ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

UNIVERSIDAD POLITÉCNICA DE MADRID

RECONOCIMIENTO DE EMOCIONES POR MEDIO DE VOZ

TRABAJO FIN DE MÁSTER

MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL

AUTOR: VANESSA VARGAS SANDOVAL

TUTOR: JAVIER DE LOPE ASIAIN

JULIO, 2019

DEDICATORIA

A mis padres y hermanos por su apoyo incondicional y por siempre creer en mí. A mis sobrinas Mishell, Camila y Paulette por alegrar mi corazón con cada muestra de cariño. A mis primos y amigos por compartir su cocimiento y levantar mi ánimo siempre que lo he necesitado. Sin su ayuda y comprensión sería imposible culminar esta meta.

AGRADECIMIENTO

A la Universidad Politécnica de Madrid y sus respectivas autoridades y profesores, por impartir su conocimiento y disciplina, de manera especial a Javier de Lope Asiaín que ha demostrado su experiencia y apoyo en el desarrollo del presente trabajo fin de máster. A mi familia por sus lecciones de vida, paciencia y apoyo incondicional.

RESUMEN

En el presente trabajo se realiza un análisis de la precisión en la clasificación de las emociones utilizando archivos de audio. Para ello se utiliza el extractor de características Mfcc (Mel-frequency cepstral coefficients), analizando los resultados con 13, 20, 27 Mfccs también conocidos como vectores acústicos. Los audios de entrada de los modelos propuestos son archivos de la base de datos RAVDESS, la cual tiene lecturas de audio de 8 diferentes tipos de emociones: neutral, calma, felicidad, tristeza, enojo, miedo, disgusto, sorpresa. Se testean modelos utilizando 8, 6 y 4 emociones, seleccionadas bajo diferentes criterios. Los clasificadores utilizados son SVM, KNN, RF y MPL. Obteniendo resultados de hasta 82% de precisión.

Palabras clave: clasificación de emociones por medio de voz, Mfcc, SVM, KNN, RF, MPL.

ABSTRACT

In this work an analysis of the accuracy in the classification of emotions is carried out using audio files. For this, the Mfcc (Mel-frequency cepstral coefficients) feature extractor is used, analyzing the results with 13, 20, 27 Mfccs, also known as acoustic vectors. The input audios of the proposed models are files of the RAVDESS database, which has audios of 8 different types of emotions: neutral, calm, happiness, sadness, anger, fear, disgust, surprise. Models are tested using 8, 6 and 4 emotions, selected under different criteria. The classifiers used are SVM, KNN, RF and MPL. Obtaining results of up to 82% accuracy.

Key words: speech emotion classification, Mfcc, SVM, KNN, RF, MPL.

TABLA DE CONTENIDOS

LISTADO DE FIGURAS

LISTADO DE TABLAS

1. CAPÍTULO 1

1.1 Introducción

Las emociones que experimenta el ser humano influyen en cómo vive e interactúa. Se puede decir que el ser humano está gobernado por sus emociones, ya que influyen en las decisiones, acciones y percepciones que tiene. Según el psicólogo Paul Ekman, existen 6 tipos de emociones básicas estas son: felicidad, tristeza, disgusto, miedo, enojo y sorpresa. También existe la teoría del psicólogo Robert Plutchik quien presenta una rueda de emociones que funciona en base de colores, su teoría dice que las emociones se pueden combinar y formar nuevas emociones (emociones compuestas). Las 8 emociones básicas que presenta Plutchik son (Figura 1): enojo, miedo, anticipación, felicidad, confianza, disgusto, tristeza y sorpresa (Donaldson, 2017).

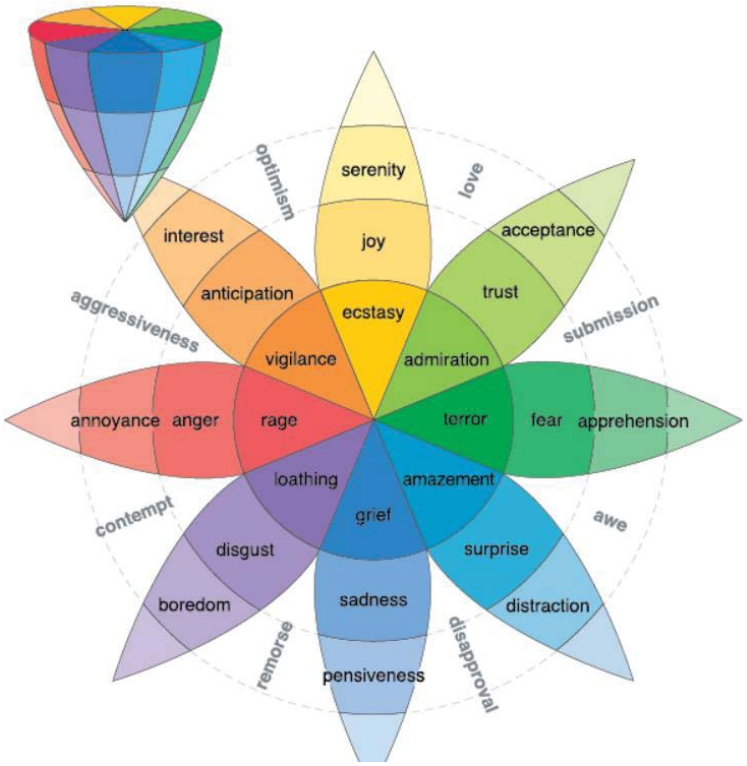


Figura 1 Estrella de emociones de Plutchik *(Donaldson, 2017)*

En la comunicación entre un humano y una máquina, las emociones juegan un papel importante. Actualmente las máquinas pueden reconocer “lo que se dice” y “quién lo dijo”, si se adiciona el reconocimiento de emociones una máquina también podrá saber “cómo se dice” para reaccionar de manera más apropiada y hacer que HCI (la interacción humana – computador) sea más agradable y natural. Otras aplicaciones que se pueden realizar con reconocimiento de emociones por medio de voz son: diagnostico psiquiátrico, juguetes inteligentes, detección de mentiras, detección oportuna de la satisfacción de un cliente, etc. (Chavan & Gohokar, 2012)

La detección de emociones en HCI es un reto difícil para el computador, especialmente cuando el reconocimiento se basa únicamente en el procesamiento de voz, pero este adquiere gran importancia ya que es un medio básico de la comunicación humana. El habla contiene tanto información lingüística como paralingüística, la primera se presenta de forma explícita, mientras que la segunda de forma implícita.

La información lingüística identifica los patrones cualitativos que el hablante ha articulado (palabras, silabas), mientras que la información paralingüística se mide por características cuantitativas que describen las variaciones en la forma en que se pronuncian los patrones lingüísticos. El último incluye variaciones de tono, acento, ritmo e intensidad, también se toman en cuenta aspectos como la calidad de voz, pausas, etc. (Anagnostopoulos, Iliou, & Giannoukos, 2012)

Las figuras 2 y 3 muestran el proceso básico para la detección de emociones por medio de voz. El primer paso es obtener la señal de voz, luego se procede a aislar información específica de la señal, principalmente de elementos como el tono, la energía, el tiempo, frecuencia y la intensidad por medio de un extractor de características. En algunos casos se han propuesto trabajos donde se reduce el conjunto de características mediante métodos de selección de características. Finalmente, al conjunto de características se lo considera la información de entrada para un método de clasificación.

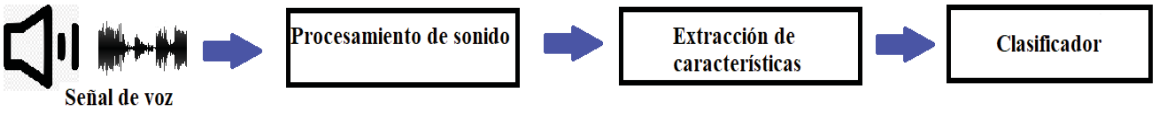


Figura 2 Proceso de reconocimiento de emociones por medio de voz- no lingüístico

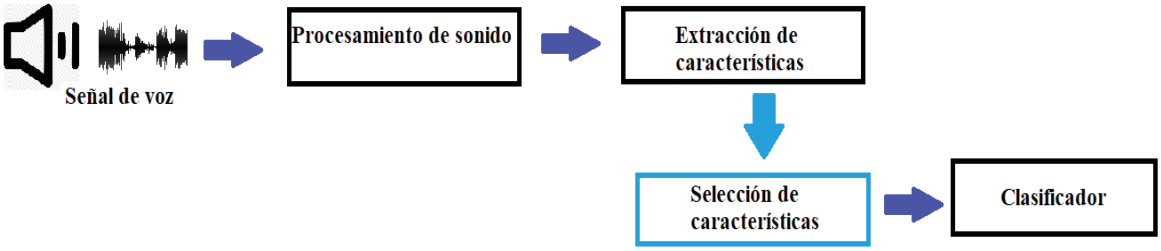


Figura 3 Proceso de reconocimiento de emociones por medio de voz - no lingüístico, con selección de características

2. CAPÍTULO 2

2.1Objetivos

2.1.1 Objetivo General

El objetivo del presente trabajo es analizar la precisión en la clasificación de emociones humanas a partir de archivos de audio, utilizando como extractor de características la herramienta MFCC (Mel-frequency cepstral coefficients).

2.1.2 Objetivos Específicos

- Realizar una revisión del estado del arte sobre las herramientas utilizadas en la detección de emociones por medio de voz.
- Elegir una base de datos apropiada que contenga audios clasificados por emociones.
- Comprender el funcionamiento del extractor de características MFFC.
- Generar data sets de las características extraídas de los audios de la base de datos de emociones que se seleccione, para entrenar modelos utilizando clasificadores supervisados.
- Testear los modelos y analizar los resultados obtenidos.

2.2 Estructura del documento

En el tercer capítulo se describirá ciertos datos preliminares necesarios para comprender el desarrollo del presente trabajo. Además, se verán las herramientas utilizadas para el reconocimiento de emociones por medio del habla en el estado del arte. En el cuarto capítulo se realizará una evaluación de los riegos que se pueden presentar al desarrollar este trabajo. Las descripciones de las herramientas que se utilizarán se presentarán en el quinto capítulo. El sexto capítulo contendrá el detalle de la creación de los modelos de reconocimiento creados, así como los resultados que estos modelos presenten al probarlos con diferentes números de coeficientes Mfcc y con distintos conjuntos de emociones. Finalmente, se concluirá sobre los resultados obtenidos y se presentarán líneas futuras para la continuidad de este trabajo.

3. CAPÍTULO 3

3.1 Datos preliminares

3.1.1 Señal de voz

El habla es la capacidad que poseen los seres humanos para comunicarse por medio de palabras. Como se había descrito antes esta posee información lingüística que permite la detección de palabras o frases y características paralingüística que permiten extraer información de la voz. La voz es el sonido producido por la vibración de los pliegues vocales de los seres vivos, tiene características como timbre, intensidad, calidad.

Las personas (articuladores) pueden producir sonidos de voz distinguibles variando la fonación, que es el proceso que convierte la presión del aire de los pulmones en vibraciones audibles, cuando el aire pasa a través de los pliegues vocales hace que estos actúen como vibrador durante la fonación (Ver figura 4).

