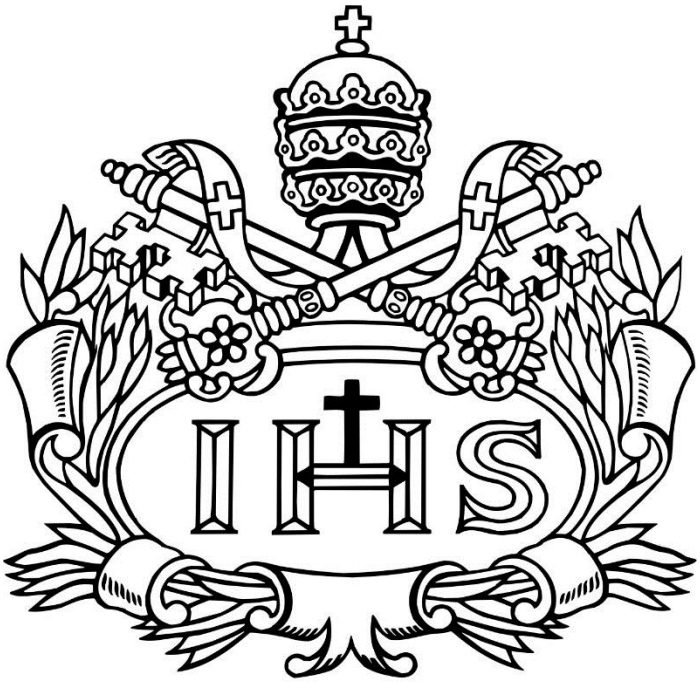
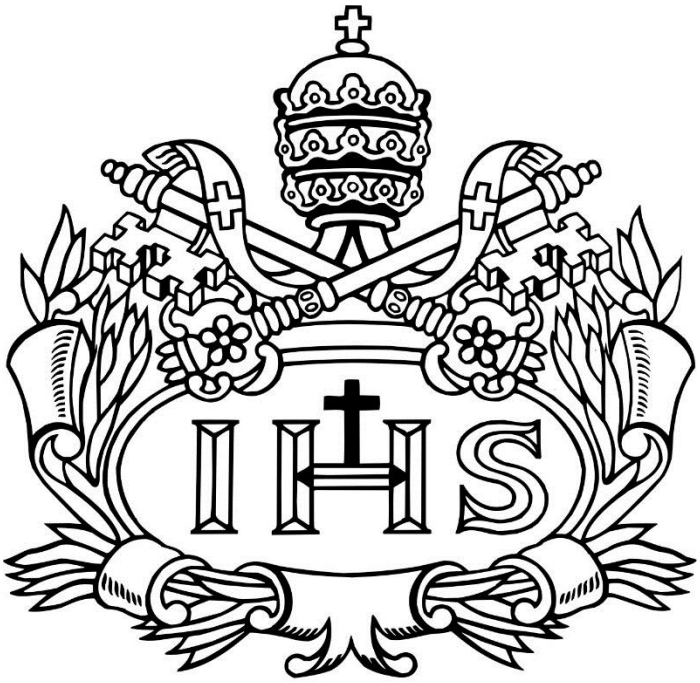
DETECCIÓN DE ESTADOS DE ÁNIMO MEDIANTE EL PROCESAMIENTO DE SEÑALES ACÚSTICAS



PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
CARRERA DE INGENIERÍA ELECTRÓNICA
BOGOTÁ D.C.
2017

DETECCIÓN DE ESTADOS DE ÁNIMO MEDIANTE EL PROCESAMIENTO DE SEÑALES ACÚSTICAS



Pontificia Universidad
JAVERIANA
Bogotá

Autor

Esteban Orozco Castaño

Director

Ing. Jairo Alberto Hurtado Londoño, PhD.

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
CARRERA DE INGENIERÍA ELECTRÓNICA
BOGOTÁ D.C.
2017

AGRADECIMIENTOS

Quiero agradecer a mis padres y a mi hermano.

Su apoyo incondicional es un motivador indiscutible.

- *Esteban Orozco C.*

Tabla de Contenido

Tabla de Figuras

Lista de Tablas

1 INTRODUCCIÓN

El reconocimiento de emociones en el habla (*speech emotion recognition*) es un campo de investigación relativamente nuevo inducido por el deseo de investigadores de lograr una interacción automática y "natural" vía máquina-humano. Este deseo ha provocado una enorme investigación en el área de reconocimiento del habla donde se han logrado avances considerables; detección de voz, detección de locutor y conversión de discurso a palabras son algunos ejemplos de estos avances. Sin embargo, el reconocimiento de emociones sigue siendo un reto por varias razones: no existe un consenso general entre psicólogos o investigadores acerca de las emociones, no es claro qué características son las más adecuadas para la detección de emociones, y finalmente, las diferencias culturales, de lenguaje, de estilos de habla y acentos son factores relevantes a la hora de considerar el tratamiento de las señales [1]. Reconocer emociones en el habla, especialmente emociones como furia, es de especial interés en aplicaciones puntuales como es el caso de los *call center*.

