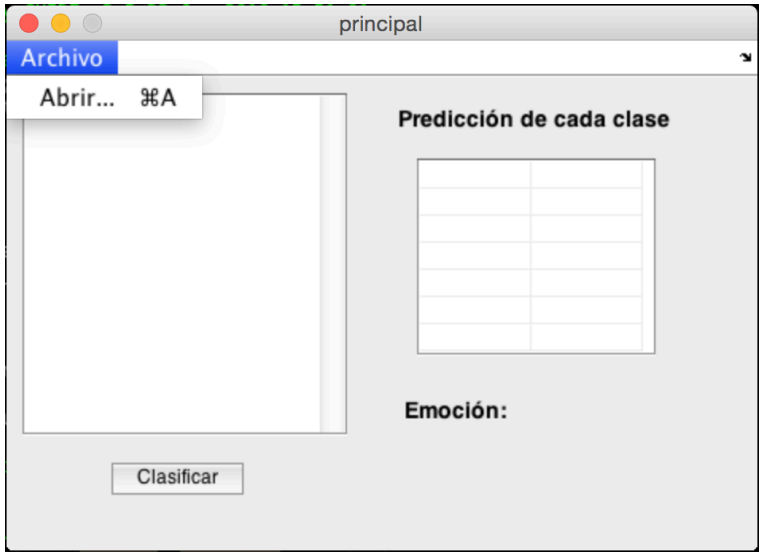


*Figura 4.14. Interfaz de la aplicación para detectar emociones*

En la Figura 4.15 se observa una simulación, de cómo se ejecutará la aplicación, del resultado de predicción de un ejemplo al cargar tres archivos de audio. El archivo "f1f.001.wav" pertenece a la clase felicidad dado que como se observa en la tabla de predicción de cada clase, la emoción felicidad tiene un dato de 0.8100 de predicción sobre las demás clases y al ser el dato mayor se considera que el archivo de audio de audio pertenece a esa emoción.

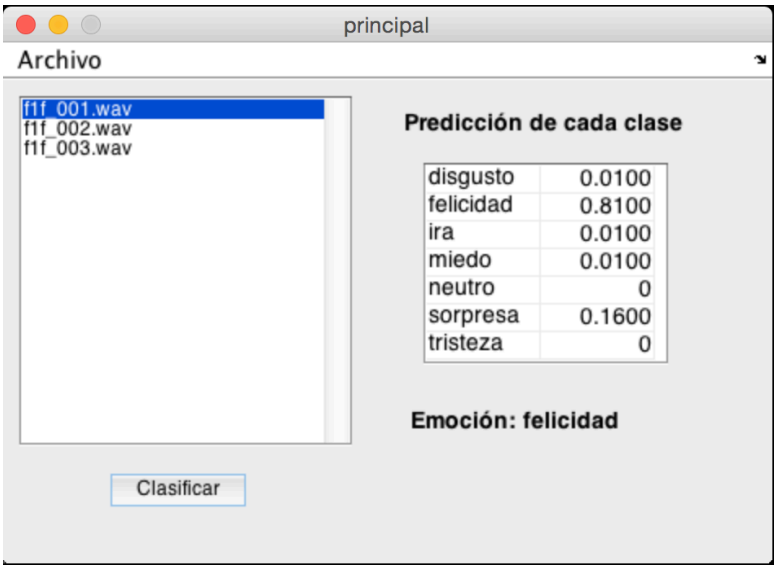


*Figura 4.15. Interfaz de la aplicación resultados de clasificación*

# Capítulo 5 Resultados

En este capítulo se presentan los resultados obtenidos de las pruebas realizadas a los subconjuntos de características con cada uno de los clasificadores. Este capítulo se encuentra dividido de la siguiente manera: se describe el método y las medidas de evaluación, se presentan los resultados de subconjuntos de características sin funciones Delta, resultados de subconjuntos aplicando funciones Deltas, los mejores resultados, pruebas adicionales y comparación de resultados.

## 5.1 Medidas de evaluación

La evaluación de las pruebas se realizó en cada uno de los 180 subconjuntos de características con la función *Cross 10 fold Validation* de WEKA. Como se explicó en el subcapítulo 4.5 se divide un subconjunto de características en 10 partes donde nueve se utilizan para entrenamiento y uno para pruebas. Para la evaluación se utilizan las siguientes medidas:

- Precisión. Es la proporción del número de instancias de la clase x entre el número de instancias que fueron clasificadas como de la clase x, es decir, mide la exactitud que tiene el clasificador para la clase x. El valor de 1 significa mayor precisión.
- Exhaustividad. Es la proporción del número de instancias que fueron clasificados como clase y de entre el número de instancias de la clase y, es decir, mide la sensibilidad que tiene el clasificador en la clase y. El valor de 1 significa mayor exhaustividad.
- Medida F. Es una medida combinada de precisión y exhaustividad. Representa la media para los datos mencionados. El valor de 1 significa un óptimo resultado.

Durante el proceso de entrenamiento en esta investigación, WEKA puede tener dos estados los cuales le ayudan a obtener las medidas mencionadas, estas son: verdadero-positivo y falso-positivo. El verdadero-positivo se da, por ejemplo, cuando se toma un valor de la parte de *Cross fold Validation* que se utiliza para pruebas que pertenece a la clase tristeza y WEKA, al momento de clasificar, da el resultado de tristeza. Para el falso-positivo se da cuando el mismo dato que pertenece a la clase tristeza WEKA lo clasifica y da como resultado alegría.

## 5.2 Evaluación con subconjuntos de características sin funciones Delta

Para esta sección solo se realizaron pruebas para los 90 subconjuntos que contienen características sin deltas. Utilizando los cuatro clasificadores: MultilayerPerceptron, NaiveBayes, RandomForest y SMO. En la Tabla 5.1 se observan los mejores resultados obtenidos aplicados al algoritmo MultilayerPerceptron en las dos bases de datos emocionales

que se utilizaron en esta investigación. Utilizando la Medida-F como referencia. Esta tabla contiene los resultados para los cinco rangos que se manejaron (20, 25, 30, 25 y 40 ms) así como también para cada una de las tres funciones de ventaneo (Blackman, Hamming y Hanning). Esto se realizó para observar que rango y que función de ventaneo presentan los mejores resultados.

| Rango | Ventana | Emo_voz.mx1 (Medida-F) | EmoWisconsin (Medida-F) |
|---|---|---|---|
| 20 ms | Blackman | 0.718 | 0.399 |
| | Hamming | 0.715 | 0.42 |
| | Hanning | 0.718 | 0.407 |
| 25 ms | Blackman | 0.692 | 0.391 |
| | Hamming | 0.74 | 0.41 |
| | Hanning | 0.732 | 0.399 |
| 30 ms | Blackman | 0.738 | 0.405 |
| | Hamming | 0.735 | 0.405 |
| | Hanning | 0.716 | 0.407 |
| 35 ms | Blackman | 0.724 | 0.403 |
| | Hamming | 0.73 | 0.398 |
| | Hanning | 0.726 | 0.417 |
| 40 ms | Blackman | 0.698 | 0.394 |
| | Hamming | 0.701 | 0.394 |
| | Hanning | 0.716 | 0.396 |

*Tabla 5.1. Mejores resultados del clasificador MultilayerPerceptron con características sin funciones Delta*

Como se observa en la tabla anterior para el corpus Emo_voz.mx1 el mejor resultado se obtiene utilizando la función de ventaneo Hamming con un rango de 25 ms y como medida-F: 0.74 y para el corpus EmoWisconsin utilizando la función Hamming con un rango de 20 ms obteniendo un resultado de 0.42. Para los dos casos los mejores resultados se dan cuando se utiliza la función Hamming con los dos rangos menores de los cinco que se utilizaron.