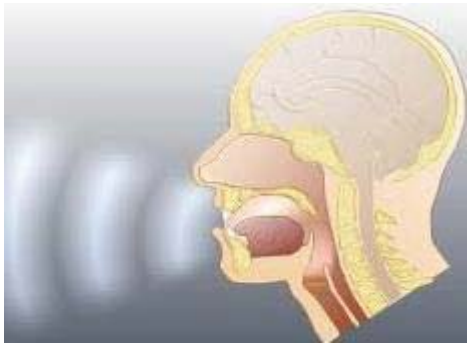


Figura 4 Generación de la señal de voz

Los pliegues vocales además poseen su propia frecuencia de resonancia que determina el tono de voz de la persona. Una frecuencia de resonancia es una frecuencia natural de vibración que se encuentra determinada por las características físicas de los pliegues. En un hombre adulto los pliegues vocales miden usualmente entre 17 y 23 mm de longitud y en una mujer adulta varía entre 12.5 y 17 mm, cada pliegue puede expandirse entre 3 o 4 mm por medio de los musculo de la laringe. Debido a las características antes mencionadas la

voz masculina oscila entre 110 Hz y la femenina entre 210 Hz. Cada fonación permite que se escape una pequeña bocanada de aire produciendo un sonido audible en la frecuencia correspondiente a la expansión de los pliegues vocales, ha este proceso se lo llama voz (Nave, 2016).

3.1.2 Procesos de reconocimiento de emociones por medio del habla

Como se ha mencionado en la parte introductoria el reconocimiento de emociones por medio del habla se ha llevado a cabo a través de métodos de procesamiento que utilizan información lingüística (explicita), paralingüística (implícita) y la combinación de ambas. La figura 5 muestra los tres procesos de reconocimiento de emociones que se describen a continuación:

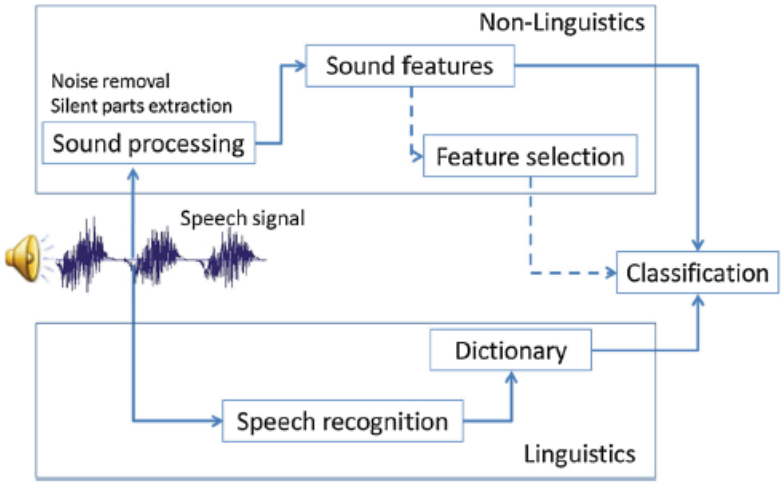


Figura 5 Proceso de reconocimiento de emociones por medio del habla *(Anagnostopoulos, Iliou, & Giannoukos, 2012)*

En el procesamiento de información lingüística se pueden identificar las siguientes herramientas: audio (información de entrada), un sistema ASR (identificación de palabras o frases) y un diccionario lingüístico. Mientras que en el procesamiento de información paralingüística se identifica 4 herramientas que son: señal de voz humana, extractor de características, selector de características (opcional) y un clasificador. En los siguientes numerales de describen brevemente cada una de estas herramientas.

- Proceso con información paralingüística

    1. Recibe como entrada una señal de voz (sonido).
    2. Procesa la señal para eliminar ruidos y sectores de la señal en el que no exista ningún tipo de información. Los extractores de características suelen realizar este proceso.
    3. A través de un extractor de características se extrae características de la señal de voz transformándola a información cuantitativa con la que se pueda trabajar posteriormente. La información de elementos como tono, energía, tiempo e intensidad son considerados generalmente para extraer estas características de las señales de voz.
    4. La selección de características es un paso opcional que ha sido propuesto en ciertos trabajos.
    5. El conjunto de características obtenido del numeral 3 o del 4, si se utiliza un selector de características, es el dato de entrada para entrenar/testear los modelos propuesto utilizando un clasificador.

- Proceso con información lingüística

    1. Recibe como entrada una señal de audio (habla).
    2. Utiliza un sistema de reconocimiento ASR (Automatic Speech Recognition), mismo que identifica y procesa el habla para reconocer palabras o frases del audio. Básicamente convierte el habla en texto.
    3. Se usa un diccionario lingüístico que contenga palabras o frases específicas que puedan ser relacionadas con cada emoción. Es importante tener un diccionario actualizado para asegurar un reconocimiento de emociones exitoso.
    4. La clasificación se realiza en base al mayor número de palabras/frases detectas que correspondan a una clase (emoción), con la ayuda del diccionario lingüístico creado en la fase de entrenamiento (Figura 6).

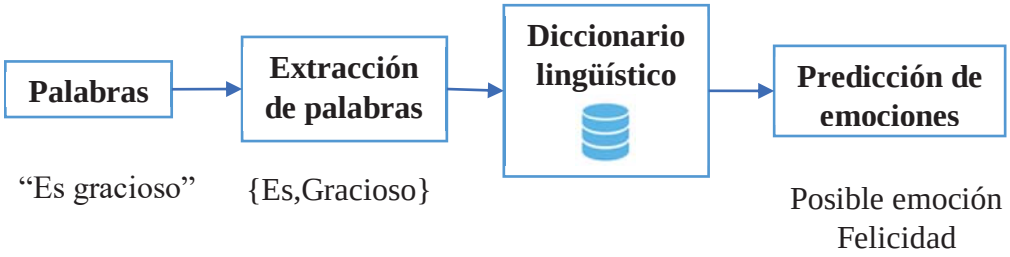


Figura 6 Detección de emociones con información lingüística

- Proceso con información lingüística y paralingüística

1. Recibe como entrada una señal de audio (habla).
2. Procesamiento de información paralingüística.
3. Procesamiento de información lingüística.

Generalmente cuando se trabaja con ambos modelos de procesamiento, se considera a la clasificación lingüística como apoyo para la clasificación paralingüística.

3.2 Estado del arte

3.2.1 Base de Datos de emociones

El reconocimiento de emociones por medio del habla utiliza archivos audio de bases de datos como datos de entrada en sus modelos. Estas bases de datos son difíciles de crear ya que deben contener audios de personas (hombres y mujeres) que representen cada emoción de forma natural.

Existen bases de datos que utilizan información de programas de televisión, programas de radio o centros de llamadas que son grabados en situaciones reales, lo cual permite captar el habla espontanea de las personas. Una las principales desventajas de estas bases de datos es que no se pueden distribuir fácilmente por problemas de copyright.

Otras bases de datos usan grabaciones realizadas por actores, siendo esta una forma sencilla de obtener bases de datos de emociones. Sin embargo, surgen ciertos cuestionamientos relacionados con la naturalidad de los resultados, pueden existir diferencias significativas entre el habla actuada y el habla espontanea si los actores no pueden capturar completamente las emociones representadas o las pueden exagerar de manera que no tengan coincidencia con las emociones captadas en un ambiente real.

Usualmente los estudios de reconocimiento de emociones por medio del habla trabajan con bases de datos grabadas por actores (ver anexo 1) debido a que existe mayor accesibilidad a estas bases de datos. Algunas emociones que se representan usualmente en estas bases son: enojo, tristeza, felicidad, neutralidad, miedo, disgusto y sorpresa (Anagnostopoulos, Iliou, & Giannoukos, 2012).

A continuación, se describen las características de cinco bases de datos de emociones encontradas en la revisión del estado del arte de esta área (Anexo 1).

- Berlin Emo, una base de datos alemana. Grabada por diez diferentes actores con diez diferentes textos, trabaja con 6 emociones. Una de las más populares en el estado del arte. (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005)

- DES (Danish Emotional Speech) una base de datos elaborada en la universidad de Dinamarca. Grabada por 4 actores un total de 500 segmentos de voz representando cinco emociones. (Pan, Shen, & Shen, 2012)

- Thai DB (Audiovisual Thai Emotion Database) grabada por 6 estudiantes, 972 palabras más comunes en Tailandia. Trabaja con seis emociones. (Seehapoch & Wongthanavasu, 2013)

- Base de datos eNTERFACE’05 trabaja con seis emociones. Contiene cuarenta y dos temas (escenarios) cada tema posee cinco oraciones grabadas por cada emoción. (Shing Ooi, Phooi Seng, Ang, & Chew, 2014)

- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) grabada por 24 actores, contiene un total de 7356 archivos de audio. Trabaja con 8 emociones. (Livingstone, Steven, & Frank, 2018). Se describe las características de la base de datos en detalle más adelante.

3.2.2 Procesamiento del audio

En el método de detección de emociones con información paralingüística, el procesamiento del audio es fundamental ya que permite obtener una representación robusta y adecuada de la señal de voz (Anagnostopoulos & Vovoli, 2009).

Generalmente una señal de voz grabada con un micrófono tiene ruido de fondo y ruidos de primer plano. El procesamiento implica la reducción de ruido, ecualización (recortar frecuencias concretas), filtrado de paso bajo (permitir el paso de las frecuencias más bajas y atenuar las frecuencias más altas) y eliminación de eventos silenciosos.

3.2.3 Extractores de características

La extracción de características es el proceso en el cual se obtiene un conjunto de características cuantificables a partir de una señal de audio. Estos extractores permiten obtener información de características específicas de la señal de audio convirtiéndolas en vectores acústicos. El conjunto de estos vectores se convierte en el dato de entrada para los clasificadores en el proceso de detección de emociones con información paralingüística. Es importante que estos conjuntos sean robustos contra el ruido y de fácil adaptabilidad (Babaee, Badrul Anuar, Abdul Wahab, Shamshirband, & T. Chronopoulos, 2015).

Las características se pueden dividir en:

1. Características temporales

Las características temporales se extraen directamente de la señal de audio, sin procedimientos previos (Chen, Mao, Xue, & Cheng, 2012) (Seehapoch & Wongthanavasu, 2013). Por lo cual, la complejidad computacional de estas características tiende a ser baja. Algunos ejemplos de extractores de características temporales son:

- Zero Crossing Rate: Representa el número de veces que la señal de audio pasa por cero en una unidad de tiempo o cuantas veces cambia la señal.

- Short-term energy: representa la intensidad de la señal en una unidad de tiempo.

2. Características espectrales

La señal de audio principalmente la señal del habla es generalmente representada por características espectrales o cepstrales. El cálculo de un Cepstral se da mediante tres procesos (Figura 7): transformada de Fourier (DFT), logaritmo y la inversa de la transformada de Fourier (FFT) que permiten identificar la frecuencia básica y la purificación discreta de una señal de audio.

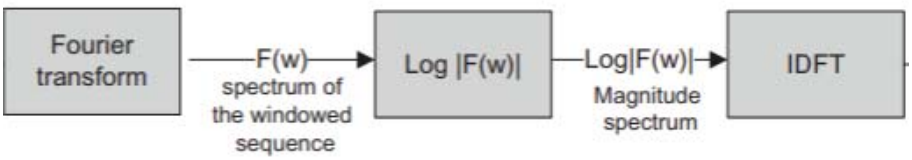


Figura 7 Computación cepstrum *(Babaee, Badrul Anuar, Abdul Wahab, Shamshirband, & T. Chronopoulos, 2015)*

A continuación, se describen ejemplos de extractores de características espectrales.