Si bien las emociones no están bien definidas, son innatas al hombre. Las características típicas de las reacciones faciales generadas por cada emoción son una muestra de este hecho; por otra parte, las personas distinguen las emociones cuando las sienten, por tanto, es claro que existen. Ahora, dado que éstas influyen en la conducta del hombre [1], son un área de interés para los investigadores. Dicho esto, es coherente que exista entre los *call center* (en especial aquellos de recuperación de cartera) un especial interés por tales investigaciones. Es importante resaltar, la ira o furia son la emoción más indeseada en el ejercicio de las actividades propias de los *call center* entre la definición de emociones básicas hecha por Ekman (Ira, Asco, Miedo, Alegría, Tristeza, Sorpresa) expuestas en [2]. Bajo el supuesto de que en algún *call center* la ira se presenta con una mayor probabilidad sobre las demás emociones de alta excitación, este trabajo justifica su desarrollo en la detección de esta, o de su ausencia.

El objetivo de este trabajo de grado se basa en el desarrollo del prototipo funcional de un sistema de detección automática de alteración de emociones mediante el procesamiento de señales acústicas generadas durante interacciones humanas. Para esto, se realizó la consolidación de una base de datos de locuciones en español, con su respectiva etiqueta (alta o baja excitación emocional, alternativamente, presencia o ausencia de ira) y compuesta de características usadas en reconocimiento de emociones por habla (como características continuas del habla), estas son, *pitch* y Coeficientes Cepstrales en las Frecuencias de Mel o MFCC, por sus siglas en inglés, se probaron sistemas de clasificación como SVM (del inglés *Support Vector Machine*) y KNN (*k n-nearest neighbors*) para esta base de datos, y finalmente se evaluó el desempeño del sistema propuesto con los indicadores de errores de entrenamiento y de prueba.

2 MARCO TEÓRICO

Existen variantes en el objetivo de detectar emociones en el habla, esencialmente esto se puede lograr mediante la selección de características y clasificadores adecuados. En este proceso, el mayor reto se encuentra en la selección de las características correctas, pues bajo el imaginario de un sistema compuesto de características pobres, inadecuadas o corruptas se concluirá que no será suficientemente apto la realización de su labor, mientras que, de no tener un clasificador adecuado se podrá explorar más fácilmente diferentes opciones de clasificación. Es así como, a continuación, se presenta el marco teórico suficiente para comprender la totalidad de este informe, donde se incluye conceptos teóricos clave acerca de las características y clasificadores utilizados, los modelos psicológicos de las emociones, así como también una breve descripción del aparato fonador humano y las respuestas fisiológicas humanas ante estímulos emocionales.

2.1 Modelos Psicológicos de las Emociones

Se ha establecido que el habla es un evento acústico que contiene información importante sobre el funcionamiento del sistema nervioso central, y por lo tanto contiene información sobre el estado emocional de un individuo [3].

Una de las discusiones actuales más controvertidas acerca de las emociones consiste en la existencia, o no, de emociones básicas de las cuales se derivan el resto de emociones, con esto en mente, Ekman, previamente mencionado, define 6 emociones básicas ampliamente aceptadas entre psicólogos, aunque existe una variedad de otras concepciones acerca de características básicas, como también existe un modelo diferente para describir las emociones, el modelo continuo de las emociones. Para una aproximación más profunda sobre diferentes propuestas se invita al lector consultar en [3], [4], donde podrá encontrar tablas comparativas entre modelos de emociones básicas propuestas por diferentes autores, un modelo continuo que utiliza las características activación, valencia y dominancia como características, y finalmente, las características principales de las emociones básicas definidas por Ekman.

Existe una creencia de que las emociones pueden ser caracterizadas por tres dimensiones: activación, valencia y dominancia. Activación, se refiere a la cantidad de energía requerida al expresar cierta emoción, también entendida como alta o baja excitación; la valencia describe qué tan negativa o positiva es una emoción específica, y la dominancia describe el grado de control del individuo sobre la situación que causa la emoción [2].

Es importante comentar la naturaleza relativa de etiquetado de emociones, en vista de la falta de consenso sobre la definición de las mismas; también es importante decir que la diferencia entre estas no podría establecerse propiamente en términos absolutamente cuantitativos, sino que serían cualitativamente diferentes [4], sin embargo, un modelo matemático es realizado con el objeto de caracterizar cada una de estas y de esta manera poder hacer uso de herramientas que proporciona el procesamiento de señales para identificar estos cambios emocionales de interés.