En la Tabla 5.2 se observan los mejores resultados aplicados al algoritmo de clasificación NaiveBayes. Utilizando los mismos datos de referencia que la tabla anterior.

| Rango | Ventana | Emo_voz.mx1 (Medida-F) | EmoWisconsin (Medida-F) |
|---|---|---|---|
| 20 ms | Blackman | 0.531 | 0.351 |
| | Hamming | 0.531 | 0.376 |
| | Hanning | 0.564 | 0.387 |
| 25 ms | Blackman | 0.546 | 0.38 |
| | Hamming | 0.575 | 0.382 |
| | Hanning | 0.544 | 0.371 |
| 30 ms | Blackman | 0.553 | 0.356 |
| | Hamming | 0.563 | 0.362 |

| | | | |
|---|---|---|---|
| | Hanning | 0.557 | 0.371 |
| 35 ms | Blackman | 0.559 | 0.363 |
| | Hamming | 0.558 | 0.349 |
| | Hanning | 0.561 | 0.36 |
| 40 ms | Blackman | 0.574 | 0.369 |
| | Hamming | 0.566 | 0.375 |
| | Hanning | 0.553 | 0.379 |

*Tabla 5.2. Mejores resultados del clasificador NaiveBayes con características sin funciones Delta*

En este caso el mejor resultado aplicado al corpus Emo_voz.mx1 se obtiene utilizando la función Hamming con un rango de 25 ms y medida-F de 0.575. En el caso del corpus EmoWisconsin se obtiene utilizando la función Hanning con un rango de 20 ms y una medida-F de 0.387. Para este caso la función Hamming en uno de los casos sigue arrojando el mejor resultado y se siguen tomando los rangos 20 y 25 ms como los mejores.

Para la Tabla 5.3 los mejores resultados que presentan se obtienen aplicando el clasificador RandomForest. Utilizando los mismos datos que las tablas anteriores.

| **Rango** | **Ventana** | **Emo_voz.mx1 (Medida-F)** | **EmoWisconsin (Medida-F)** |
|---|---|---|---|
| 20 ms | Blackman | 0.734 | 0.42 |
| | Hamming | 0.741 | 0.418 |
| | Hanning | 0.748 | 0.423 |
| 25 ms | Blackman | 0.737 | 0.406 |
| | Hamming | 0.752 | 0.416 |
| | Hanning | 0.756 | 0.411 |
| 30 ms | Blackman | 0.747 | 0.423 |
| | Hamming | 0.743 | 0.402 |
| | Hanning | 0.75 | 0.421 |
| 35 ms | Blackman | 0.735 | 0.408 |
| | Hamming | 0.729 | 0.411 |
| | Hanning | 0.734 | 0.404 |
| 40 ms | Blackman | 0.749 | 0.405 |
| | Hamming | 0.737 | 0.405 |
| | Hanning | 0.746 | 0.411 |

*Tabla 5.3. Mejores resultados del clasificador RandomForest con características sin funciones Delta*

En base a la tabla anterior el mejor resultado para el corpus Emo_voz.mx1 se obtiene utilizando la función de ventaneo Hanning con un rango de 25 ms y una medida-F de 0.756. Para el caso del corpus EmoWisconsin el mejor resultado se da utilizando las funciones Blackman con un rango de 30 ms y Hanning con un rango de 20 ms, esto se debe a que en ambos casos dan como resultado una medida-F de 0.423.

En la última tabla de esta fase de pruebas se muestran los mejores resultados aplicados al algoritmo SMO utilizando los mismos datos de descripción que las tablas anteriores.

| Rango | Ventana | Emo_voz.mx1 (Medida-F) | EmoWisconsin (Medida-F) |
|---|---|---|---|
| 20 ms | Blackman | 0.777 | 0.426 |
| | Hamming | 0.767 | 0.433 |
| | Hanning | 0.769 | 0.447 |
| 25 ms | Blackman | 0.785 | 0.429 |
| | Hamming | 0.772 | 0.445 |
| | Hanning | 0.76 | 0.442 |
| 30 ms | Blackman | 0.766 | 0.443 |
| | Hamming | 0.757 | 0.437 |
| | Hanning | 0.768 | 0.44 |
| 35 ms | Blackman | 0.768 | 0.422 |
| | Hamming | 0.783 | 0.449 |
| | Hanning | 0.773 | 0.44 |
| 40 ms | Blackman | 0.753 | 0.429 |
| | Hamming | 0.771 | 0.449 |
| | Hanning | 0.753 | 0.441 |

*Tabla 5.4. Mejores resultados del clasificador SMO con características sin funciones Delta*

Como se observa en la tabla anterior el mejor resultado aplicado en el corpus Emo_voz.mx1 se obtiene utilizando la función Blackman con un rango de 25 ms y una medida-F de 0.785. Para el caso del corpus EmoWisconsin se obtiene utilizando la función Hamming pero con rangos de 35 y 40 ms, esto debido a que utilizando la misma función de ventaneo pero en esos rangos se obtiene el mismo resultado de medida-F: 0.449

Para esta primera parte de pruebas los mejores resultados en la mayoría de los casos se obtienen utilizando la función de ventaneo Hamming con rangos de ventana de 20 y 25 ms. El clasificador SMO es el mejor para estas pruebas en base a sus resultados y el clasificador NaiveBayes es el que arroja menores resultados.

## 5.3 Evaluación con subconjuntos de características con funciones Deltas

En esta segunda fase pruebas se utilizaron los 90 subconjuntos restantes de los 180 subconjuntos que se obtuvieron del subcapítulo 4.4. Estos subconjuntos contienen características a las que se le aplicaron las funciones Delta y Delta-Delta. Como se explicó anteriormente, las pruebas en los dos tipos de conjuntos de características se realizan para observar si el aplicar las funciones Delta mejora el resultado predictivo. Las pruebas se

realizaron también para los cuatro clasificadores: MultilayerPerceptron, NaiveBayes, Random Forest y SMO.

El proceso de pruebas también se realizó como las pruebas anteriores. Ejecutando cada algoritmo clasificador con los cinco rangos de ventanas (20, 25, 30, 35 y 40 ms) y por cada función de ventaneo (Blackman, Hamming y Hanning). En la Tabla 5.5 se muestran los mejores resultados obtenidos aplicando el clasificador MultilayerPerceptron a las dos bases de datos emocionales utilizando la medida-F como referencia.

| Rango | Ventana | Emo_voz.mx1 (Medida-F) | EmoWisconsin (Medida-F) |
|---|---|---|---|
| 20 ms | Blackman | 0.725 | 0.398 |
| | Hamming | 0.721 | 0.416 |
| | Hanning | 0.759 | 0.413 |
| 25 ms | Blackman | 0.722 | 0.389 |
| | Hamming | 0.758 | 0.401 |
| | Hanning | 0.742 | 0.397 |
| 30 ms | Blackman | 0.724 | 0.404 |
| | Hamming | 0.73 | 0.379 |
| | Hanning | 0.751 | 0.396 |
| 35 ms | Blackman | 0.721 | 0.398 |
| | Hamming | 0.725 | 0.409 |
| | Hanning | 0.74 | 0.401 |
| 40 ms | Blackman | 0.729 | 0.407 |
| | Hamming | 0.752 | 0.402 |
| | Hanning | 0.729 | 0.409 |

*Tabla 5.5. Mejores resultados del clasificador MultilayerPerceptron con características aplicando funciones Delta*

Como se observa en la tabla anterior el mejor resultado para el corpus Emo_voz.mx1 fue de 0.758 utilizando la función de ventaneo Hamming con un rango de 25 ms y para el corpus EmoWisconsin fue de 0.416 utilizando la función Hamming con un rango de 20 ms. Tanto en estas pruebas como en las de características sin funciones Delta los mejores resultados se obtienen utilizando las mismas funciones de ventaneo y los rangos. Pero en el caso del corpus Emo_voz.m1 este resultado es mejor, dado que el anterior fue de 0.74. Caso contrario para el corpus EmoWisconsin donde el anterior resultado fue de 0.42 comparado con los 0.416 de esta nueva prueba.