- MFCC (Mel - frequency cepstral coefficients): Calcula las bandas de frecuencia en base a la escala de frecuencia de Mel (es una escala musical perceptual del tono). Usado en (Shing Ooi, Phooi Seng, Ang, & Chew, 2014) (Kishore & Satish, 2013)

- Spectral centroid: Se denomina Spectral centroid al punto medio de la energía espectral durante la distribución de la señal.

- Linear prediction: Es una técnica que permite calcular la potencia espectral de una señal. Usado en (Kishore & Satish, 2013) (Seehapoch & Wongthanavasu, 2013).

3. Características prosódicas.

Las características prosódicas muestran información que los oyentes pueden reconocer en una señal de audio como: tono, frecuencia, volumen, intensidad, duración y ritmo (Wang, Du, & Zhan, 2008) (Navas, Hernáez, & Luengo, 2006).

Las características prosódicas también son consideradas características supra – segméntales. Las características suprasegmentales no segmentan la señal si no trabajan sobre la expresión entera. Mientras que las características temporales y espectrales son consideradas características segméntales. Las características segméntales se calculan una vez cada pequeño segmento de tiempo por lo general de 20 a 50 mseg usando técnicas windowing. (Ver Figura 8).

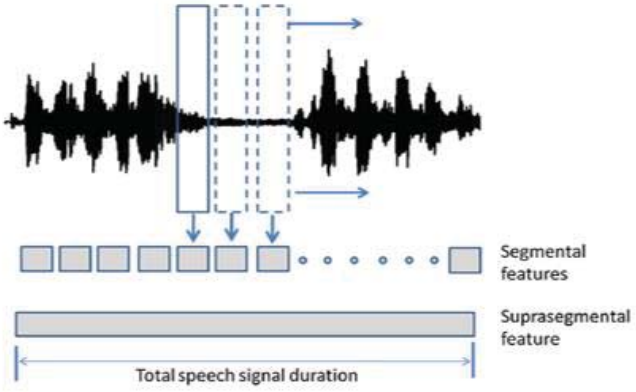


Figura 8 Características Segméntales and Suprasegmental *(Anagnostopoulos, Iliou, & Giannoukos, 2012)*

3.2.4 Selector de características

El objetivo de trabajar con un selector de características es reducir el tamaño del conjunto de características que se ha obtenido con el extractor. Para ello se selecciona un subconjunto con las características relevantes eliminando las irrelevantes.

La selección de características ha demostrados ser efectiva ya que al reducir (eliminar datos irrelevantes) la dimensionalidad de los datos de los conjuntos de características aumenta la precisión del aprendizaje de los clasificadores (Rong, Li, & Phoebe Chen, 2008). Algoritmos como Forward Selection, Greedy Algorithm, Restricted Forward Selection, Ensemble Random Forest to Trees, etc., son usados para seleccionar características en los modelos de reconocimiento de emociones por medio de voz.

3.2.5 Clasificadores

Una vez obtenidos los conjuntos de características de los audios, el siguiente paso es usar un clasificador para entrenarlo y posteriormente testear el modelo de manera que se pueda evaluar su precisión en la detección de emociones. Existen dos tipos de clasificadores utilizados para detectar emociones por medio de voz (Anexo 1), clasificadores supervisados y clasificadores no supervisados.

- Clasificadores supervisados: en un clasificador supervisado los datos de entrenamiento deben pertenecer a una clase específica (datos etiquetados) y el algoritmo aprende de las características de estos datos encontrando patrones por clase. Clasificadores como Support vector machine (SVM) (A, Roy, & Selvi, 2013), Random forest (RF) (Rong, Chen, Chowdhury, & Li, 2007), etc. son utilizados en el proceso de detección de emociones por medio de voz. El clasificador SVM es uno de los más utilizado en los estudios realizados, obteniendo resultados de precisión de hasta 97% (Anexo 1).

- Clasificadores no supervisados: los clasificadores no supervisados crean agrupaciones con los datos de entrada en base a sus similitudes, por lo cual los datos de entrenamiento no deben estar etiquetados. Algunos de los clasificadores no supervisados usados en la detección de emociones por medio de voz son: Gaussian mixture model (GMM) (Neiberg, Elenius, & Laskowski, 2006), Hidden-markov-model (HMM) (Yun & Yoo, 2009), Convolutional neural network (CNN) (Lim, Jang, & Lee, 2016), etc. Los modelos que utilizan estos clasificadores presentan resultados de precisión de hasta 95% de precisión (Anexo1).

3.2.6 Sistema de reconocimiento automático de palabras

Un sistema ASR, es un proceso que permite a una máquina interpretar el habla humana, es decir convierte la señal del habla en una cadena de palabras (Forsberg, 2003) (Navas, Hernáez, & Luengo, 2006) (Schuller, Villar, Rigoll, & Lang, 2005). Los humanos no usan solo sus oídos cuando escuchan, sino también su conocimiento sobre el hablante, el tema, conexiones gramaticales, modismos, etc. Por lo cual un humano puede interpretar rápidamente lo que escucha, esto no sucede con la máquina. Cuando se trabaja con ASR varios retos se hacen presentes, a continuación, se enumeran algunos de ellos:

1. Los seres humanos comúnmente hablan con errores gramaticales y semánticos.

2. El habla no tiene pausas naturales entre cada palabra, generalmente estas pausas aparecen después de una frase u oración.
3. Existen varios dialectos tanto por región, como dialectos sociales.
4. Las palabras homófonas de los lenguajes.

A continuación, se nombran varios kits de herramientas que permiten convertir el audio en texto como: CMU Sphinx, Google Cloud Speech API, Microsoft Bing Voice Recognition, IBM Speech to Text, etc. (Python Software Foundation, 2019)

3.2.7 Diccionarios lingüísticos

Un diccionario lingüístico es una recopilación de palabras o frases que representan una clase. Una vez que se convierte el audio en texto mediante un sistema ASR se puede identificar palabras o frases que se presentan frecuentemente (medir la cantidad de veces que los términos se repiten) en una clase o dominio e insertarlas en un diccionario (vector) que represente la clase, es importante descartar las palabras vacías (silencio) o stopwords.

Una vez que se tiene los vectores con las palabras o frases que representen a cada clase, se puede testar el modelo. El sistema ASR convierte al audio en texto y con ayuda del diccionario se puede comparar las palabras del audio ingresado con las que contienen los vectores de cada clase y el audio pertenecerá a la clase con la que posea más palabras o frases en común.

Se puede además trabajar con reglas gramaticales que permitan identificar verbos, conectores, sustantivos, adjetivos, etc. para eliminar palabras que no aporten valor a los diccionarios. Otro aspecto que se puede considerar es los sinónimos de las palabras que están en los diccionarios y en el caso de las frases se puede considerar el parafraseo (Las paráfrasis son oraciones o frases que transmiten el mismo significado utilizando una redacción diferente).

4. CAPÍTULO 4

4.1 Evaluación de Riesgos

Como se ha descrito previamente, las herramientas básicas que se necesitan en el proceso de reconocimiento de emociones por medio de voz son:

- Base de datos de emociones
- Extractor de características
- Clasificador

La selección del extractor de características se realiza en base a una breve revisión sistemática del tema (Anexo 1), donde se puede observar que el extractor Mfcc es frecuentemente utilizado en varios modelos propuestos, mismos que han obtenido buenos resultados (superiores al 75% de precisión).

Varios clasificadores serán probados para comparar los resultados de precisión que se obtengan con cada uno de ellos. Se debe variar los parámetros de entrada del clasificador hasta encontrar la mejor combinación estos, para que se entregue buenos resultados.

Tanto el clasificador como el extractor de características dependen de la calidad de los datos de entrada, es decir la base de datos de emociones seleccionada. El desafío es encontrar una base de datos que proporcione:

- Archivos de calidad, es decir, sin ruido, variaciones en tono, intensidad, que representen correctamente las diferentes emociones, etc. Para que el extractor de características pueda proporcionar un conjunto de datos robusto.

- Suficientes muestras para que el conjunto de datos de entrenamiento sea apropiado y el clasificador pueda distinguir emociones eficientemente.

5. CAPITULO 5

5.1 Herramientas utilizadas

5.1.1 Base de datos

La base de datos RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) (Livingstone & Russo, 2018) se ha seleccionado para el desarrollo del presente trabajo. RAVDESS posee la licencia Creative Commons de uso no comercial. Contiene 7356 registros, cada uno de estos validado en tres aspectos: intensidad, autenticidad y emocional.

En la creación de esta base de datos participaron 24 actores profesionales, 12 mujeres y 12 hombres (Ver figura 9), todos ellos con un acento estadounidense neutral. Los actores realizaron 104 grabaciones distintas, que consistían en 60 enunciados hablados y 44 enunciados cantados. La tabla 1 detalla todas las características de estos archivos.

Figura 9 Ejemplo de grabaciones de 6 emociones en audio. *(Livingstone & Russo, 2018)*

No	Característica	Opciones
1	Modalidad	1 audio y video 2 solo video 3 solo audio
2	Canal Vocal	1 voz 2 canción
3	Emoción	1 neutral 2 calma 3 felicidad 4 tristeza 5 enojo 6 miedo 7 disgusto 8 sorpresa
4	Intensidad emocional	1 neutral 2 fuerte
5	Frases	1 “Kids are talking by the door” 2 “Dogs are sitting by the door”
6	Se realiza una repetición por cada grabación	

Tabla 1 Descripción de los archivos de audio de la base de datos RAVDESS

Los archivos de solo audio se dividen en dos carpetas, “Speech file” (archivos de habla) y “Song file” (audios de canciones). En este caso, como el objetivo del presente trabajo es reconocer emociones por medio de voz, únicamente se usará los archivos de audio de la carpeta “Speech file”.

Se tiene entonces 1440 archivos de audio en total, 192 por cada emoción excepto por la emoción neutral que posee 96 audios, ya que estos se han grabado únicamente con intensidad neutral. La mitad de estos audios pertenecen a voces masculinas y la otra mitad a voces femeninas. La extensión de los audios es .WAV y tienen una duración aproximada de 3 segundos.

5.1.2 Extractor de características Mfcc

Los Mfcc (coeficientes espectrales de las frecuencias en la escala Mel) son una técnica de extracción de características muy usada en el reconocimiento de voz. Mfcc muestra las características de la señal de voz relacionadas al tracto vocal (numeral 2.1.1). A continuación, se describirá los pasos para obtener un vector Mfcc (Ver Figura 10):

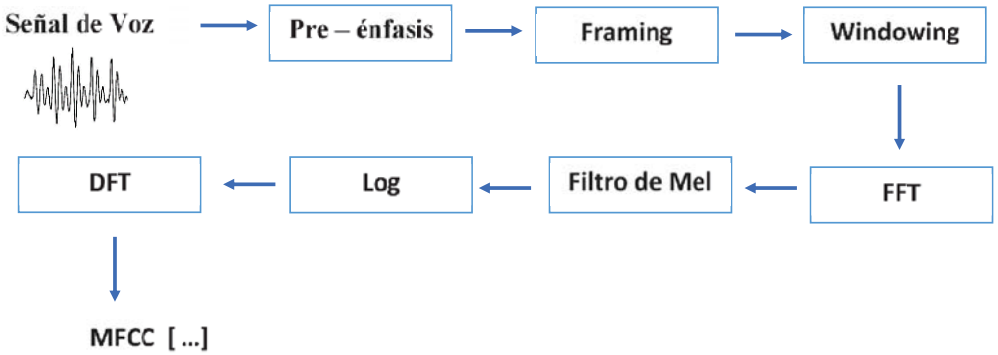


Figura 10 Proceso Mfcc

1. Pre – énfasis (pre - emphasis): la fase de pre – énfasis consiste en pasar la señal por un filtro que enfatice las frecuencias altas, esto permitirá obtener más información de la señal (las frecuencias altas contienen más información que dependen del hablante) y equilibrar el espectro de frecuencias.