2.1.1 Modelos Discretos de las Emociones

Estos modelos se basan en el modelo de emociones básicas, a partir de las cuales las demás emociones pueden describirse, es decir, son una combinación de estas. Como se ha comentado anteriormente, no existe un criterio único para definir qué emociones forman este conjunto. Los modelos

discretos permiten una representación más particularizada de las emociones en las aplicaciones donde solamente se requiere reconocer un conjunto predefinido de emociones [3]. La clasificación final para estos modelos será la detección de una única emoción.

2.1.2 Modelos Continuos de las Emociones

En estos modelos los estados emocionales son representados usando un espacio multidimensional continuo. Las emociones son representadas por regiones en un espacio n-dimensional [3], es decir, la clasificación no es única y puede existir más de una emoción detectada al mismo tiempo. Un ejemplo común de esta representación hace uso de la activación y la valencia para crear un espacio bidimensional en donde se encuentran las emociones, otro ejemplo menos utilizado es el modelo tridimensional, donde en adición, también se hace uso de la dominancia para crear un espacio en tres dimensiones. En estos modelos se debe seleccionar regiones en donde las emociones se encuentran más intensamente individualmente y la posición del patrón a evaluar en este espacio bidimensional o tridimensional definirá de alguna forma la intensidad de emociones presentes en la clasificación.

2.2 Respuestas Fisiológicas Humanas ante Estímulos Emocionales

De acuerdo a algunos estudios psicológicos realizados por Williams y Stevens [5], se encontró que el sistema nervioso simpático se excita frente a estímulos emocionales de alta excitación como la ira, produciendo incrementos en la frecuencia cardiaca, presión arterial, incrementos en la presión subglótica y cambios en la profundidad de los movimientos respiratorios. En consecuencia, el habla se torna más fuerte, más rápida, y es enunciada con mayor energía de alta frecuencia, tiene *pitch* promedio más alto, y rango de *pitch* más amplio [1].

2.3 Aparato Fonador Humano

El aparato fonador humano es el conjunto de órganos encargados de generar el sonido que se produce al hablar. El aire impulsado por el diafragma, proveniente de los pulmones, pasa por la laringe en donde se encuentran los pliegues o cuerdas vocales (ver Ilustración 1 *Aparato Fonador Humano. Tomado de [6]*), responsables de la modulación de este flujo de aire, transformándolo en la señal sonora de frecuencia fundamental f_0 *pitch*; posteriormente, este flujo de aire sube por la faringe y finalmente pasa a través del tracto nasal y bucal (únicos en cada persona), que actúan como resonadores acústicos o filtros de la señal [6], [7].

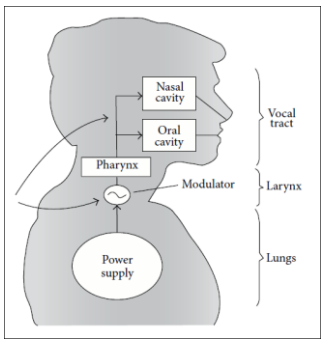


Ilustración 1 Aparato Fonador Humano. Tomado de [6].

Un modelo matemático que representa lo anteriormente dicho viene dado por las siguientes ecuaciones:

$$s(n) = e(n) * h(n)$$

$$S(k) = E(k)H(k)$$

donde,

$s(n)$ es la señal de voz (convolución entre e(n) y $h(n)$),

$e(n)$ es la señal de excitación proveniente de los pliegues vocales (de frecuencia f_0),

$h(n)$ es la función respuesta impulso que emula el efecto de los tractos que actúan como resonadores,

$S(k)$ es la transformada discreta de Fourier de la señal de voz,

$E(k)$ es la transformada discreta de Fourier de la señal de excitación, y

$H(k)$ es la función filtro que hecha sobre la señal de excitación.

2.4 Características de Clasificación

Las características, referidas a un sistema de Inteligencia Artificial, son una representación matemática de los patrones a clasificar, estas compactan la información presente en cada uno de estos y son el insumo de un clasificador, tanto para su entrenamiento como para su ejercicio. En un contexto de detección de emociones en señales de voz no existe un consenso acerca de qué o cuáles grupos de características son las mejores, entendidas como aquellas que mejor representan la información emocional en la voz; existen muchos tipos de características usadas en esta tarea, a continuación, se resume los conceptos teóricos de aquellas que fueron utilizadas para este trabajo de grado.

2.4.1 Pitch

El *pitch* es una característica de la voz humana y es considerado por muchos como la frecuencia fundamental f_0 de la señal producida por el flujo de aire a través de la glotis ubicada en la laringe en el aparato fonador humano. En la Ilustración 2 *Pulso glótico, tomado de [6]*se muestra la típica forma de un pulso glótico, inicialmente los pliegues vocales están cerrados y la amplitud de la señal es cero; en seguida, estos se despliegan lentamente hasta su máximo — en donde el flujo de aire es máximo — y, en consecuencia la amplitud de la señal, finalmente, estos se cierran rápidamente, y terminan con un ciclo glotal, conocido también como periodo de *pitch*; el recíproco de este se conoce como frecuencia fundamental [6].