En la Tabla 5.6 se observan los mejores resultados aplicando el clasificador NaiveBayes. Utilizando los mismos datos de referencia que las pruebas anteriores para las dos bases de datos emocionales.

| Rango | Ventana | Emo_voz.mx1 (Medida-F) | EmoWisconsin (Medida-F) |
|---|---|---|---|
| 20 ms | Blackman | 0.564 | 0.319 |
| | Hamming | 0.567 | 0.314 |
| | Hanning | 0.575 | 0.307 |
| 25 ms | Blackman | 0.572 | 0.312 |
| | Hamming | 0.582 | 0.334 |
| | Hanning | 0.587 | 0.317 |
| 30 ms | Blackman | 0.582 | 0.319 |
| | Hamming | 0.587 | 0.314 |
| | Hanning | 0.579 | 0.323 |
| 35 ms | Blackman | 0.583 | 0.319 |
| | Hamming | 0.59 | 0.35 |
| | Hanning | 0.581 | 0.324 |
| 40 ms | Blackman | 0.57 | 0.318 |
| | Hamming | 0.586 | 0.347 |
| | Hanning | 0.568 | 0.32 |

*Tabla 5.6. Mejores resultados del clasificador NaiveBayes con características aplicando funciones Delta*

En base a los resultados de la tabla anterior se obtiene que el mejor resultado para el corpus Emo_voz.mx1 se da utilizando la función de ventaneo Hamming con un rango de 30 ms y la función Hanning con un rango de 25 ms, esto debido a que en ambos casos nos muestra el resultado de 0.587. Para el caso del corpus EmoWisconsin el mejor resultado es de 0.347 utilizando la función Hamming con un rango de 40 ms. Comparando estos resultados con los del primer conjunto de pruebas para el caso del corpus Emo_voz.mx1 este resultado fue mejor que el anterior, donde se obtuvo 0.575 utilizando la misma función de ventaneo y rango para ambos casos. Pero en el caso del corpus EmoWisconisn el anterior resultado fue mejor donde se obtuvo 0.387, pero se utilizó distinta función de ventaneo y rango.

Realizando pruebas para el clasificador RandomForest los resultados se observan en la Tabla 5.7 donde se aplica para ambos corpus.

| Rango | Ventana | Emo_voz.mx1 (Medida-F) | EmoWisconsin (Medida-F) |
|---|---|---|---|
| 20 ms | Blackman | 0.745 | 0.415 |
| | Hamming | 0.74 | 0.416 |
| | Hanning | 0.736 | 0.413 |
| 25 ms | Blackman | 0.736 | 0.418 |
| | Hamming | 0.755 | 0.412 |
| | Hanning | 0.742 | 0.415 |
| 30 ms | Blackman | 0.736 | 0.415 |
| | Hamming | 0.748 | 0.411 |
| | Hanning | 0.735 | 0.417 |
| 35 ms | Blackman | 0.714 | 0.41 |

| | | | |
|---|---|---|---|
| | Hamming | 0.73 | 0.411 |
| | Hanning | 0.73 | 0.408 |
| 40 ms | Blackman | 0.743 | 0.401 |
| | Hamming | 0.736 | 0.411 |
| | Hanning | 0.729 | 0.403 |

*Tabla 5.7. Mejores resultados del clasificador RandomForest con características aplicando funciones Delta*

Para el corpus Emo_voz.mx1 el mejor resultado fue de 0.755 utilizando la función de ventaneo Hamming con un rango de 25 ms y para EmoWisconsin fue de 0.418 usando un rango de 25 ms con la función Blackman. La comparación de resultados con las de la primera sección es contraria a las de los clasificadores anteriores. Porque para el corpus Emo_voz.mx1 el mejor resultado se dio en la anterior prueba donde se obtuvo 0.756 utilizando otra función de ventaneo pero con el mismo rango y en el caso del corpus EmoWisconsin el mejor resultado también se obtiene de las pruebas anteriores que fue de 0.423 usando otra función de ventaneo pero con el mismo rango.

Para finalizar este segundo conjunto de pruebas en la Tabla 5.8 se observan los mejores resultados aplicados también para ambas bases de datos emocionales utilizando el clasificador SMO.

| **Rango** | **Ventana** | **Emo_voz.mx1 (Medida-F)** | **EmoWisconsin (Medida-F)** |
|---|---|---|---|
| 20 ms | Blackman | 0.754 | 0.438 |
| | Hamming | 0.752 | 0.446 |
| | Hanning | 0.766 | 0.439 |
| 25 ms | Blackman | 0.775 | 0.423 |
| | Hamming | 0.775 | 0.425 |
| | Hanning | 0.781 | 0.426 |
| 30 ms | Blackman | 0.771 | 0.451 |
| | Hamming | 0.758 | 0.437 |
| | Hanning | 0.771 | 0.443 |
| 35 ms | Blackman | 0.776 | 0.42 |
| | Hamming | 0.769 | 0.443 |
| | Hanning | 0.771 | 0.434 |
| 40 ms | Blackman | 0.764 | 0.429 |
| | Hamming | 0.775 | 0.447 |
| | Hanning | 0.769 | 0.444 |

*Tabla 5.8. Mejores resultados del clasificador SMO con características aplicando funciones Delta*

Como se observa en la tabla anterior el mejor resultado para el corpus Emo_voz.mx1 es de 0.781 utilizando la función de ventaneo Hanning con un rango de 25 ms. Para el corpus EmoWisconsin el mejor resultado es de 0.451 utilizando un rango de 30 ms con la función Blackman. Comparando estos resultados con las de las pruebas anteriores, para el caso del

corpus Emo_voz.mx1 el mejor resultado se obtiene de las pruebas anteriores que fue de 0.785 utilizando otra función y rango de ventaneo, pero, para el corpus EmoWisconsin el resultado de esta prueba fue mejor que el anterior donde se obtuvo un resultado de 0.449 usando otra función y rango de ventaneo.

## 5.4 Mejores resultados

Realizadas las dos fases de pruebas (con o sin funciones Deltas) se comparan los mejores resultados de cada clasificador en cada función de ventaneo y para cada uno de los rangos utilizados. Se observó que para todos los casos el clasificador SMO dio el mejor resultado. En la Tabla 5.9 se muestran los mejores resultados de clasificación divididos por los dos conjuntos generales: con y sin funciones Deltas, y el número de características de cada uno de estos subconjuntos del corpus Emo_voz.mx1.

| Ventana | Rango | Normal | | Delta | |
|---|---|---|---|---|---|
| | | Medida-F | Características | Medida-F | Características |
| Blackman | 20 ms | 0.777 | 576 | 0.754 | 1553 |
| | 25 ms | **0.785** | 539 | 0.775 | 1313 |
| | 30 ms | 0.766 | 562 | 0.771 | 1569 |
| | 35 ms | 0.768 | 447 | 0.776 | 1683 |
| | 40 ms | 0.753 | 569 | 0.764 | 1550 |
| Hamming | 20 ms | 0.767 | 557 | 0.752 | 1568 |
| | 25 ms | 0.772 | 579 | 0.775 | 746 |
| | 30 ms | 0.757 | 573 | 0.758 | 1388 |
| | 35 ms | 0.783 | 608 | 0.769 | 1889 |
| | 40 ms | 0.771 | 603 | 0.775 | 1484 |
| Hanning | 20 ms | 0.769 | 534 | 0.766 | 1485 |
| | 25 ms | 0.757 | 615 | **0.781** | 1427 |
| | 30 ms | 0.768 | 575 | 0.771 | 1469 |
| | 35 ms | 0.773 | 570 | 0.771 | 1797 |
| | 40 ms | 0.753 | 522 | 0.769 | 1348 |

*Tabla 5.9. Mejores resultados de clasificación para Emo_voz.mx1*

Como se mencionó en el subcapítulo 4.3.5 y 4.3.6 el conjunto de características sin funciones Delta obtuvo un total de 1057 características y el segundo conjunto, que contiene las funciones Deltas, se obtuvo 3169 características. Se observa una reducción en cuanto a características después de realizar el proceso de selección de atributos. Como se observa en la

tabla anterior el mejor resultado se da cuando se utilizan características sin funciones Delta. Dando una Medida-F de 0.785 utilizando la función de ventaneo Blackman con un rango de 25 ms. Ese resultado supera al resultado 0.781 de las características con funciones Delta que utiliza la función Hanning con un rango de 0.781

En la Tabla 5.10 se muestra los mejores resultados aplicados al corpus EmoWisconsin que como sucedió en el corpus anterior el clasificador SMO dio el mejor resultado en todos los casos. En la siguiente tabla los clasificadores restantes no se presentan al arrojar resultados menores al del SMO.

| **Ventana** | **Rango** | **Normal** | | **Delta** | |
|---|---|---|---|---|---|
| | | **Medida-F** | **Características** | **Medida-F** | **Características** |
| Blackman | 20 ms | 0.426 | 278 | 0.438 | 357 |
| | 25 ms | 0.429 | 228 | 0.423 | 272 |
| | 30 ms | 0.443 | 370 | **0.451** | 430 |
| | 35 ms | 0.422 | 303 | 0.42 | 293 |
| | 40 ms | 0.429 | 279 | 0.429 | 281 |
| Hamming | 20 ms | 0.433 | 380 | 0.446 | 431 |
| | 25 ms | 0.445 | 386 | 0.425 | 1512 |
| | 30 ms | 0.437 | 325 | 0.437 | 366 |
| | 35 ms | **0.449** | 343 | 0.443 | 330 |
| | 40 ms | 0.449 | 346 | 0.447 | 425 |
| Hanning | 20 ms | 0.447 | 352 | 0.439 | 417 |
| | 25 ms | 0.442 | 368 | 0.426 | 272 |
| | 30 ms | 0.44 | 396 | 0.443 | 1067 |
| | 35 ms | 0.44 | 376 | 0.434 | 1025 |
| | 40 ms | 0.441 | 338 | 0.444 | 369 |

*Tabla 5.10. Mejores resultados de clasificación para EmoWisconsin*

En base a la tabla anterior se observa que para este caso, las características con funciones Delta dieron el mejor resultado. Donde se obtiene 0.451 utilizando la función de ventaneo Blackman con un rango de 30ms en comparación con el resultado de 0.449 de las características sin funciones Delta.