   La energía de los componentes de alta frecuencia en las señales de voz es generalmente baja. El pre énfasis se utiliza para aumentar la energía de los componentes de alta frecuencia. El filtro de pre – énfasis se puede aplicar a una señal X mediante la siguiente ecuación:

$$Y(t) = X(t) - a\,X(t-1)$$

   Donde a suele tomar valores en el intervalo [0.95, 0.97]

También podremos ver, en la figura 11, un ejemplo (amplitud - frecuencia) de una señal de voz antes y después de la fase de pre – énfasis. Los componentes de alta frecuencia de la señal de voz poseen una amplitud pequeña respecto a los componentes de baja frecuencia. Podemos concluir que el filtro aumenta los componentes de alta frecuencia y atenúa los componentes de baja frecuencia

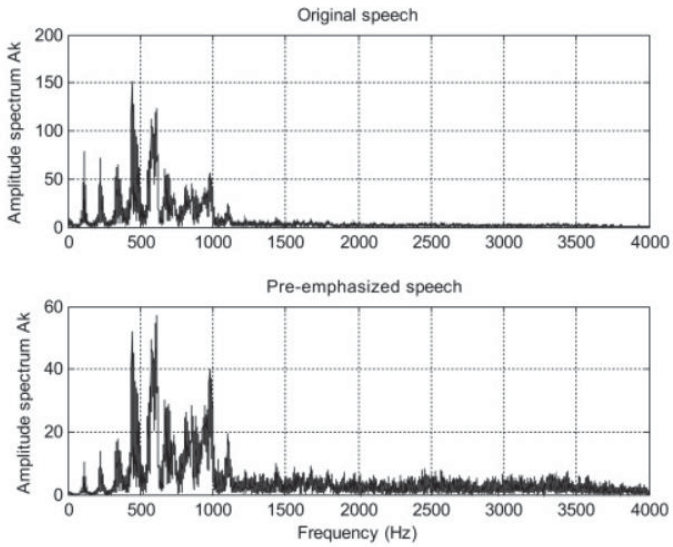


Figura 11 Señal de voz antes y después de pre – énfasis *(Tan & Jiang, 2013)*

2. Framing: Consiste en dividir la señal en varios frames de trozos de tiempo corto. El ancho de los frames generalmente son de 20 a 30 ms con una superposición entre 40% - 60%. Cada frame posee N puntos de muestras. Por lo tanto, los frames no pueden ser muy cortos ya que no se obtendrá muestras suficientes para una estimación espectral confiable y no puede ser demasiado largo ya que la señal cambia demasiado.

   La figura 12 muestra una señal de voz dividida por varios frames, donde se puede observar también la superposición (rojo) entre ellos.

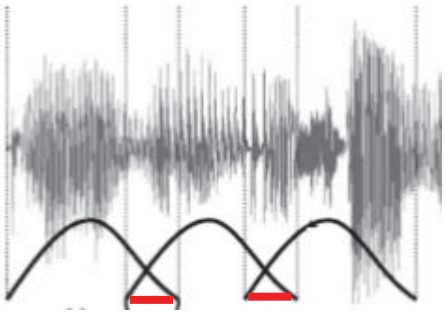


Figura 12 Framing Mfcc *(YANKAYIŞ, 2019)*

3. Windowing: En la mayoría de los sistemas donde se tiene superposición de frames se debe suavizar la transición de un frame a otro, es decir se quiere mantener la continuidad entre frames. Por ello en esta fase cada frame es multiplicado con una función de Hamming Windowing. Básicamente lo que hace la función es reducir el trozo de la señal de voz a cero tanto al principio como al final de cada frame (Ver figura 13).

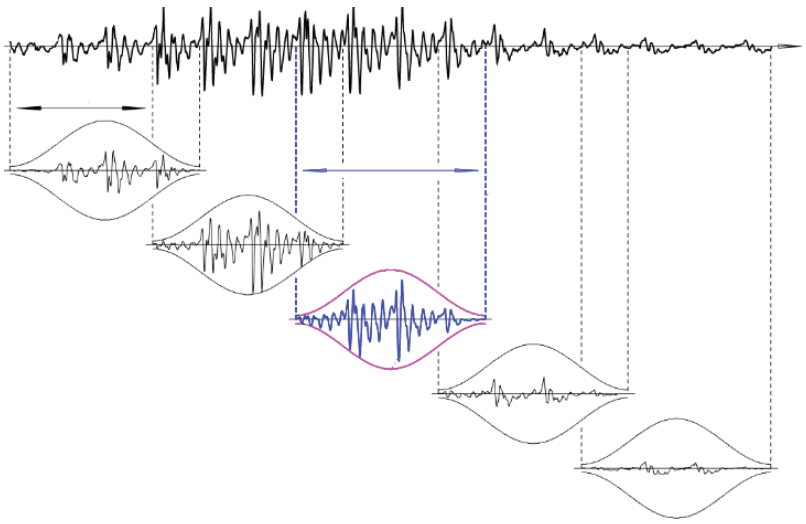


Figura 13 Hamming Windowing *(Shmyrev, 2016)*

La función Hamming Windowing está dada por la siguiente ecuación:

$$W[n] = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) & \text{donde } 0 \leq n \leq N-1 \\ 0 & otro\ caso \end{cases}$$

Siendo W (n) = Hamming Windowing

N = número de muestras en cada frame

La señal de salida Y será igual a la señal de entrada de cada frame X multiplicada por la función Hamming Windowing.

$$Y(n) = X(n) * W(n)$$

4. FFT (Fast Fourier Transform): Este proceso permite extraer la magnitud de frecuencia de cada frame. X[k] es un número complejo que representa la magnitud y la fase de ese componente de frecuencia en la señal original, X[k] se define de la siguiente ecuación:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-2\frac{\pi}{N}kn} \qquad k= 0, 1\ldots N\text{-}1$$

Donde:

x [n] = señal Hamming Windowing

5. Procesamiento de Mel – Filter bank: MFCC utiliza la escala de frecuencia de Mel, la cual que se basa en percepciones auditivas humanas. El ser humano es más discriminativo en las frecuencias más bajas y menos en las frecuencias más altas. Entonces se tiene dos tipos de filtros de Mel, los primeros están espaciados linealmente a baja frecuencia por debajo de 1000 Hz y los segundos están espaciados

logarítmicos por encima de 1000 Hz. Por ello, existen más filtros en las regiones de baja frecuencia y menos filtros en las regiones de alta frecuencia (Ver figura 14).

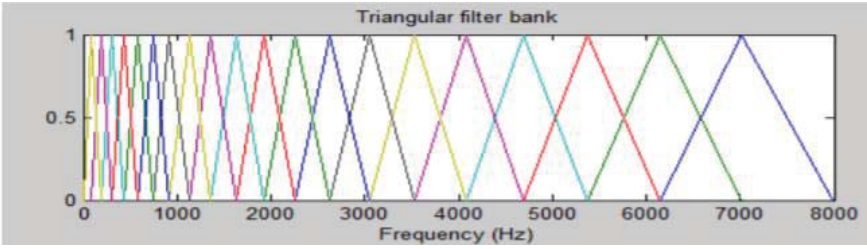


Figura 14 Banco de filtros de Mel *(Pal Singh & Rani)*

El objetivo de esta fase es multiplicar la magnitud de frecuencia, obtenida en la fase anterior, con un conjunto de 20 - 40 filtros de Mel, cada salida de filtro es la suma de los componentes filtrados (Ver figura 15).

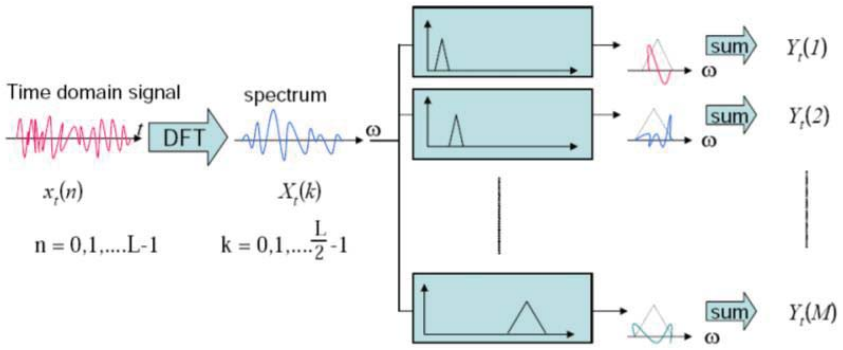


Figura 15 Fase FFT del proceso Mfcc

Se puede transformar la frecuencia (f) a frecuencias en escala de Mel (m) utilizando la siguiente ecuación:

$$m = 2595 * \log 10 \left(1 + \frac{f}{700}\right)$$

6. Log: El objetivo de este punto es calcular el logaritmo de la magnitud de frecuencia en escala de Mel elevada al cuadrado (Ver figura 16). Esta fase hace que las estimaciones de frecuencia sean menos sensibles a ligeras variaciones de la señal.

Por ejemplo, la variación en la potencia de los audios debido a la distancia entre el locutor y el micrófono.

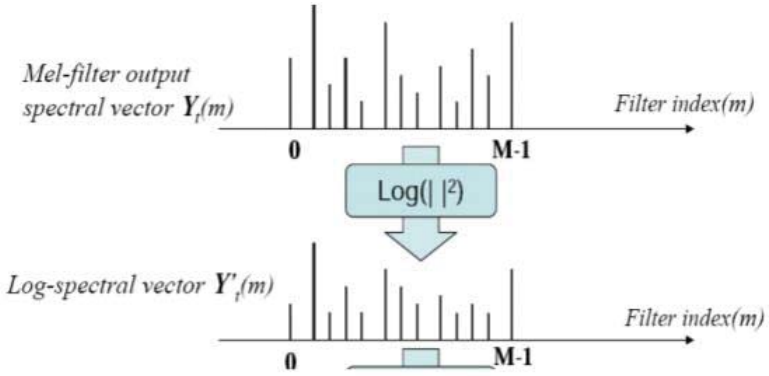


Figura 16 Fase log del proceso de MFcc

7. DCT (Discrete cosine transform): esta fase permite transformar la señal del dominio de frecuencia al dominio de tiempo. El resultado de esta conversión se llama Mfcc, el conjunto de coeficientes Mfcc es conocido como vector acústico. Esta transformación está dada por la fórmula:

$$y_t[k] = \sum_{m=1}^{M} \log(|Y_t(m)|^2)\cos\left[k * (m - 0.5) * \frac{\pi}{M}\right]$$

$$k = 1, 2, \ldots, J$$

Donde:

M = número de filtros de Mel

J = número de Mfcc

6. CAPÍTULO 6

6.1 Descripción de modelos creados

Inicialmente se probarán 8 modelos, cada modelo utilizará un clasificador diferente, con 20 coeficientes Mfcc extraídos por cada audio, utilizando las 8 emociones que ofrece el data set. Se eliminarán aquellos modelos obtengan resultados de precisión muy bajos. Luego, se compararán los resultados de los modelos elegidos variando los datos de entrada de los clasificadores al extraer 13, 20 y 27 coeficientes por audio. Cuando se obtengan estos resultados se utilizará el número de coeficientes Mfcc y los modelos que entreguen los mejores resultados para las pruebas posteriores.