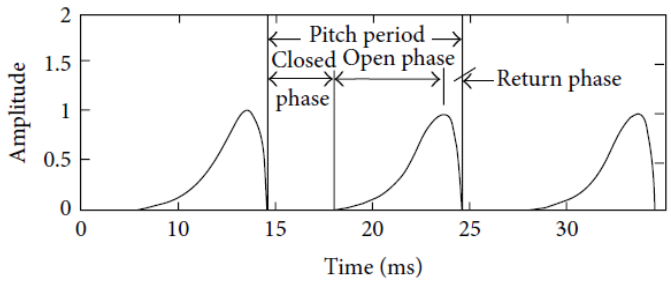


Ilustración 2 Pulso glótico, tomado de [6].

La estimación de esta característica se puede realizar usando diferentes métodos, entre ellos, el método de autocorrelación, el método cepstral, el método del producto armónico del espectro HPS [8], por sus siglas en inglés, entre otros.

A modo de información, el presente trabajo se realizó la estimación del *pitch* mediante los dos primeros métodos previamente enunciados y, una vez se encontró un resultado satisfactorio y eficaz para las condiciones de la base de datos, se seleccionó el método de estimación de autocorrelación; el pilar del fundamento matemático de esta estimación, es en esencia, desarrollado por Naotoshi Seo en [8].

2.4.2 Coeficientes Cepstrales en las Frecuencias del Mel

Los parámetros MFCC son un tipo particular de coeficientes cepstrales formulados por Davis y Mermelstein [9], ampliamente utilizados en la actualidad en varias aplicaciones, como por ejemplo, identificación de locutor, verificación de locutor o reconocimiento del discurso. Estudios demuestran que la percepción de volumen que se transduce en el oído interno humano depende de la frecuencia en una escala no lineal [10]. La motivación de los MFCC es emular este comportamiento del sistema auditivo humano, para esto, se hace uso de una nueva escala de frecuencia no lineal denominada MEL caracterizada por tener mayor resolución en bajas frecuencias de forma análoga a como sucede en el oído interno humano [7], [10].

A continuación, se presentan brevemente los conceptos teóricos necesarios para un análisis de estos coeficientes:

2.4.2.1 Ventaneo

Es la multiplicación de una señal en tiempo por varias ventanas (señales de energía finita y de soporte compacto) independientes; estas ventanas pueden, o no, encontrarse traslapadas entre sí y forman una colección de señales típicamente más pequeñas que la señal original; cada una de estas nuevas señales es llamada trama.

2.4.2.2 Pre-énfasis

El pre-énfasis es una técnica utilizada sobre las señales de audio con el objetivo de compensar la atenuación de aproximadamente 20 dB/década que se produce en el proceso fisiológico del mecanismo de producción del habla [7].

Sea $x(n)$ una señal de tiempo discreto arbitraria y sea $y(n) = x(n) - \alpha x(n-1)$, entonces $y(n)$ se define como la señal $x(n)$ expuesta a un filtro de pre-énfasis.

2.4.2.3 Banco de Filtros de Mel

El banco de filtros de Mel es una colección de ventanas triangulares de área unidad [7], típicamente las primeras 10 ventanas están linealmente espaciadas desde 0 Hz hasta 1 kHz, y en adelante, 10 logarítmicamente espaciadas según la escala de Mel desde 1 kHz hasta 5,5 kHz [11] (aunque otras implementaciones son hechas hasta 4 kHz [10]). La conversión de frecuencia en hertz a Mels viene dada por la ecuación $mel(f) = 2595 \log\left(1 + \frac{f}{700}\right)$

donde,

f es la frecuencia en hertz, y

mel es la frecuencia expresada en Mels.