En los dos corpus los mejores resultados se presentan cuando se utiliza el tipo de ventana Blackman con rangos de ventanas 25 y 30 ms. 0.785 para Emo_voz.mx1 y 0.451 para EmoWisconsin tomando la medida F como referencia.

Localizado el mejor subconjunto de características de cada corpus que en conjunto con el algoritmo SMO arrojaron los mejores resultados. Se realizó una prueba para determinar qué características de cada subconjunto son los que aportan información más relevante. Se aplicó el método InfoGainAttributeEval de WEKA en conjunto con un método de ranqueo para ordenar las características de acuerdo a su ganancia de información. En la Tabla 5.11 se muestran las primeras 20 mejores características de cada subconjunto de las bases de datos emocionales.

| Número | Emo_voz.mx1 | EmoWisconsin |
|---|---|---|
| 1 | Energy_peakMean | Delta_Delta_MFCC01_std |
| 2 | Volume_quartil3 | Delta_Volume_std |
| 3 | Volume_percentil95 | Delta_Delta_Energy_iqr13 |
| 4 | Energy_iqr13 | Delta_Energy_std |
| 5 | Energy_iqr23 | Delta_Delta_MFCC02_quartil1 |
| 6 | Energy_std | Volume_percentil95 |
| 7 | Volume_percentil98 | Delta_Delta_Energy_percentil95 |
| 8 | Energy_percentil98 | Delta_ChromaVector10_variance |
| 9 | Volume_std | Delta_Energy_percentil98 |
| 10 | MFCC01_quartil3 | ChromaVector10_percentil98 |
| 11 | Volume_mean | Delta_Energy_quartil1 |
| 12 | Volume_maximo | MFCC01_mean |
| 13 | Energy_range | Delta_MFCC01_quartil3 |
| 14 | Energy_maximo | Volume_mean |
| 15 | MFCC01_mean_geo | Delta_ChromaVector10_percentil95 |
| 16 | MFCC01_maximo | Delta_MFCC01_range |
| 17 | Volume_range | Delta_Delta_SpectralCentroid_percentil95 |
| 18 | Volume_peakMean | Energy_range |
| 19 | Energy_iqr12 | Delta_MFCC03_std |
| 20 | Volume_iqr13 | Delta_Delta_MFCC02_percentil95 |

*Tabla 5.11. Primeras 20 mejores características de las dos bases de datos*

En base a la tabla anterior se observa que la energía, volumen y los MFCCs son características que más información relevante aportan y que son consideras claves para la detección de emociones por medio de la voz. También se observa que hay similitudes en las mejores características de una base de datos actuada y una inducida.

En las siguientes figuras se muestran graficas donde se muestran las respuestas que tienen las funciones Blackman, Hamming y Hanning en dos características de voz. En la Figura 5.1 se aplican las funciones a la energía con un rango de 25 ms y en la Figura 5.2 a la característica volumen. Como se observó en la Tabla 5.11 estas características aportan más información relevante para distinguir una emoción de otra. En ambas figuras se observa que la respuesta que tienen los datos obtenidos de la función Blackman son más suaves que en el otro par de funciones donde sus resultados son muy similares.

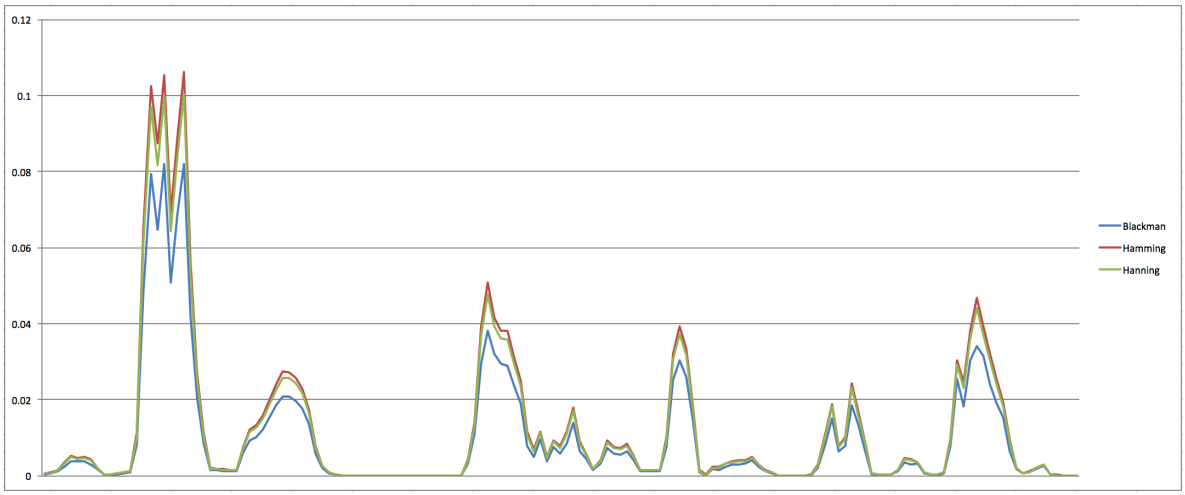


*Figura 5.1. Respuestas de las funciones de ventaneo a la característica Energía*

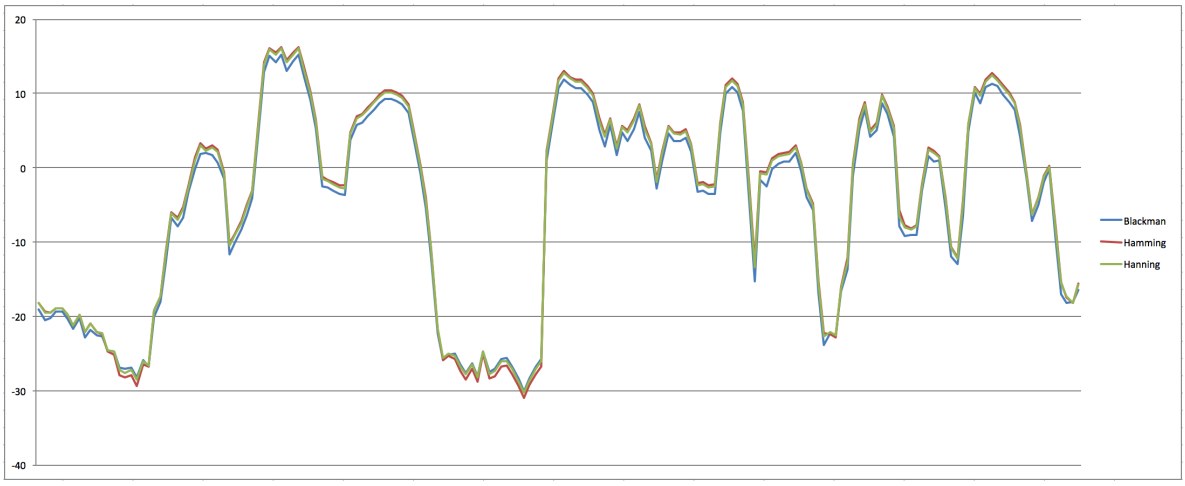


*Figura 5.2. Respuestas de las funciones de ventaneo a la característica Volumen*

## 5.5 Pruebas adicionales

Con las pruebas realizadas de la sección anterior se determinó que el ventaneo Blackman con rangos de 25 y 30 ms son los que mejores resultados. Siendo estos:

- Emo_voz.mx1: 0.785 utilizando el clasificador SMO con 539 características.
- EmoWisconsin: 0.451 utilizando el clasificador SMO con 430 características.