A continuación, se seleccionarán conjuntos de 6 y 4 emociones para analizar los resultados con los modelos seleccionados. Primero se seleccionarán emociones para crear estos conjuntos en base a criterios que se describirán posteriormente. Luego, se crearán conjuntos de emociones en base a la tasa de verdaderos positivos que estas obtengan al analizar la matriz de confusión obtenida al testear las 8 emociones.

6.2 Resultados de pruebas

6.2.1 Clasificación con 20, 13 y 27 vectores acústicos por audio.

Inicialmente se ha probado los modelos con 8 clasificadores (ver Figura 17), donde se puede observar que 4 de ellos entregaron resultados de precisión bajos (color naranja) por lo cual serán omitidos en las siguientes pruebas.

Para todas las pruebas que se realizaran en adelante, el porcentaje de datos tomado para entrenamiento es 80% y 20% para testeo, este procedimiento se repetirá 100 veces y se conseguirá la media de la precisión obtenida en cada procedimiento.

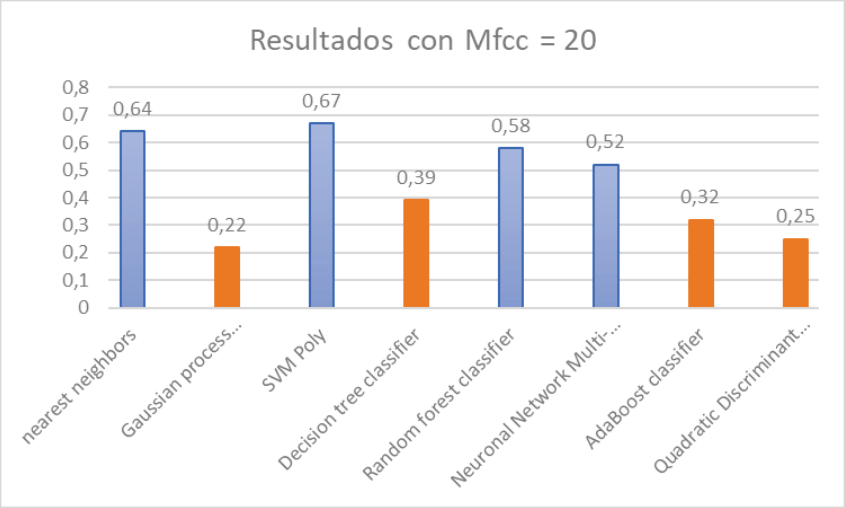


Figura 17 Resultados 8 clasificadores con Mfcc = 20

Trabajaremos entonces con los clasificadores SVM (support vector machine) con núcleo lineal y poly, k-nearest neighbors, Random forest y Neuronal Network Multi-layer Perceptron, que han sido los clasificadores que han entregado los mejores resultados. Para la implementación de los clasificadores se ha utilizado el toolkit de Python scikit. Los parámetros de cada clasificador son los que el toolkit auto asigna, exceptuando los siguientes:

- KNN:

  n_neighbors (número de vecinos) = 1

  weights (función de peso) = ‘distance’

  algoritmo = ‘brute’ (algoritmo de fuerza bruta)

  p = 1 (manhattan_distance)

- Random Forest:

  random_state (generar números aleatorios) = 20

  max_features (el número de características a considerar) = 'log2'

  Bootstrap (muestras para construir el árbol) = False (todo el data set).

- Multi - layer Perceptron:

  hidden_layer_sizes (número de neuronas en la capa oculta) = 200.

Ahora se probarán lo 4 clasificadores seleccionados variando el número de coeficientes de Mfcc, para observar que impacto tiene el reducir o aumentar coeficientes en la precisión de la clasificación. Para la extracción de características se ha utilizado la librería librosa, esta es una librería de audio de Python. Los valores de los parámetros de entrada del extractor de características Mfcc fueron:

- **y** = (datos de entrada)
- **sr** = 22050 (sampling rate of y)
- **S** = None (log power Mel spectrogram)
- **n_mfcc** = 20 / 13 / 27 (number of MFCCs to return)
- **dct_type** = 2 (Discrete cosine transform (DCT) type)
- **norm** = 'ortho' (Normalization)

Los conjuntos de datos Mfcc entregados son vectores acústicos de dimensiones [1536,2800], [1536,1820], [1536,3780] cuando n_mfcc igual a 20, 13, 27 respectivamente. Se toman estos vectores como el conjunto de datos de entrada para probar los clasificadores.

La tabla 2 muestra los resultados de precisión (%) obtenidos con 13, 20 y 27 coeficientes Mfcc y la media de tiempo en segundos (t(s)) que toma al clasificador entregar los resultados.

Clasificador	**Mfccs 20**		**Mfccs 13**		**Mfccs 27**	
	t (s)	%	t (s)	%	t (s)	%
SVM Lineal	5,9	64	3,7	62	8,5	65
SVM Poly	7,5	67	4,1	65	9,1	68
KNN	1,2	64	0.79	63	1,7	65
RF	1,2	58	1,1	58	1,3	57
MLP	5,8	51	3,6	46	7,6	54

Tabla 2 Resultados de la clasificación con Mfccs = 20 / 13 / 27

Como se puede observar la precisión más alta que se ha obtenido en la clasificación es de 68% cuando se utilizan 27 vectores acústicos por audio con el clasificador SVM (Poly). En general los resultados obtenidos con 27 vectores acústicos son mejores que cuando se usan 13 o 20 (Ver figura 18), pero el tiempo que tarda en realizar la clasificación significante (Ver Figura 19).

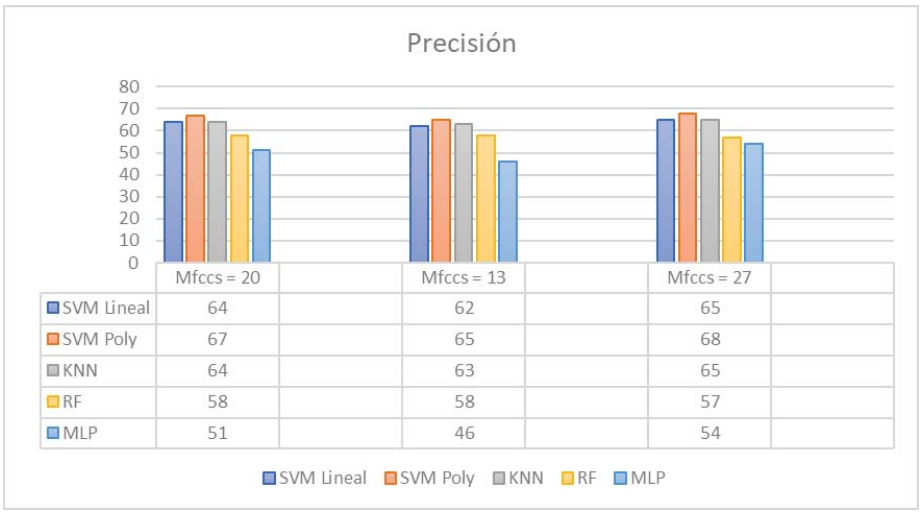


Figura 18 Precisión obtenida en la clasificación con Mfccs = 20, 13 y 27

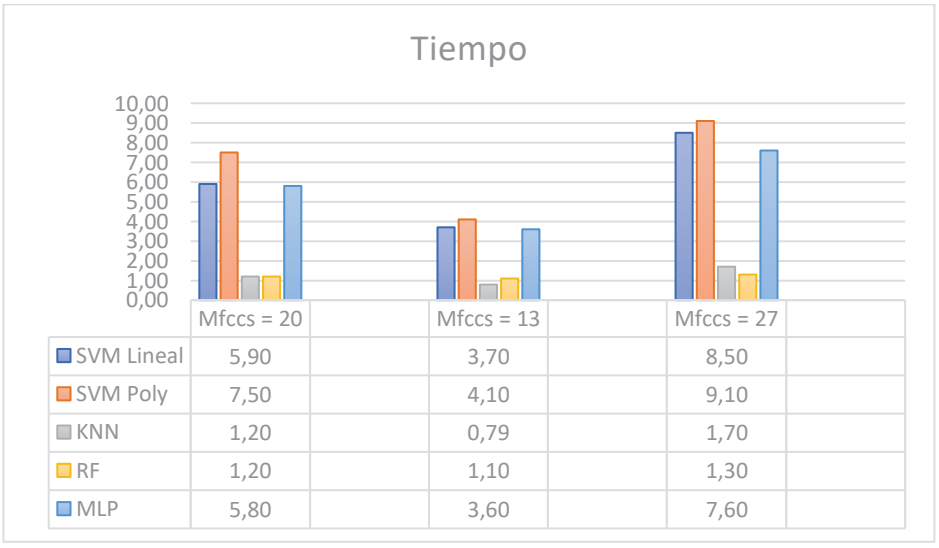


Figura 19 Tiempo transcurrido durante la clasificación con Mfccs = 20, 13 y 27

Como muestra la figura 18 la precisión varía entre 1% – 3 % cuando se trabaja con 20 y 27 vectores acústicos, mientras que con 20 y 13 vectores la precisión varía entre 2% - 5%. Aunque se obtiene mejores resultados con 27 vectores, la figura 19 muestra que el tiempo de detección también aumenta.

La mayoría de los resultados de precisión con 27 coeficientes Mfcc aumenta únicamente un 1%, por lo que no justifica trabajar con 27 vectores acústicos si los resultados con 20 vectores son muy similares y toma menos tiempo. Podemos concluir que usar 20 vectores acústicos permite obtener un buen resultado de precisión en menor tiempo.

6.2.2 Clasificación con 8, 6 y 4 emociones.

En esta sección, para realizar las pruebas se trabaja con 20 vectores acústicos por audio ya que, como se ha visto en el numeral anterior entrega buenos resultados en menor tiempo. Ahora se analizará la precisión que arrojan las clasificaciones variando el número de emociones en el data set de entrada. La selección de emociones se realiza en base a los siguientes criterios:

- Seleccionar 6 emociones omitiendo aquellas que son parecidas según la percepción humana y que posiblemente se puedan distinguir mejor si se tiene información visual y no solo auditiva. Estas emociones son: neutral – calma y enojo – disgusto (Ver figura 20).

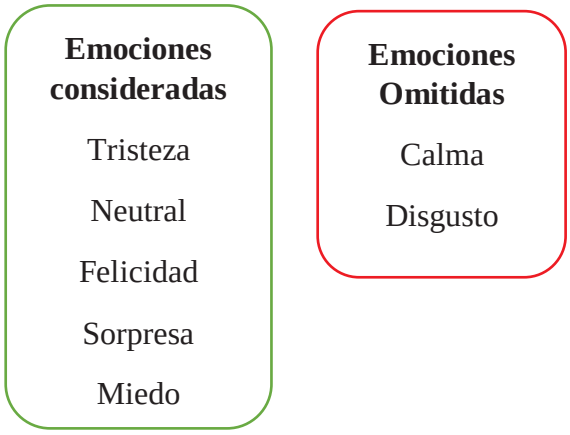


Figura 20 Emociones consideradas y emociones omitidas – testeo con 6 emociones

- Trabajar con 4 emociones que se encuentren opuestas en la estrella de emociones de Plutchik (Ver figura 21 y 22). Al ser emociones que se encuentren en direcciones opuestas es más fácil distinguir unas de otras para una persona y se espera que los vectores acústicos de estas emociones también sean distinguibles de modo que se pueda mejorar los resultados de clasificación.