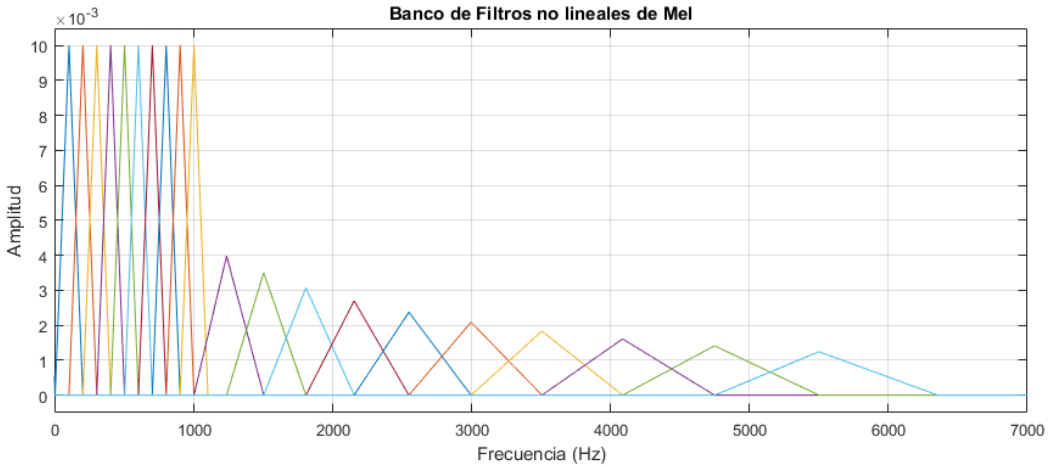


Ilustración 3 Típico banco de filtros en la escala de Mel.

En la Ilustración 3 se puede apreciar la forma de un típico banco de filtros utilizados en la extracción de los MFCC.

La extracción de los coeficientes MFCC estuvo basada en la teoría expuesta en [1], [7], [10], [12], que se detallará más adelante, en resumen, los pasos para su obtención son:

1. Ventaneo de la señal.
2. Pre-énfasis sobre cada ventana obtenida.
3. A cada ventana aplicarle la Transformada de Fourier Discreta.
4. Obtener el valor absoluto del resultado anterior.
5. Aplicar el Banco de Filtros de Mel sobre las señales espectrales anteriores.
6. Calcular el logaritmo de las energías de cada ventana obtenida al aplicar el banco de filtros.
7. Calcular la Transformada de Coseno Discreta sobre los logaritmos anteriormente calculados (otras implementaciones usan la Transformada Inversa de Fourier).

2.5 Clasificadores

Los clasificadores, en un contexto de Inteligencia Artificial, son algoritmos encargados de discriminar un patrón entre clases o etiquetas; para esto, no procesan propiamente el patrón, procesan sus características. Entre clasificadores, existen los de tipo aprendizaje supervisado; estos son algoritmos que previa operación, han sido entrenados, es decir, muestras de patrones con su respectiva etiqueta han sido procesados por este con el objetivo de ajustar sus reglas de decisión. Para este trabajo de grado fueron utilizados dos tipos de clasificadores supervisados, las máquinas de vectores de soporte o SVM y los algoritmos de los k vecinos más cercanos o KNN, a continuación se presenta un breve resumen conceptual acerca de estos.

2.5.1 Máquina de Vectores de Soporte SVM (Support Vector Machines)

Las máquinas de vectores de soporte son algoritmos de aprendizaje supervisado, estos basan su clasificación en el uso de un hiperplano discriminador. Los SVM, en consecuencia, son aptos únicamente en la clasificación de dos clases. La condición de clasificación de los SVM viene dada por la expresión

$W^T x + b$ [13]. W es el vector normal al hiperplano y su función es direccionarlo, b es una constante con la función de posicionar el hiperplano sobre el hiperespacio en una posición diferente al origen (para $b \neq 0$); W y b son seleccionados de manera que este hiperplano esté en la posición que minimiza algún error (función de costo), finalmente, x es el vector de características del patrón a clasificar. Así, las condiciones de clasificación son descritas por las siguientes expresiones:

$$\boldsymbol{W}^T\boldsymbol{X}_i + b > 0 \qquad para\ y_i = +1$$
$$\boldsymbol{W}^T\boldsymbol{X}_i + b < 0 \qquad para\ y_i = -1$$

También existen implementaciones de SVM no lineales mediante el uso del "truco del Kernel", para este caso, la clasificación se hace sobre hiperplanos ubicados en dimensiones generalmente superiores a las dimensiones del patrón, en donde se supone existirá una clasificación lineal [13], [14]. A continuación, en la Ilustración 4, se presenta gráficamente un ejemplo del hecho anteriormente expuesto.

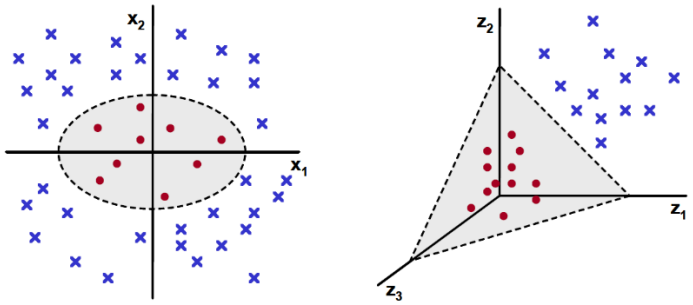


Ilustración 4 Muestra de transformación de dimensión.

Izquierda: patrones en 2D no separables por algún hiperplano, derecha: patrones en 3D separables por algún hiperplano.