Para complementar esta investigación se realizaron otros dos tipos de pruebas. Se observaron si ambos corpus comparten características y modelo de entrenamiento, dado que cada corpus es de naturaleza diferente: una es actuada y la otra inducida. Además, de conocer si hubo una mejora en la predicción. La primera consistió en realizar la unión e intersección del mejor conjunto de características de ambos corpus; para el caso de la unión se obtuvo un total de 882 características y para la intersección 87 características. Los algoritmos de clasificación del software WEKA son los mismos que se utilizaron en la sección de evaluación de clasificadores.

Estos algoritmos se aplican a los nuevos conjuntos de características. Se observa que no se mejora la eficiencia de predicción dado que los resultados se ubican por debajo de los resultados con los que ya se contaba. En la Tabla 5.12 se observa el clasificador, el mejor resultado para cada corpus y el mejor resultado que ya se tenía tomando la Medida-F como referencia. En la mayoría de los casos el clasificador SMO es el que sigue arrojando los mejores resultados.

| Base de datos | Unión | Intersección | Mejor resultado |
|---|---|---|---|
| Emo_voz.mx1 | 0.778 / SMO | 0.722 / RandomForest | 0.785 / SMO |
| EmoWisconsin | 0.438 / SMO | 0.388 / Multilayer Perceptron | 0.451 / SMO |

*Tabla 5.12. Comparación de mejores resultados de los conjuntos unión e intersección*

La segunda prueba consistió en realizar experimentos cruzados con los mejores conjuntos de características de ambos corpus y aplicando los cuatro algoritmos de clasificación. El mejor conjunto del corpus Emo_voz.mx1, el cual tiene 539 características, se toma para entrenar y probar el corpus EmoWisconsin. También el mejor conjunto del corpus EmoWisconsin, el cual tiene 430 características, se entrenó y se probó al corpus Emo_voz.mx1. Cómo se observa en la Tabla 5.13 no se logró una mejora en el modelo de predicción.

| Base de datos | Cruzados | Mejor resultado |
|---|---|---|
| Emo_voz.mx1 | 0.718 / SMO | 0.785 / SMO |
| EmoWisconsin | 0.426 / SMO | 0.451 / SMO |

*Tabla 5.13. Comparación de resultados de experimentos cruzados*

## 5.6 Comparación de resultados

En las investigaciones de las bases de datos que se utilizaron para el proceso de extracción de características y entrenamiento; tienen una sección donde utilizan su propia metodología para realizar pruebas de clasificación. En esta sección se realiza una comparación de sus resultados con los obtenidos en esta investigación.

El objetivo de la investigación del corpus Emo_voz.mx1 fue crear un corpus emocional en idioma Español de México y posteriormente su evaluación. Para esto extrajeron 14 características de audio: 12 MFCC, frecuencia fundamental F0 y los coeficientes de la energía. Para el proceso de clasificación utilizaron el algoritmo j48. Guardando los resultados en una matriz de confusión. A dicha matriz se le aplicó el método *Ward* con el cual se obtiene un dendrograma. Este dato proporciona los niveles jerárquicos. Posteriormente, se utiliza la distancia euclidiana para determinar la similitud entre las clases y se utilizan las Máquinas Vectores de Soporte para realizar la clasificación.

El resultado que proporciona esta investigación la define como el porcentaje de instancias clasificadas incorrectamente. En la Tabla 5.14 se observa la comparación del resultado que ellos mencionan con el que se obtuvo en esta investigación. Se mencionó que el mejor resultado para este corpus en este trabajo de tesis fue de 0.758 para la medida-F. Además de las medidas de evaluación que se manejan para los resultados de este capítulo, WEKA proporciona otras, entre ellas está el porcentaje de instancias clasificadas incorrectamente.

| Nombre | % incorrectos |
|---|---|
| Emo_voz.mx1 | 33.59 |
| Detección del estado emocional mediante la voz en Español de México | 21.47 |

*Tabla 5.14. Comparación de resultados del corpus Emo_voz.mx*

Como se observa en la tabla anterior el porcentaje de esta investigación es menor que el del corpus, por lo tanto, el porcentaje de instancias clasificadas correctamente en esta investigación es mayor que el resultado que presenta el corpus.

La Tabla 5.15 muestra la comparación del resultado del corpus EmoWisconsin con el de esta investigación tomando la medida F como atributo de comparación. La aportación principal de la investigación de este corpus fue la creación de una base de datos emocional inducida en Español de México. Dado que actualmente existen pocos corpus de esta variedad del lenguaje Español. El proceso de clasificación tiene mucha similitud con el propuesto en esta investigación. Primero extrajeron 6,552 características usando la herramienta OpenEAR: energía, probabilidad de sonoridad, MFCC, energía espectral en bandas, flujo espectral, máximo y mínimo espectral, tasa de cruce por ceros, contorno F0, espectro MEL, punto rollof espectral, centroide espectral. Para el proceso de selección de características se utilizó un proceso basado en *Cfs Subset* usando un algoritmo genético como método de búsqueda y utilizando el algoritmo Máquina Vector de Soporte para la clasificación. Evaluado con *10-fold cross validation* tomando la medida-F como referencia.

| Nombre | Medida F |
|---|---|
| EmoWisconsin | 0.407 |
| Detección del estado emocional mediante la voz en Español de México | 0.451 |

*Tabla 5.15. Comparación de resultados del corpus EmoWisconsin*

En base a la tabla anterior se observa que el resultado de esta investigación fue superior al que presenta el corpus. Utilizando el mismo método de evaluación y función de clasificación; el cual están basados en Máquina Vector de Soporte.

## 5.7 Experimento

Se realizó una aplicación en Matlab, el cual, mediante un archivo de voz permite conocer el estado emocional. Para esto, se extraen las 539 características del subconjunto y el modelo del clasificador SMO que presentó el mejor resultado en el proceso de pruebas como se observa en el subcapítulo 5.4 el cual fue de medida F: 0.785 Para observar cómo se comporta la aplicación con nuevos datos se realizó un nuevo experimento. Se hicieron grabaciones de audio a compañeros del CENIDET. En esta tarea se tuvo el apoyo del departamento de Desarrollo Académico del CENIDET quienes nos ayudaron a inducir emociones en los voluntarios. Se contó con la presencia de cuatro personas: dos mujeres y dos hombres.

Se realizaron dos tipos de pruebas: un juego de mesa y videos. En el juego de mesa se utilizó el Jenga Pink. Este juego trata de ir sacando bloques de una torre por turnos y colocarlos en su parte superior. Se juega con 54 bloques de madera que se ubican en formación cruzada por niveles de tres bloques juntos hasta conformar una torre de 18 niveles de altura. Cada bloque contiene una pregunta que el participante debe leer y responder. Algunas preguntas de ejemplo son:

- ¿Quién es tu actor favorito?
- ¿Cuál es tu restaurante favorito?
- ¿Cuál es tu color favorito?
- ¿Quién es tu amor secreto?
- ¿Cuál ha sido el mayor ridículo que has hecho?
- ¿Con quién ha sido tu primer beso?

En el caso de los videos se tomaron en cuenta los siguientes. La explicación de cada uno de estos videos se encuentra en el Anexo 2:

- ALMA. Cortometraje de animación en 3D dirigido por Rodrigo Blass
- Aprobación. Cortometraje dirigido por Kurt Kuenne
- El sueño del caracol. Cortometraje dirigido por Ivan Sáinz Pardo
- ¿Qué es esto? Cortometraje dirigido por Constantin Pilavios

Las grabaciones se realizaron en un ambiente moderadamente ruidoso, utilizando una laptop con un micrófono Perfect Choice omnidireccional y la herramienta Speech File System Versión 4.10 (Speech Filing System) con una frecuencia de muestreo de 16 KHz. Se grabaron tres sesiones de juego donde participaron los cuatro voluntarios. Posteriormente se realizaron dos sesiones para ver los cortometrajes. Las dos sesiones se realizaron en parejas, primero las dos mujeres y después los dos hombres. A cada uno se les mostro un cortometraje para que al terminar dieran su opinión de dicho video.