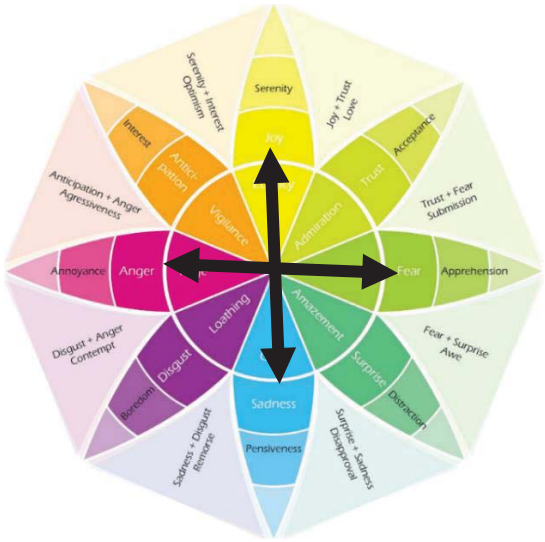


Figura 21 Estrella de emociones de Plutchik - selección de 4 emociones (Infographics, 2015)

Emociones consideradas	**Emociones Omitidas**
Enojo	Calma
Tristeza	Disgusto
Felicidad	Neutral
Miedo	Sorpresa

Figura 22 Emociones consideradas y emociones omitidas – testeo con 4 emociones

Los resultados de las pruebas realizadas se muestran a continuación:

La tabla 3 muestra los resultados de precisión (%) de las clasificaciones con 8, 6 y 4 emociones y la media tiempo en segundos (t(s)) que tarda 100 clasificaciones. Como se observa en la figura 23 la diferencia entre los resultados de precisión cuando se trabaja con 8 y 6 emociones varia de 3 % hasta 10%, la diferencia con 8 y 4 emociones varía entre 6% - 10% y la diferencia entre 6 y 4 emociones oscila entre 3% - 5%.

Algo interesante que se puede destacar de esta clasificación es que el clasificador MPL no incrementa la precisión de la clasificación cuando se utilizan 6 y 4 emociones, si no que entrega el mismo resultado. Pero su precisión cuando se trabaja con 6 o 4 emociones aumenta el 10% que cuando se trabaja con 8 emociones, siendo este el valor más alto de cambio que se puedo observar.

Los clasificadores SVM (Poly), SVM (Linear) aumentan la precisión de la clasificación de forma similar, cuando trabajan con 6 emociones su precisión aumenta 3% y cuando trabajan con 4 emociones su precisión aumenta 8% respecto a la clasificación con 8 emociones.

Clasificador	**8 emociones**		**6 emociones**		**4 emociones**	
	t (s)	%	t (s)	%	t (s)	%
SVM Lineal	5,9	64	3,4	67	1,6	72
SVM Poly	7,5	67	3,5	70	1,8	75
KNN	1,2	64	0,72	67	0,3	70
RF	1,2	58	0,9	62	0,5	65
MLP	5,8	51	6,4	61	2,8	61

Tabla 3 Resultados de clasificaciones con 8, 6 y 4 emociones

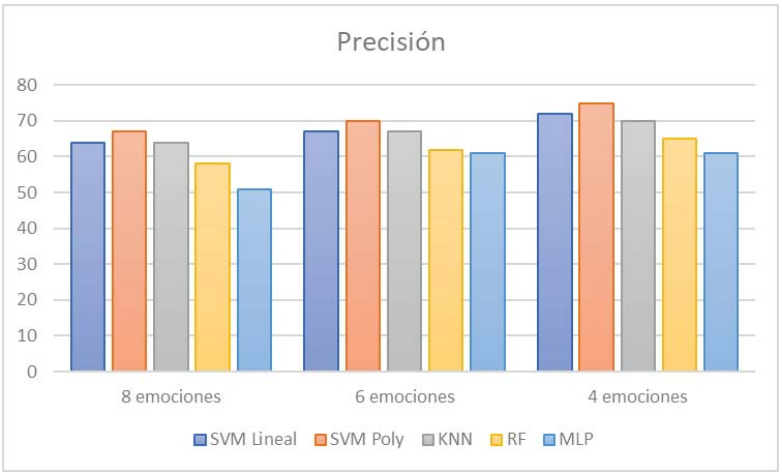


Figura 23 Precisión obtenida en la clasificación con 8,6 y 4 emociones.

En cuanto a los tiempos que toma la clasificación disminuye considerablemente cuando disminuye el número de emociones (Ver figura 24), esto podría deberse principalmente a que disminuyen en número de datos de entrada, estos datos son 1536 con 8 emociones, 1152 con 6 emociones y 768 con 4 emociones.

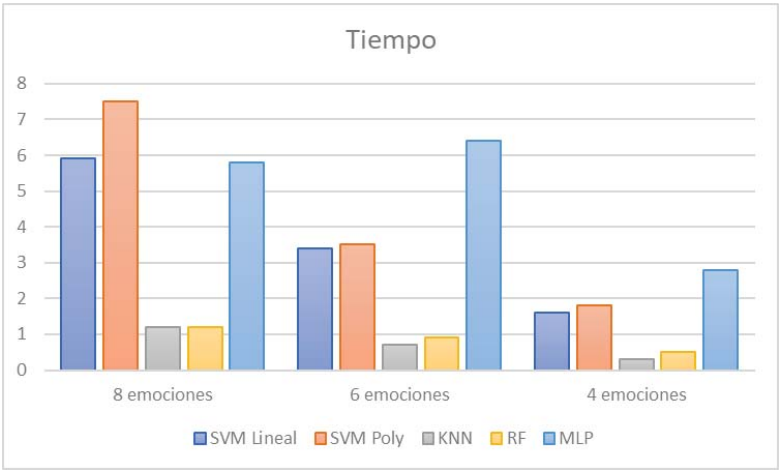


Figura 24 Tiempo transcurrido durante la clasificación con 8,6 y 4 emociones.

6.2.3 Análisis de matriz de confusión

En esta sección analizaremos las matrices de confusión de la clasificación de emociones con el clasificador SVM – Poly ya que con este se han obtenido mejores resultados de precisión en las pruebas previas. La tabla 4 muestra la matriz de confusión que se ha obtenido al realizar una clasificación con 8 emociones. Aclaramos que en esta fase se analiza un proceso de clasificación, contrario a los resultados previos donde se realizaban 100 clasificaciones.

Matriz de Confusión

	Calma	**Disgusto**	**Enojo**	**Felicidad**	**Miedo**	**Neutral**	**Sorpresa**	**Tristeza**
Calma	**29**	0	0	0	0	0	0	3
Disgusto	0	**29**	8	1	1	0	1	3
Enojo	0	3	**29**	3	2	0	2	1
Felicidad	5	2	5	**23**	0	2	1	3
Miedo	4	2	1	1	**18**	1	2	2
Neutral	0	0	0	0	0	**38**	0	2
Sorpresa	0	2	1	7	5	4	**19**	1
Tristeza	9	4	0	2	1	4	0	**19**
Precisión General	0.66							

Tabla 4 Matriz de confusión - 8 emociones – SVM Poly

Se observa en la matriz de confusión de 8 emociones (Tabla 4) que varios ejemplos de enojo son detectados como disgusto, por lo cual la inferencia que se realizó en la sección 3.2.2 sobre estas dos emociones fue correcta, contrario a lo que pasa con las emociones calma y neutral que no tienen una mala clasificación entre ambas. También se puede observar que

algunos de los ejemplos de calma son detectados como tristeza y felicidad y ejemplos de felicidad son detectados como sorpresa. Ahora calcularemos la precisión de cada emoción y la tasa de verdaderos positivos (TVP) por cada emoción, los resultados se muestran en la tabla 5.

- Precisión: Es la calidad de la respuesta positiva del clasificador. (Se toma en cuenta que a la clase no se le asignen elementos de otra clase)

$$Precisión = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos}$$

- TVP: Mide la eficiencia en la clasificación de todos los elementos que son de la clase. (Se toma en cuenta que a otras clase no se le asignen elementos de la clase)

$$TVP = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos}$$

Emoción	Precisión	TVP
Calma	0,91	0,62
Disgusto	0,67	0,66
Enojo	0,73	0,66
Felicidad	0,64	0,62
Miedo	0,58	0,67
Neutral	0,95	0,78
Sorpresa	0,49	0,76
Tristeza	0,49	0,56

Tabla 5 Precisión y Tasa de verdaderos positivos - 8 emociones - SVM Poly

Se puede observar en la figura 25 que clase que obtiene mayor TVP es Neutral, seguido por la clase Sorpresa. Mientras que las clases que mejor precisión obtienen es Neutral y Calma. Podemos concluir en la base de datos (RAVDESS) que hemos utilizado, la clase que mejor se clasifica tanto en precisión como TVP es Neutral.

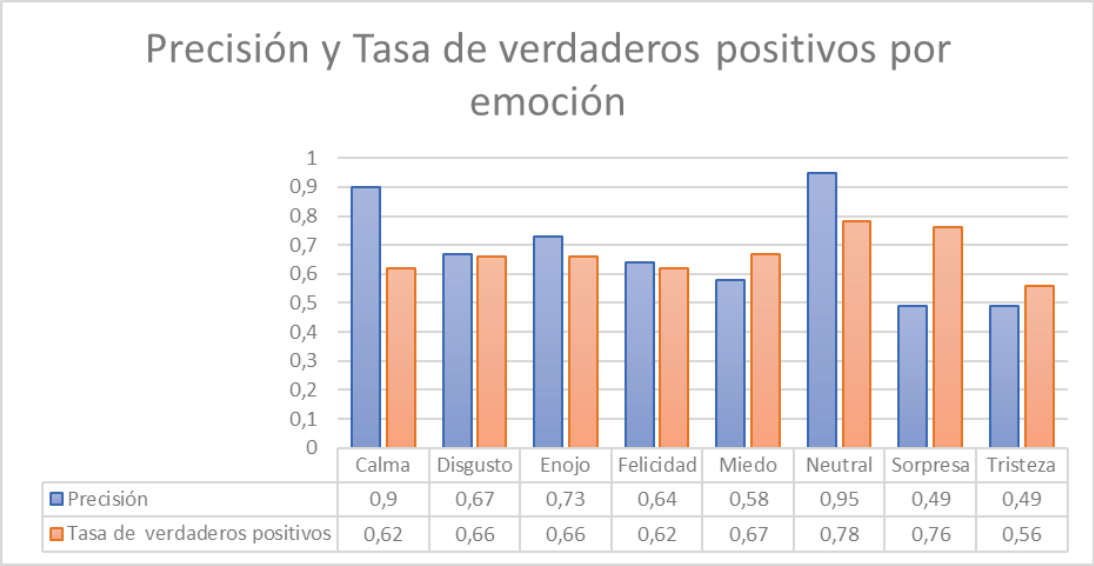


Figura 25 Precisión y Tasa de verdaderos positivos - 8 emociones - SVM Poly

Por otro lado, se puede observar en la figura 25 que aunque la clase Sorpresa tiene una alta TVP, esta posee un bajo valor de precisión esto se debe a que se le asignan 20 ejemplos que no pertenecen a esta, 7 de estos 20 elementos pertenecen a la clase Felicidad. Esto quiere decir que los ejemplos de felicidad tienden a confundirse con ejemplos de sorpresa.