2.5.2 K Vecinos más Cercanos KNN

El método de clasificación de k vecinos más cercanos es un método de clasificación supervisada no paramétrico, este método consiste en contar los k patrones más cercanos en el hiperespacio de características del patrón a clasificar, el resultado de la clasificación será igual a la etiqueta que más se repita en el conteo anteriormente descrito. Para estimar la cercanía, generalmente se usa la distancia euclidiana. En la Ilustración 5 se ejemplifica gráficamente este método.

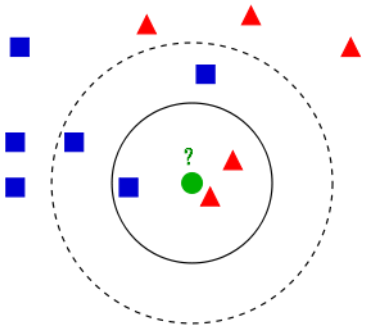


Ilustración 5 Ejemplo de decisión para k =3 en un knn.

3 OBJETIVO DEL PROYECTO

Los pilares de este trabajo de grado yacen en el cumplimiento del siguiente objetivo general, así como en el cumplimiento de los objetivos específicos que, integrados, dan cumplimiento al mismo.

3.1 Objetivo General

Desarrollar el prototipo funcional de un sistema de detección automática de alteración de emociones mediante el procesamiento de señales acústicas generadas durante interacciones humanas.

3.2 Objetivos Específicos

1. Consolidar una base de datos de locuciones en español con su respectiva etiqueta (alta o baja excitación emocional), que se componga de características usadas en reconocimiento de emociones por habla tales como características continuas del habla (*Continuous speech features*) que permita evaluar diferentes sistemas de reconocimiento de emociones.
2. Identificar entre los sistemas de clasificación SVM (del inglés *Support Vector Machine*) y ANN (*Artificial Neural Network*), cuál provee mejores características de reconocimiento con la base de datos disponible, basado en criterios convenidos con el supervisor.
3. Evaluar el desempeño del sistema con indicadores como errores de entrenamiento y de prueba para los audios en alta y baja calidad.

El desarrollo del prototipo definitivo consiste de un clasificador tipo knn, para $k = 4$, seleccionado una vez se encontró su desempeño y el desempeño de otros clasificadores, entre ellos los clasificadores tipo SVM. Se ha creado un método automático para transformar una base de datos de audios en una base de datos de características de estos, los errores de entrenamiento (14,74%) y de prueba han sido encontrados (21,27%) y serán analizados en este libro más adelante.

4 DESARROLLO

Las partes y etapas que conforman el desarrollo de este proyecto son Investigación, Pre-procesamiento, Extracción de Características, Entrenamiento y Evaluación, y finalmente, Consolidación del Sistema.

4.1 Investigación

La investigación hecha en este proyecto se basó en el entendimiento de los pilares del sistema propuesto; estos son, características de la base de datos y características a utilizar para la clasificación.

4.1.1 Base de Datos

Toda base de datos es un pilar en el desarrollo de un sistema de Inteligencia Artificial (AI), por lo que una investigación previa al desarrollo del sistema es pertinente; múltiples consideraciones fueron encontradas para su selección. Se encontró que dos de los factores críticos en una base de datos de audios de emociones es indudablemente la procedencia de los audios y la naturaleza de grabación de los mismos. Grabaciones de voz con contenido emocional actuado, inclusive hechas por profesionales, no garantizan la reacción natural del sistema fisiológico humano durante la experimentación de emociones, como cambios en características de la voz, que son, precisamente, las que este trabajo de grado pretende detectar. Cabe aclarar que grabaciones actuadas, que por lo general tienden a ser más exageradas que grabaciones de emociones reales [15], podrían ser de igual manera válidas, tema que no compete investigar en este trabajo. Otra pregunta que debe hacerse es sobre el control de las grabaciones, audios conseguidos en condiciones controladas son útiles en análisis científicos y experimentaciones; sin embargo, estos podrían reducir la validez de los datos [1]. Varias alternativas de bases de datos de audios con contenido emocional se pueden encontrar en internet, inconvenientemente, los idiomas de grabación de estos audios están en su gran mayoría en inglés, alemán y algunos pocos en mandarín.

En el desarrollo de este proyecto se decidió trabajar con grabaciones de auténticas reacciones emocionales interpretadas y etiquetadas por un profesional en función. Los interlocutores partícipes de estas grabaciones desconocían el propósito de estas; esta base de datos corresponde a registros de audios hechos por la empresa Visión Gerencial (*contact center* de Medellín, Colombia), estos fueron registrados en un típico espacio de trabajo de un *call center* de recuperación de cartera, donde existe ruido tanto acústico como eléctrico, interferencia proveniente de las llamadas hechas por los tele-operarios próximos al sitio de grabación, y demás señales acústicas indeseables producidas por el recinto o factores externos.