Después de la grabación se procedió a realizar cortes de manera manual en los archivos de audio utilizando el software Audacity Versión 2.1 (Audacity Team, 2016). Esto con el fin de eliminar pausas. Se obtuvo un total de 87 pistas de los archivos de audio del Jenga Pink y 50 pistas de los archivos de los videos. Las pistas de audio fueron analizadas por la aplicación que se desarrolló en este módulo. En estas pruebas el comportamiento de los participantes al responder las preguntas mostraban indicios de alegría, duda y en algunos casos se mostraban indiferentes. En la Tabla 5.16 se muestra los resultados de clasificación de los archivos de audio del juego Jenga Pink. Se aprecia la emoción encontrada, el número de pistas que pertenecen a cada emoción y el rango de predicción de las pistas en cada emoción.

| **Jenga** | | **Disgusto** | **Miedo** | **Neutro** | **Sorpresa** | **Tristeza** |
|---|---|---|---|---|---|---|
| Juego 1 | Núm. de pistas | 14 | 2 | 22 | 1 | - |
| | Rango de predicción | 0.226 – 0.334 | 0.253 – 0.256 | 0.239 – 0.449 | 0.267 | - |
| Juego 2 | Núm. de pistas | 3 | - | 9 | - | - |
| | Rango de | 0.29 – 0.33 | - | 0.250 – | - | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| | predicción | | | 0.477 | | |
| Juego 3 | Núm. de pistas | 15 | 1 | 18 | 1 | 1 |
| | Rango de predicción | 0.249 – 0.318 | 0.240 | 0.248 – 0.429 | 0.225 | 0.265 |

*Tabla 5.16. Resultados de clasificación para los archivos de Jenka Pink*

En la Tabla 5.17 se observa los resultados de los archivos de audio de los cortometrajes donde se muestra la emoción encontrada, el número de pistas que pertenecen a una emoción y el rango de predicción de las pistas en cada emoción. Por el contenido que muestra cada cortometraje cada una inducía emociones diferentes. Para ALMA: miedo y tristeza. Aprobación: alegría y tristeza. El sueño del caracol: alegría y tristeza. ¿Qué es esto?: enojo y tristeza.

| **Videos** | | **Emociones** | | | | |
|---|---|---|---|---|---|---|
| | | **Disgusto** | **Miedo** | **Neutro** | **Sorpresa** | **Tristeza** |
| ALMA | Núm. de pistas | 6 | 3 | 6 | - | 1 |
| | Rango de predicción | 0.224 – 0.339 | 0.265 – 0.31 | 0.217 – 0.452 | - | 0.31 |
| Aprobación | Núm. de pistas | - | 4 | 8 | - | 2 |
| | Rango de predicción | - | 0.263 – 0.34 | 0.263 – 0.390 | - | 0.287 – 0.335 |
| El sueño del caracol | Núm. de pistas | 5 | 2 | 6 | 2 | 1 |
| | Rango de predicción | 0.253 – 0.333 | 0.298 – 0.312 | 0.243 – 0.43 | 0.29 – 0.31 | 0.347 |
| ¿Qué es esto? | Núm. de pistas | 3 | 1 | - | - | - |
| | Rango de predicción | 0.248 – 0.290 | 0.324 | - | - | - |

*Tabla 5.17. Resultados de clasificación para los archivos de los Cortometrajes*

Como se aprecian en las tablas anteriores solo se tomaron en cuenta las emociones que muestra la aplicación para cada una de las pistas de audio, las cuales son: disgusto, miedo, neutro, sorpresa y tristeza. Los rangos de predicción para cada emoción son muy bajo quedando por debajo del 50%. Incluso la aplicación muestra resultados en emociones donde en las pruebas del juego y cada uno de los cortometrajes no se habían detectado. Además, estos datos se grabaron en un ambiente con poco ruido de fondo por lo cual esto pudo haber influido en los resultados. Como un primer acercamiento la aplicación necesita mejorar la

predicción con archivos de naturaleza inducida como lo fue con estos nuevos datos del experimento realizado.

# Capítulo 6 Conclusiones y trabajos futuros

En este capítulo se presentan las conclusiones generales, contribuciones y trabajos futuros de la metodología propuesta en esta investigación.

## 6.1 Conclusiones

La metodología que se presentó en el capítulo 4 de este documento de investigación cumple con el objetivo de realizar una aplicación que permita detectar el estado emocional de una persona por medio de su voz. Dicha metodología se compone de dos fases: Extracción y Clasificación. Cada una de estas fases se encuentra divida en módulos y submódulos los cuales son secuenciales y se complementan para definir el conjunto de técnicas, métodos y procesos explicados en la metodología.

Durante la realización de esta tesis se buscó y se localizaron pocas bases de datos emocionales en México. Al encontrar únicamente cinco donde solo se obtuvo acceso a dos: una inducida y otra actuada. Estos corpus incluyen acentos regionales del Sur-Centro y Este del país. Se construyeron conjuntos de características acústicas y se seleccionó el mejor conjunto apropiado para la tarea del reconocimiento de emociones en la voz con ayuda de métodos de evaluación y búsqueda.

Los rasgos que mostraron el mejor rendimiento eran claramente la energía, volumen y MFCCs. Esto se observa en la Tabla 5.11. Como se observó en algunas investigaciones del capítulo 3 tanto la característica energía y los MFCCS dieron los mejores resultados de clasificación incluso cuando se utilizaban solo una característica para pruebas. Estas características tienen una mejor ganancia de información con respecto a las demás. La función de ventaneo que dio los mejores resultados fue Blackman.

El mejor método de clasificación fue un algoritmo basado en Máquinas Vectores de Soporte (SMO como implementación en WEKA). Utilizando este algoritmo de aprendizaje se obtuvieron los mejores resultados para ambos corpus. Esto se aprecia en las Tablas 5.9 y 5.10. Este método resultó ser robusto para estas tareas de clasificación. Los algoritmos basados en Máquinas Vectores de Soporte en la mayoría de los casos siempre han mostrado ser los mejores para las tareas de clasificación para problemas binarios o multiclase. Esto se observa en la mayoría de los trabajos presentados en el capítulo 3.

Se construyeron nuevos conjuntos de características como la intersección, unión y cruzados de los mejores subconjuntos seleccionados para cada base de datos por separado. Estas nuevas características no mejoraron el rendimiento de la clasificación de los modelos entrenados, pero la tasa de exactitud no disminuyó lo suficiente.

Optimizando el conjunto de características propuesto en esta investigación y con el mejor método de clasificación desarrollado con estos experimentos; es posible generar mejores resultados que otros autores que utilizan las mismas bases de datos.

En las pruebas realizadas a los corpus del idioma Español de México los resultados para al del tipo actuado fueron buenos pero para al del tipo inducido no. En el experimento que se realizó con cuatro alumnos del CENIDET donde se obtuvo datos del tipo inducido, los resultados de clasificación que se obtuvieron estuvieron por debajo del 50% de precisión. Se necesita mejorar la predicción de clasificación realizando más pruebas en datos inducidos o naturales para mejorar el resultado predictivo implementando más características de voz o agregar nuevos módulos al proceso de extracción de características.

### 6.1.1 Contribuciones

En esta sección se listan las contribuciones que se obtuvieron con el desarrollo de esta investigación:

- Una aplicación que permite conocer el estado emocional de las personas por medio de su voz en idioma Español de México.
- Otro acercamiento de realizar la detección de emociones en idioma Español de México. Como se menciona en el Capítulo 3 existen pocas investigaciones enfocadas en esta variante del idioma Español.
- Análisis de las tres funciones de ventaneo: Blackman, Hamming y Hanning.
- Análisis de subconjuntos de características donde se observó que la energía, el volumen y los MFCCs son características que aportan mayor información relevante al momento de clasificar emociones.

## 6.2 Trabajos futuros

A continuación se listan los trabajos futuros relacionados a la metodología de solución presentada en esta investigación:

- Integrar más características de voz: En esta investigación se realizaron pruebas con algunas características de voz como un acercamiento en este campo; pero existen otras las cuales pueden ser de utilidad.
- Elaborar una base de datos emocional inducida: Como se mencionó en el Capítulo 3 existen pocas bases de datos emocionales en idioma Español de México y las que existen la mayoría son de tipo actuado. Para el entrenamiento de un sistema este tipo de datos nos darán resultados alejados de la realidad por lo que es conveniente realizar

pruebas en bases de datos inducidas. Dado que este tipo es lo más cercano a un dato natural.