6.2.4 Selección de emociones de acuerdo con las tasas de verdaderos positivos (TVP) obtenidas.

En esta sección se realizará pruebas de clasificación seleccionando las emociones que mejor TVP hayan obtenido, intentando así encontrar el conjunto de emociones con el que el clasificador entregue mejores resultados. Esto ayudará como base para trabajos futuros si se desea utilizar las mismas herramientas (base de datos, extractor de características y

clasificador). Las pruebas se realizarán bajo la modalidad que se ha trabajado en el numeral 3.2.2.

Según la tabla 5 podemos elegir las 6 emociones que tienen mejor TVP en orden descendente:

1. Neutral
2. Sorpresa
3. Miedo
4. Enojo
5. Disgusto
6. Calma

Veamos los resultados que se obtienen al realizar la clasificación con estas 6 emociones:

Clasificador	**6 emociones**	
	t (s)	%
SVM Linear	5,7	75
SVM Poly	3,5	77
KNN	1,1	75
RF	1,7	70
MLP	7,6	69

Tabla 6 Resultados de la clasificación con 6 Emociones con mejor TVP

Veamos los resultados que se obtienen al realizar la clasificación con estas 4 emociones:

1. Neutral
2. Sorpresa
3. Miedo
4. Enojo

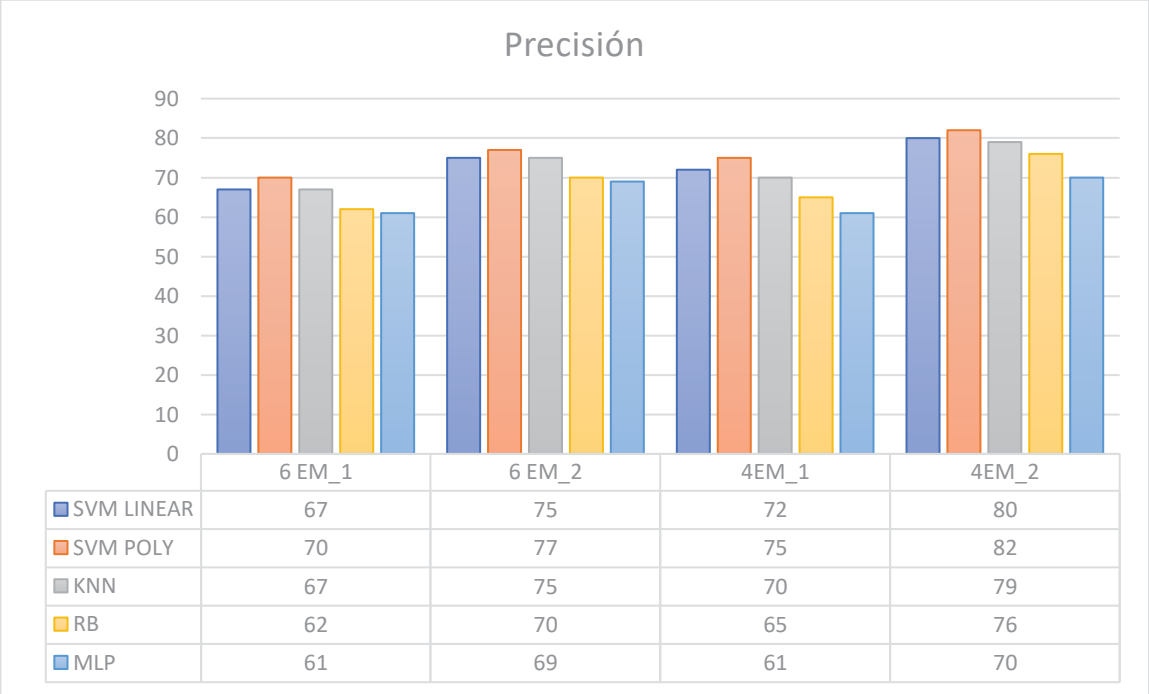
Clasificador	4 emociones	
	t (s)	%
SVM Linear	2,4	80
SVM Poly	1,15	82
KNN	0,3	79
RF	0,5	76
MLP	3,8	70

Tabla 7 Resultados de la clasificación con 4 Emociones con mejor TVP

Como se puede observar en las tablas 6 y 7 la precisión de la clasificación con 6 y 4 emociones con mejor TVP es mayor que la precisión obtenida en las clasificaciones realizadas en el numeral 3.2.2. Estos resultados serán comparados y analizados en la siguiente sección.

6.2.5 Comparación de resultados

La figura 26 permite apreciar las diferencias obtenidas con cada modelo propuesto. Nombraremos al conjunto de 6 y 4 emociones seleccionadas bajo los criterios presentados en la sección 3.2.2 como 6EM_1 y 4EM_1 respectivamente, mientras que los conjuntos de 6 y 4 emociones seleccionados en base a tasa de verdaderos positivos serán 6EM_2 y 4EM_2 respectivamente.



	6 EM_1	6 EM_2	4EM_1	4EM_2
SVM LINEAR	67	75	72	80
SVM POLY	70	77	75	82
KNN	67	75	70	79
RB	62	70	65	76
MLP	61	69	61	70

Figura 26 Comparación de resultados de precisión.

Se puede ver en la figura 26 que el clasificador que entrega los resultados más bajos en todos los escenarios es la red neuronal Multi – Layer Percetron así como el clasificador SVM con kernel Poly es el clasificador que ha entregado mejores resultados. Los clasificadores SVM entregan en general buenos resultados lo que corresponde con el estado del arte revisado.

En cuanto selección del conjunto de emociones tanto de 6 y 4, los resultados de la clasificación de las emociones con mejor tasa de verdaderos positivos como se esperaba obtuvieron mejores resultados llegando a alcanzar hasta un 82% de precisión con el clasificador SVM (Poly), clasificando 4 emociones (enojo, miedo neutral y sorpresa).

Los conjuntos de emociones que seleccionamos en la sección 3.2.2 alcanzaron resultados máximos de precisión de 70 % y 75% cuando se trabajó con 6 y 4 emociones respectivamente. Mientras que los máximos resultados de precisión obtenidos con los conjuntos de emociones con mejor TVP (sección 3.2.4) son 77% y 82% cuando se trabaja con 6 y 4 emociones respectivamente.

Conclusiones

- El reconocimiento de emociones por medio de voz es un problema difícil de resolver, ya que incluso el ser humano en ciertas ocasiones no puede distinguir fácilmente las emociones de otro, especialmente porque las expresiones de estas emociones suelen variar de persona a persona. Por esta razón tampoco se puede esperar una clasificación cien por ciento correcta de una máquina, pero sin duda una eficiente detección de emociones por medio de voz aporta una mejora importante al HCI de manera que las personas puedan tener una interacción más natural con las máquinas.

- En términos de la acústica, las técnicas de procesamiento del habla ofrecen información paralingüística valiosa derivada principalmente de características prosódicas (acentos, tonos y entonación) y espectrales (frecuencia y energía) de los audios.

- Trabajar con el extractor de características Mfcc ha permitido obtener un número considerable de características por audio. Se pudo además ver en los resultados que trabajar con 13, 20 o 27 vectores acústicos no modificaban drásticamente los resultados de precisión de la clasificación.

- Podemos nombrar las ventajas de la base de datos RAVDESS, esta contiene las 6 emociones básicas mencionadas por los psicólogos Ekman y Plutchik, contiene grabaciones de audio de 12 hombre y 12 mujeres (es importante tener datos de ambos sexos debido a las diferencias que existen en ambos por la morfología laríngea) y es una base de libre acceso con licencia Creative Commons.

- La base de datos RAVDESS que se ha seleccionado para el desarrollo del presente trabajo contiene las 6 emociones básicas que proponen tanto Ekman como Pultchik. Se han analizado 4 tipos de conjuntos de datos tanto de 6 y 4 emociones. En los conjuntos de emociones seleccionado bajo los criterios presentados en la sección 3.2.2 que se apegan a la apreciación huma, el conjunto de 6 emociones posee 5 de

las 6 emociones básicas y el conjunto de 4 emociones posee 4 de las 6 emociones básicas. Estos han obtenido resultados de precisión de 70% con 6 emociones y 75% con 4 emociones. Mientras que los conjuntos de emociones seleccionados por obtener mejor tasa de verdaderos positivos, el conjunto de 6 emociones posee 4 emociones básicas y el conjunto de 4 emociones posee 3 emociones básicas y han obtenidos resultados de precisión de 77% y 82% respectivamente.

- En base a los resultados obtenidos se pudo observar que la emoción con menor tasa de verdaderos positivos (menor eficiencia en la clasificación de las muestras de la clase) es Felicidad. Pudiendo ser esta la causa por la cual existe mucha diferencia en la clasificación de las emociones seleccionadas y las que obtuvieron mejor tasa de verdaderos positivos ya que en esta última la emoción fue omitida.

- RAVDESS ha sido utilizada para la detección de emociones por medio voz en (Iqbal & Barua, 2019), donde se ha evaluado la precisión de detección de 4 emociones (Enojo, felicidad, tristeza y neutral) utilizando el clasificador SVM, extrayendo 34 características por audio. Cuando se evalúan audios de emociones grabadas solo por hombres se obtiene una precisión de 79%, la precisión con emociones grabadas solo por mujeres es de 79% y con la combinación de ambos 81 %.

  Los resultados de precisión obtenidos en (Iqbal & Barua, 2019) son muy parecidos a los resultados que hemos obtenido al testear nuestro modelo con las 4 emociones con mejor tasa de verdaderos positivos (precisión igual a 82%). La única emoción que tienen en común estos conjuntos de emociones es “neutral”, todas las demás no corresponden. Esta última puede ser la principal diferencia de resultados con el conjunto de emociones seleccionado bajo criterios ya que la emoción neutral es la emoción que mejor se clasifica y no ha sido considerada en este último conjunto.

  También se puede destacar que se han obtenido resultados parecidos pese a que en el presente trabajo se ha utilizado un único extractor de características, mientras que en el estudio previo se utiliza 34 extractores.

- Un reto importante en la aplicación de detección de emociones por medio de voz es conseguir audios limpios ya que en general los entornos tienden a ser ruidosos. Además, para el entrenamiento de los clasificadores tendría gran importancia obtener muestras de emociones reales ya que en la mayoría de las bases de datos de emociones estas suelen ser grabadas por actores que no experimentan las emociones de forma real.

Líneas Futuras

Para trabajos futuros en la detección de emociones se puede proponer lo siguiente:

- Combinar extractores de características para generar un conjunto robusto de datos de entrada para el entrenamiento de los clasificadores.

- Combinar los datos de entrada de voz con datos adicionales como imágenes captadas por cámara, señales electro-encefalográficas captadas por equipos como Emotiv o frecuencia cardiaca captada por pulsómetros. Se debe aprovechar el uso de estas herramientas ya que todos estos aspectos varían en el humano cuando experimenta diferentes emociones indistintamente de su cultura e idioma.