La base de datos de grabaciones de voz se compone de cuatro llamadas realizadas por una de las tele-operarias del *call center* ya mencionado. Una de estas grabaciones, con contenido emocional de ira por parte de ella, de un (1) minuto y cuarenta y dos (42) segundos de duración y las otras tres llamadas, etiquetadas con ausencia de contenido de ira por parte de ella; estas tienen duraciones de cuatro (4) minutos y cuarenta y siete (47) segundos, cinco (5) minutos y trece (13) segundos y tres (3) minutos y cuarenta (40) segundos (ver Tabla 1 *Base de datos proporcionada por Visión Gerencial*). Para un total de quince (15) minutos y veintidós (22) segundos de duración. Todas estas, a una frecuencia de muestreo F_s de 44100 Hz y almacenadas en formato WAV.

	Audio Furia	Audio sin Furia 1	Audio sin Furia 2	Audio sin Furia 3
Duración	1:42	4:47	5:13	3:40

Tabla 1 Base de datos proporcionada por Visión Gerencial.

Partiendo del hecho que la detección o verificación de locutores no es competencia de este trabajo de grado, se aclara que esta base de datos, anteriormente presentada, debe ser tratada de alguna manera para eliminar la voz de segundos locutores, para aplicaciones de verificación de locutor se sugiere al lector revisar la documentación hecha en [16].

En resumen, la base de datos recibida, propiedad del *contact center* previamente mencionado, se resume así: existe un único locutor, perteneciente a una voz femenina que corresponde a una de las tele-operarias de la empresa, 4 grabaciones (con frecuencia de muestreo de 44100 Hz) de voz de esta locutora fueron etiquetadas por un grupo de profesionales dentro de la empresa, la función principal de este grupo dentro de la empresa es monitorear y analizar llamadas aleatoriamente sobre los aproximados 100 puestos de trabajo de tele-operarios dentro de la empresa. De estas grabaciones, 3 fueron consideradas como mayoritariamente sin contenido de furia y 1 de ellas como mayoritariamente con contenido de furia.

A propósito se ha trabajado con una muestra muy superior de audios con ausencia de furia sobre audios con presencia de furia, esto, motivado por la intención de reflejar la frecuencia de ocurrencia en la vida real de estos eventos. Algunos investigadores prefieren trabajar con equilibradas cantidades de locuciones para diferentes etiquetas para evaluar con propiedad la precisión de clasificación [1].

4.1.2 Selección de Características para Entrenamiento y Prueba del Sistema

Mencionadas anteriormente las características seleccionadas son el *pitch* o frecuencia fundamental f_0 y los MFCC, motivados por la documentación expuesta en [1], [9], [17]–[19].

La mayoría de investigadores creen que características continuas del habla como el pitch contienen mucha información acerca del contenido emocional en una locución [1], las características derivadas de estas han sido fuertemente utilizadas en reconocimiento de emociones en el habla, entre ellas se encuentran, la media del pitch, la mediana, el máximo, el mínimo, el rango (máximo - mínimo), desviación estándar, entre otras [1].

Las características espectrales pueden ser extraídas en una variedad de coeficientes, sin embargo, para aprovechar mejor la distribución espectral sobre el espectro audible de frecuencia, usualmente, el espectro estimado es filtrado a través de un banco de filtros pasa-banda y, las características espectrales son posteriormente extraídas del resultado de este filtrado (tal como lo hacen los MFCC) [1], dado que la percepción humana no sigue una escala lineal.

4.2 Pre-procesamiento

El pre procesamiento hecho en las señales de audio consistió en dos etapas, la primera fue la eliminación de las voces de segundos interlocutores en la señal, con el objetivo de tener una señal con voces de una única persona, esto se hizo con el apoyo del editor de audio Audacity. El resultado de esta etapa es una nueva base de datos (ver Tabla 2 *Base de datos pre procesada – etapa 1*) de cinco (5) minutos y cuarenta y cuatro (44) segundos de duración. La segunda etapa consistió en seleccionar la detección de presencia de voz del interlocutor en cuestión, para esto se ha utilizado el concepto de energía de una ventana junto con un umbral fijo, en otras palabras, si la energía por muestra contenida en la ventana en cuestión es menor que cierto umbral, entonces la ventana no se considera con contenido de voz, la selección de este umbral se hizo con base en un análisis estadístico simple sobre la base de datos de audios para determinar el punto de trabajo. La función encargada de realizar el procedimiento de detección de voz puede encontrarse con el nombre detecta_voz.m, finalmente la función segmentar_audio.m se encarga de agrupar todas las ventanas conexas donde se detectó voz por la función anteriormente mencionada.