- Agregar una sección de eliminación de ruido: Antes de realizar el proceso de extracción de características es conveniente agregar un módulo de eliminación de ruido en el audio para obtener solo la voz y así tener un mejor resultado de predicción.
- Realizar una aplicación de detección en tiempo real: La aplicación que se realizó en esta investigación toma archivos de audio grabados para que posteriormente puedan ser procesados. Es conveniente tener una aplicación que cuando una persona esté hablando al mismo tiempo este detectando la emoción e incluso los cambios que pueda tener de una emoción a otra.

## **Anexo 1. Palabras, frases y párrafos de la base de datos SES** (Montero Martínez, 2003)

**Palabras**

1) *la yema*
2) *jardín*
3) *huevo*
4) *se cayó*
5) *la llave*
6) *el bolsillo*
7) *la puerta*
8) *cerrojo*
9) *veinte*
10) *el final*
11) *chico*
12) *esquina*
13) *ropa*
14) *se cambió*
15) *fruta*
16) *no queda*
17) *niño*
18) *coche*
19) *gregoriano*
20) *le gusta*
21) *la deuda*
22) *cero*
23) *experiencia*
24) *vivirás*
25) *gozan*
26) *reina*
27) *salud*
28) *llegó*
29) *cerrado*
30) *Arrizabalaga*
31) *reyerta*

**Frases**

1) *No queda fruta los viernes*
2) *¿Ya se cambió de ropa?*
3) *¿Hay algún chico en la esquina?*
4) *El final del siglo veinte.*
5) *¿La puerta tiene cerrojo?*
6) *Tengo la llave en el bolsillo.*
7) *¿Se cayó en el jardín?*
8) *¿Rompió la yema del huevo?*
9) *Gozan de perfecta salud.*
10) *Vivirás una feliz experiencia.*
11) *Dejaron la deuda al cero*
12) *Le gusta mucho el gregoriano.*
13) *Yo llevo al niño en el coche.*
14) *Llegó la reina del puño cerrado.*
15) *Arrizabalaga dejará la reyerta.*

**Párrafos**

1) *Los participantes en el Congreso marcharon después a El Escorial. Se trasladaron allí en un amplio autobús, en el que un guía iba explicando los monumentos relevantes del recorrido. La visita al monasterio fue comentada por el mismo guía que debía saber mucho sobre El Greco, en cuyo cuadro "el martirio de San Mauricio" se extendió ampliamente; no debía ser igual su conocimiento del resto de los cuadros que componían la pinacoteca, sobre los cuales pasó como un rayo, dando lugar a sonrisas cómplices.*
2) *Sergio era un joven serio y trabajador que vivía cerca de la hospedería del Monasterio de Guadalupe, en las Villuercas, comarca perteneciente a la provincia de Cáceres. Se*

*ganaba la vida vendiendo recuerdos alusivos a la Virgen Morenita, desde llaveros a platos con la imagen grabada en esmalte vidriado. Tenía un problema y era que su tiendecita era de mala construcción y estaba en una parte del pueblo muy empinada, fenómeno por otra parte normal en aquel lugar. Había mucho turismo en la zona. Sergio tuvo la mala suerte de perder su tienda en las últimas inundaciones, pues un corrimiento de tierras se la llevó por delante, con lo cual se le acabó su modo de vida.*

3) *Pablo estudiaba en la Universidad Politécnica de Madrid y estaba deseando regresar a Medellín; echaba de menos los productos de la matanza y los quesos frescos que hacía su abuela. Ya faltaba poco para las vacaciones; entonces volvería a las orillas del Guadiana, bajo los chopos. Su deseo era tan grande que a veces se le hacían años los pocos días que faltaban.*

4) *La vida diaria a menudo no es tan fácil, aunque estemos en el final del siglo veinte. Sobre todo cuando los dos en la pareja trabajan. Siempre hay que preguntarse si ya se cambió la ropa, si la puerta tiene el cerrojo o si tengo la llave en el bolsillo. Yo llevo al niño en el coche. Todos los días; al colegio. Pero ¿quién hace la compra? Al final de la semana todo se acaba. No queda fruta los viernes. Los sábados dejaron la cuenta al cero. Y los domingos, aunque te dices que vivirás una feliz experiencia, la cosa no es tan sencilla: el niño sale con sus amigos. ¿Hay algún chico en la esquina? ¿Se cayó en el jardín? Desde luego, siempre gozan de perfecta salud y yo estoy aquí preocupándome por nada. Definitivamente, vivir no es tan sencillo ni al final del siglo veinte.*

## Anexo 2. Descripción de los cortometrajes

| Nombre | Descripción | Duración |
|---|---|---|
| ALMA | En un pueblo está nevando. Un niño se acerca a una pared a escribir ALMA. Al voltear ve en una ventana de una tienda un muñeco idéntico a él. Decide entrar a la tienda y observa varios muñecos con forma de niños. Busca el muñeco que vio y lo encuentra arriba de un mueble. Cuando lo alcanza y lo toca; su alma pasa al muñeco. Posteriormente se da cuenta que él es el muñeco. Al final otra muñeca se coloca en la ventana en espera de otra víctima. | 5:30 |
| Aprobación | Un joven se dedica a motivar a los demás. Involucrando desde personas normales, hasta actrices y oficiales de policías; cuando decide sacar una foto para su licencia de conducir conoce a una chica que es la camarógrafa. Ella le dice que no tiene que sonreír. Cuando se da cuenta que no le hace caso decide seguir tomando fotos a las demás personas. Él decide hacerla sonreír y la sigue a todos lados, pero al darse cuenta que no puede lograrlo él se convierte en una persona seria y triste. Esto ocasiona que las motivaciones las de sin ganas y provoca que lo despidan. Después encuentra otro motivo para ser feliz y esto es tomando fotos a otras personas tratando de motivarlas. Observa una licencia la cual tiene la foto sonriendo, se asombra y va en busca de la chica pero al llegar le comentan que la despidieron al sacar las fotos de las personas sonriendo. Decide buscarla y la encuentra tomando fotos a personas sonriendo ella le comenta el porqué de su seriedad y el la comprende. Al final los dos quedan juntos. | 17:10 |
| El sueño del caracol | En una cafetería una chica se encuentre leyendo un libro cuando conoce a un chico. Este al ver la mirada de la chica decide irse y ella lo sigue hasta una librería. En ese lugar ella toma un libro para disimular pero él se da cuenta entonces ella decide comprar el libro el cual lleva por título caracoles. Cuando regresa a su cuarto empieza a recordar al chico y cada día lo va a visitar a la librería. Ella está enamorada de él y empieza a practicar frente espejo para invitarlo a salir, | 15:05 |

| | | |
|---|---|---|
| | Cuando llega a la librería no lo ve entonces pregunta por él y le dan la triste noticia que el chico había fallecido. Ella llega triste a su cuarto y se da cuenta que en un libro se encuentra una carta del chico donde menciona que él está enamorado de ella. | |
| ¿Qué es esto? | En un patio se encuentra un padre y su hijo. El padre al ver un pájaro en el jardín le pregunta a su hijo ¿qué es eso? él le contesta que es un gorrión. El padre continúa preguntado ¿qué es eso? y el hijo continúa respondiendo. Conforme pasa el tiempo el hijo empieza a perder la paciencia hasta que le grita. El padre se levanta y entra a su casa a traer una nota se la entrega a su hijo y le dice que lo lea en voz alta. Él lo hace y la nota dice que su papa se había sentado en el parque con su hijo que había cumplido tres años. El niño le pregunto 21 veces ¿qué es eso? y el padre siempre respondía que era un gorrión y por cada pregunta lo abrazaba y lo besaba. Entonces el hijo se da cuenta de su error y abraza a su padre. | 5:31 |

# Referencias

Ávila, E., & Quintana, P. (1994). Codificación, sintesís y reconocimiento de voz. *Universidad de Gran Canaria* .

Albornoz, E. M., Crolla, M. B., & Milone, D. H. (2008). Recognition of emotions in speech. *XXXVI Conferencia Latinoamericana de Informática* , 1120-1129.

*Audacity Team*. (20 de enero de 2016). Obtenido de http://www.audacityteam.org/

Boersma, P., & Weenink, D. (s.f.). Obtenido de Praat: doing phonetics by computer: http://www.fon.hum.uva.nl/praat/

Bojanić, M., & Delić, V. (2013). Automatic emotional speech recognition in Serbian language. *21st Telecommunications forum TELFOR 2013* , 459-465.