A. Anexos

A.1 Tabla de resultados del estado del arte

Clasificador	Resultados	Feature Speech / Características de voz	Base de datos	Referencia
SVM	~ 90%	LLDs	Berlin EMO	(Vlasenko, Schuller, Wendemut, & Rigoll, 2007)
SVM	87.5%	MFCC	Berlin EMO	(Schuller, Müller, Lang, & Rigoll, 2005)
SVM	89%	MFCC	Berlin EMO DSPLAB	(Yang,, Ji, & Liu, 2009)
GMM	90%	MFCC	ISL Meeting Corpus	(Neiberg, Elenius, & Laskowski, 2006)
GMM	81%	Sequential Forward Floating Selection (SFFS)	Berlin EMO	(Atassi & Esposito, 2003)
GMM	74.6%	SFFS	Berlin EMO	(Lugger & Yang, 2007)
HMM	89%	WTM (Winner- Take-Most)	Berlin EMO	(Yun & Yoo, 2009)
HMM	~87%	LDA (Linear Discriminate Analysis)	5 emociones	(Yu, 2008)

Clasificador	Resultados	Feature Speech / Características de voz	Base de datos	Referencia
HMM	86%		7 emociones	(Schuller, Rigoll, & Lang, Hidden Markov model-based speech emotion recognition, 2003)
ANN	83.2%	MFCC	Berlin EMO	(Iliou & Anagnostopoulos, 2009)
RF	80.6%	MFCC	3 emociones	(Rong, Chen, Chowdhury, & Li, 2007)
GMM	79%	SBC (Sub band based cepstral)	6 emociones	(Kishore & Satish, 2013)
SVM	95.1%	MFCC, MEDC (energy spectrum dynamic coefficients), Energy	Berlin EMO	(Pan, Shen, & Shen, 2012)
GMM	95.83%	MFCC	4 emociones voz femenina	(Rao, et al., 2012)
CNN	~93.7%	Spectrogram fragments	Berlin EMO	(Huang, Dongz, Mao, & Zhan, 2014)
CNN	88%	STFT	7 emociones	(Lim, Jang, & Lee, 2016)
SVM	97.7%	MFCC	Berlin EMO	(A, Roy, & Selvi, 2013)
SVM	98%	MFCC, Energy	Thai DB	(Seehapoch & Wongthanavasu, 2013)

Clasificador	Resultados	Feature Speech / Características de voz	Base de datos	Referencia
CNN	84.3%	Spectrogram fragments	Berlin EMO	(Badshah, Ahmad, Rahim, & Baik, 2017)
RBF NN	75.8%	MFCC, BDPCA+LDA	eNTERFACE'05	(Ooi, Seng, Ang, & Chew, 2014)
RBF NN (Radial Basic Function)	79.5%	Periodicity Histogram	Danish DB	(Chavan & Gohokar, 2012)
GMM	74.45%	MFCC y ACFC (Auto Correlation Function Coefficient)	Berlin EMO	(Ning An, Wang, Ren, & Li, 2013)
SVM	80.6%	ZCR (zero-crossing rate), RMS, HNR (harmonic-to-noise), PITCH, MFCC	Berlin EMO	(Garg, Kumar, & Sinha, 2013)

Tabla 8 Revisión del estado del arte de los últimos 18 años en detección de emociones por medio de voz

Referencias

A, M., Roy, S., & Selvi, T. (2013). SVM Scheme for Speech Emotion Recognition using. *International Journal of Computer Applications MFCC Feature*.

Anagnostopoulos, C., & Vovoli, E. (2009). Sound Processing Features for Speaker-Dependent and Phrase-Independent Emotion Recognition in Berlin Database. *Information Systems Development*, 413 - 421.

Anagnostopoulos, C.-N., Iliou, T., & Giannoukos, I. (2012). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Springer Science+Business Media Dordrecht* , 155-177.

Atassi, H., & Esposito, A. (2003). A Speaker Independent Approach to the Classification of Emotional Vocal Expressions. *20th IEEE International Conference on Tools with Artificial Intelligence*.

Babaee, E., Badrul Anuar, N., Abdul Wahab, A. W., Shamshirband, S., & T. Chronopoulos, A. (2015). An Overview of Audio Event Detection Methods from Feature Extraction to Classification. 1087-6545.

Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. (2017). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. *IEEE*.

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

Chavan, V. M., & Gohokar, V. (2012). Speech Emotion Recognition by using SVM-Classifier. *International Journal of Engineering and Advanced Technology (IJEAT)*.

Chavan, V., & Gohokar, V. (2012). Speech Emotion Recognition by using SVM-Classifier. *International Journal of Engineering and Advanced Technology (IJEAT)*, 2249-8958.

Chen, L., Mao, X., Xue, Y., & Cheng, L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing*, 1154 - 1160.

Dave, N. (2013). Feature Extraction Methods LPC, PLP and MFCC. *INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN*.

Donaldson, M. (27 de 04 de 2017). *sixseconds The emotional Intelligence network*. Obtenido de Plutchik's Wheel of Emotions – 2017 Update: https://www.6seconds.org/2017/04/27/plutchiks-model-of-emotions

Forsberg, M. (2003). Why speech recognition difficult. *Researchgate*, 1 - 10.

Garg, V., Kumar, H., & Sinha, R. (2013). Speech Based Emotion Recognition Based on Hierarchical Decision Tree with SVM, BLG and SVR Classifiers. *IEEE*.

Huang, Z., Dongz, M., Mao, Q., & Zhan, Y. (2014). Speech Emotion Recognition Using CNN. *ACM*.

Iliou, T., & Anagnostopoulos, C. (2009). Comparison Of Different Classifiers for Emotion Recognition. *Proceedings of panhellenic conference in informatics*, 102-106.

Infographics. (2015). *Infographics*. Obtenido de Robert Plutchik's Wheel of Emotions (like colors, 8 primary emotions can be mixed to describe all human emotions): https://www.reddit.com/r/Infographics/comments/380zah/robert_plutchiks_wheel_of_emotions_like_colors_8/

Iqbal, A., & Barua, K. (2019). A Real-time Emotion Recognition from Speech using Gradient Boosting. *International Conference on Electrical, Computer and Communication Engineering (ECCE)*.

Kishore, K., & Satish, K. (2013). Emotion Recognition in Speech Using MFCC and Wavelet Features. *IEEE International Advance Computing Conference*.

Lim, W., Jang, D., & Lee, T. (2016). Speech Emotion Recognition using Convolutional and Recurrent Neural Networks. *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*.

Livingstone, S., & Russo, F. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English.

Livingstone, Steven, R., & Frank, A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) . *PLoS ONE. Zenodo*.

Lugger, M., & Yang, B. (2007). An incremental analysis of different feature groups in speaker independent emotion. *Proceedings of international congress phonetic sciences*, 2149-2152.

Mitrović, D., Zeppelzauer, M., & Breiteneder, C. (2010). Advances in Computers. *Features for Content-Based Audio Retrieval*, 71-150.

Navas, Hernáez, & Luengo. (2006). An objective and subjective study of the role of semantics and prosodic. *IEEE Trans Audio Speech Lang Process*, 1117 - 1127.

Nave, ©. (2016). *Vocal Sound Production*. Obtenido de HyperPhysics University, Georgia State: http://hyperphysics.phy-astr.gsu.edu/hbase/Music/voice.html

Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion Recognition in Spontaneous Speech Using GMMs. *NTERSPEECH*, 809-812.

Ning An, Q. Z., Wang, K., Ren, F., & Li, L. (2013). Speech Emotion Recognition using Combination of Features. *Fourth International Conference on Intelligent Control and Information Processing (ICICIP)*.

Ooi, C. S., Seng, K. P., Ang, L. M., & Chew, L. W. (2014). A new approach of audio emotion recognition. *Elsevier*.

Pal Singh, P., & Rani, P. (s.f.). An Approach to Extract Feature using MFCC. *International organization of Scientific Research*, 21-25.

Pan, Y., Shen, P., & Shen, L. (2012). Speech Emotion Recognition Using Support Vector Machine. *International Journal of Smart Home*.

Python Software Foundation. (2019). *Python*. Obtenido de SpeechRecognition 3.8.1: https://pypi.org/project/SpeechRecognition/

Rao, K. S., Kumar, T. P., Anusha, K., Leela, B., Bhavana, I., & Gowtham, S. (2012). Emotion Recognition from Speech. *International Journal of Computer Science and Information Technologies*.

Rong, J., Chen, Y., Chowdhury, M., & Li, G. (2007). Acoustic Features Extraction for Emotion Recognition. *IEEE International Conference on Computer and Information Science, nº 6th*, 419-424.

Rong, J., Li, G., & Phoebe Chen, Y.-P. (2008). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing and Management*, 315 - 328.

Schuller, B., Müller, R., Lang, M., & Rigoll, G. (2005). Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. *INTERSPEECH,*.

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model-based speech emotion recognition. *Proceedings of international conference on multimedia and expo*, 401-404.

Schuller, Villar, Rigoll, & Lang. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. *In: Proceedings of 9th Eurospeech–Interspeech*, 805 - 809.

Seehapoch, T., & Wongthanavasu, S. (2013). Speech Emotion Recognition Using Support Vector Machines. *5th International Conference on Knowledge and Smart Technology (KST)*.

Seehapoch, T., & Wongthanavasu, S. (2013). Speech Emotion Recognition Using Support Vector Machines. *International Conference on Knowledge and Smart Technology*.

Shing Ooi, C., Phooi Seng, K., Ang, L.-M., & Chew, L. (2014). A new approach of audio emotion recognition. *ELSEVIER - Expert Systems with Applications*, 12.

Shmyrev, N. (8 de 01 de 2016). *stackoverflow*. Obtenido de How to split speech data on frames and compute MFCC: https://stackoverflow.com/questions/34672182/how-to-split-speech-data-on-frames-and-compute-mfcc

Signal Processing. (11 de 06 de 2019). Obtenido de Why is each window/frame overlapping?: https://dsp.stackexchange.com/questions/36509/why-is-each-window-frame-overlapping

Tan, L., & Jiang, J. (2013). *Digital Signal Processing - Fundamentals and Applications.* ScienceDirect.

Vlasenko, B., Schuller, B., Wendemut, A., & Rigoll, G. (2007). Turn-level:emotion recognition from speech considering static and dynamic processing. *Proceedings 2nd international conference on affective computing and intelligent interaction*, 139-147.

Wang, Du, & Zhan. (2008). Adaptive and optimal classification of speech emotion recognition. *Proceedings*, 407 - 411.

Wikipedia. (10 de 06 de 2019). *Escala Mel*. Obtenido de CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=617721

Yang,, C., Ji, L., & Liu, G. (2009). Study to Speech Emotion Recognition Based on TWINsSVM. *Fifth International Conference on Natural Computation*.

YANKAYIŞ, M. (10 de 06 de 2019). *FEATURE EXTRACTION MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)*. Obtenido de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.701.6802&rep=rep1&type=pdf

Yu, W. (2008). Research and implementation of emotional feature classification and recognition in speech signal. *Proceedings of international symposium on intelligent information technology application*, 471-474.

Yun, S., & Yoo, C. (2009). Speech emotion recognition via amax-margin framework incorporating a loss function. *Proceedings IEEE international conference on acoustics, speech and signal processing*, 4169-4172.