Breiman, L. (2001). Random Forests. *Machine Learning* , 5-32.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A Database of German Emotional Speech. *Proceedings Interspeech* , 1517-1520.

Caballero Morales, S. (2013). Recognition of emotions in Mexican Spanish speech: an approach based on acoustic modelling of emotion-specific vowel. *Scientific World Journal* , -.

Chambi, J. (2013 de septiembre de 2013). *Blog de Big Analítica en Perú*. Recuperado el 10 de octubre de 2014, de Perú Analítica: http://www.peruanalitica.com/2013/09/arboles-de-decision/

Chiriacescu, I. (2009). *Automatic Emotion Analysis Based on Speech (Tesis de maestría)*. Delft, Nederland: Delft University of Technology.

Echeverry Correa, J. D., & Morales Pérez, M. (2008). Reconocimiento de emociones en el habla. *Tecnológicas* , 113-130.

Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to Audio Analysis: A MATLAB Approach.* Oxford, UK: Elsevier.

Grant, D., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, (págs. 404-411).

Han, Z., Lun, S., & Wang, J. (2012). A Study on Speech Emotion Recognition Based on CCBC and Neural Network. *International Conference on Computer Science and Electronics Engineering* (págs. 144-147). Hangzhou: IEEE.

ift. (12 de mayo de 2016). *Informe Estadístico 4to Trimestre 2015.* Recuperado el 28 de mayo de 2016, de Instituto Federal de Telecomunicaciones: http://www.ift.org.mx/estadisticas/informe-estadistico-4to-trimestre-2015

INEGI. (8 de diciembre de 2015). *Encuesta Intercensal 2015.* Recuperado el 11 de junio de 2016, de Instituto Nacional de Estadística y Geografía: http://www.inegi.org.mx/est/contenidos/Proyectos/encuestas/hogares/especiales/ei2015/

Jovičić, S., Kašić, Z., Djordjević, M., & Rajković, M. (2004). Serbian emotional speech database: design, processing and evaluation. *Proc. SPECOM'2004* , 77-81.

Kuang, Y., & Li, L. (2013). Speech Emotion Recognition of Decision Fusion Based on DS Evidence Theory. *Software Engineering and Service Science (ICSESS)* , 795-798.

Long Pao, T., Hsiang Wang, C., & Ji Li, Y. (2012). A Study on the Search of the Most Discriminative Speech Features in the Speaker Dependent Speech Emotion Recognition. *Fifth International Symposium on Parallel Architectures, Algorithms and Programming* (págs. 157-162). Taipei: IEEE.

Maldonado, L. (2012). Los modelos ocultos de Markov, MOM. *TELOS* , 433-438.

Macías Kempe, R. (2008). *Corpus de voz y video para apoyar la detección de mentiras, enojo y nerviosismo (Tesis de licenciatura).* Cholula, Puebla, México: Universidad de las Américas Puebla.

Mamani Laqui, A. (17 de Noviembre de 2014). *Scribd.* Obtenido de Clasificador Naive Bayes para el reconocimiento de patrones: https://es.scribd.com/doc/178906507/Bayes

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The eNTERFACE'05 Audio-Visual Emotion Database. *Data Engineering Workshops, 22nd International Conference on* , -.

MathWorks. (2014). *MATLAB The Language of Technical Computing*. Obtenido de The MathWorks, Inc.: http://www.mathworks.com/products/matlab/

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 (Tesis doctoral). Madrid, España: Universidad Politécnica de Madrid, Departamento de Ingeniería Electrónica.

Montero, J., Gutiérrez, J., Palazuelos, S., Enríquez, E., Aguilera, S., & Pardo, J. (1998). Emotional Speech Synthesis: from speech database to TTS. *International Conference on Spoken Language Processing'98* , 923-925.

Moujahid, A., Inza, I., & Larrañaga, P. (2008). Tema 5. Clasificadores K-NN. *Departamento de Ciencias de la Computación e Inteligencia Artificial Universidad del País Vasco-Euskal Herriko Unibertsitatea* , -.

Navarrete García, J. (2003). *Mejora en el algoritmo de segmentación para el reconocimiento de caracteres de telegramas escritos por el Gral. Porfirio Díaz (Tesis de Licenciatura).* Puebla, México: Departamento de Ingeniería en Sistemas Computacionales, Universidad de las Américas Puebla.

Ortego Resa, C. (2009). *Detección de emociones en voz espontánea (Tesis de licenciatura).* Madrid, España: Universidad Autónoma de Madrid.

Ortony, A., Clore, G., & Collins, A. (1988). *The Cognitive Structure of Emotions.* New York: Cambridge University Press.

Planet, S., Iriondo, I., Martínez, E., & Montero, J. (2008). True: an online testing platform for multimedia evaluation. *Proceedings of the Second International Workshop on EMOTION: Corpora for Research on Emotion and Affect at the 6th Conference on Language Resources & Evaluation (LREC 2008).* Morocco: Marrakech.

Platt, J. (1998). Fast Training of Support Vector Machines using Sequential Minimal Optimization. En *Advances in Kernel Methods - Support Vector Learning.* B. Schoelkopf and C. Burges and A. Smola.

Peréz Espinosa, H., & Reyes García, C. A. (2010). *Reconocimiento de emociones a partir de voz basado en un modelo emocional continuo (Tesis de doctorado).* Puebla, México: Instituto Nacional de Astrofísica, Óptica y Electrónica.

Pérez Espinosa, H., Reyes García, C., & Villaseñor Pineda, L. (2011). EmoWisconsin: An Emotional Children Speech Database in Mexican Spanish. *Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II* , 62-71.

Podder, P., Zaman Khan, T., Haque Khan, M., & Muktadir Rahman, M. (2014). Comparative Performance Analysis of Hamming, Hanning and Blackman. *International Journal of Computer Applications , 96* (18), 1-7.

Podder, P., Zaman Khan, T., Haque Khan, M., & Muktadir Rahman, M. (2014). Comparative Performance Analysis of Hamming, Hanning and Blackman Window. *International Journal of Computer Applications , 96* (18), 1-7.

Rao Krothapalli, S., & G. Koolagudi, S. (2013). *Emotion recognition using speech features.* New York: Springer.

Reyes Vargas, M., Sánchez Gutiérrez, M., Rufiner, L., Albornoz, M., Vignolo, L., Martínez Licona, F., y otros. (2013). Hierarchical Clustering and Classification of Emotions in Human Speech Using Confusion Matrices. *Lecture Notes in Artificial Intelligence* , 162-169.

Solís Villarreal, J. F. (2011). *Un modelo de procesamiento de voz para clasificación de estados.* México D.F.: Instituto Politécnico Nacional.

Solís Villarreal, J. F., Yáñez Márquez, C., & Suárez Guerra, S. (2011). Reconocimiento automático de voz emotiva con memorias asociativas Alfa-Beta SVM. *Polibits* , 19-23.

*Speech Filing System*. (s.f.). Obtenido de UCL Psychology & Language Sciences: http://www.phon.ucl.ac.uk/resource/sfs/

*Swadesh lists for Spanish*. (s.f.). Obtenido de Wiktionary: http://en.wiktionary.org/wiki/Appendix:Spanish_Swadesh_list

Tang, H., Chu, S., Hasegawa-Johnson, M., & Huang, T. (2009). Emotion recognition from speech VIA boosted Gaussian mixture models. *2009 IEEE International Conference on Multimedia and Expo* (págs. 294-297). New York: IEEE.

the-ciu. (29 de enero de 2016). *Evolución del Mercado de Smartphones en México en 2015.* Recuperado el 28 de mayo de 2016, de The Competitive Intelligence Unit: http://www.the-ciu.net/nwsltr/479_2Distro.html

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication* , 1162-1181.

Whissell, C. (1989). The dictionary of affect in language. En R. Plutchik, & H. Kellerman, *The Measurement of Emotions* (págs. 113- 131). New York: Academic Press.

Witten, I., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition).* Burlington, USA: Morgan Kaufmann.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., y otros. (2006). *The HTK Book (for HTK Version 3.4).* Cambridge, UK: Cambridge University Press.

Zhou, Y., Yuan, J., Horta, T., Cai, W., Muraleedharan, R., & Kohl, J. (2013). *Bridge Project*. Obtenido de Wireless Communication and Networking Group: http://www.ece.rochester.edu/projects/wcng/project_bridge.html