Existe una aproximación para la extracción de características, esta se basa en segmentar la señal de audio en pequeñas señales que contengan fonemas para posteriormente extraer sus características [20],

esta aproximación yace en el análisis de vocales pronunciadas por locutores en presencia de diferentes emociones, sin embargo, el pobre desempeño de los algoritmos de segmentación por fonemas puede ser un problema, una alternativa es extraer el vector de características por cada segmento de voz en lugar de cada fonema [1]. Cada segmento de voz se entiende como cada segmento continuo del habla que es causado por vibraciones de los pliegues (o cuerdas) vocales y que son oscilatorios [21].

Con esto en mente, se ha seleccionado un mínimo de 100 ms de ventana para considerar su tratamiento en etapas posteriores (ver Tabla 3), es decir, extracción de características locales y posterior entrenamiento y/o prueba. El total de este procedimiento se puede encontrar en el script de Matlab con nombre base_locuciones.m y una muestra gráfica de esto es presentada en la Ilustración 6. Otros trabajos han utilizado diferentes valores para la ventana mínima válida para tratamiento, como [19], en donde han usado 40 ms y consiguen altos grados de discriminación entre ira y neutralidad para la base de datos *Berlin emotional Database* [22].

El resultado de la etapa del pre procesamiento es 71 muestras (8.27% de la base de datos) etiquetadas con contenido de furia y 788 muestras etiquetadas sin contenido emocional de furia (91.73% de la base de datos). Nótese que, nuevamente la muestra de audios con ausencia de furia es ampliamente superior en cantidad con respecto a la muestra de audios con presencia de furia.

	Audio Furia	Audio sin Furia 1	Audio sin Furia 2	Audio sin Furia 3
Duración	0:27	1:35	2:35	1:07

Tabla 2 Base de datos pre procesada – etapa 1.

	Audio Furia	Audio sin Furia 1	Audio sin Furia 2	Audio sin Furia 3
Número de segmentos válidos	71	247	358	183

Tabla 3 Base de datos de segmentos válidos para procesar.

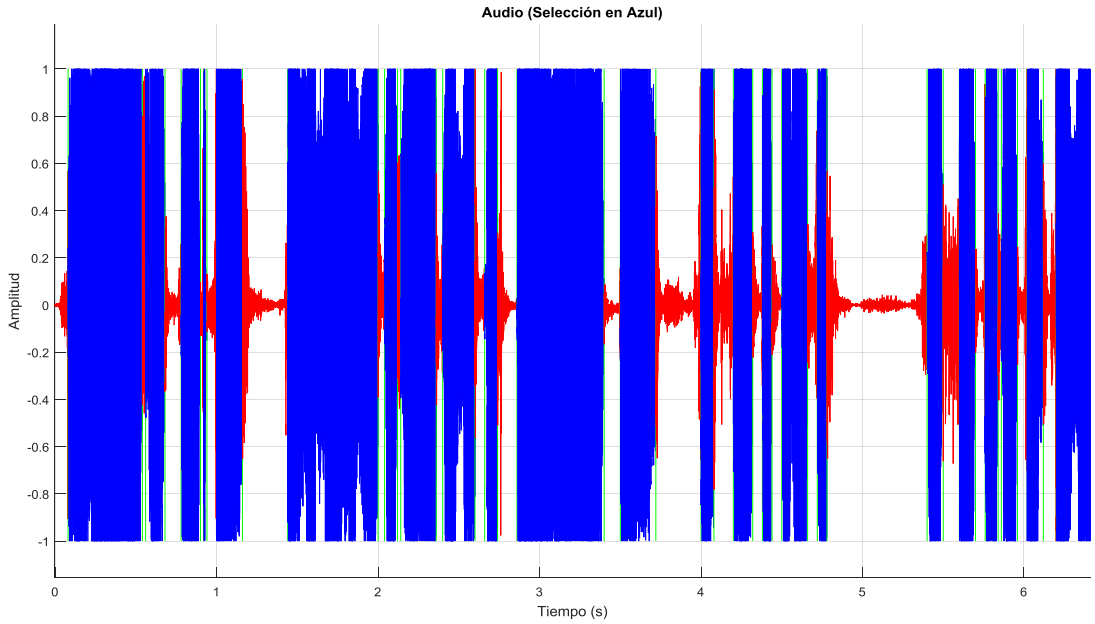


Ilustración 6 Ejemplo de pre procesamiento sobre señal discreta de audio.

4.3 Extracción de Características

Las características usadas utilizadas en el procesamiento del habla pueden ser agrupadas en 4, características continuas del habla (ejemplo, *pitch*), características de calidad del habla, características espectrales (ejemplo, MFCC) y características basadas en el operador de Teager TEO, cada una tiene pros y contras, sin embargo, en reconocimiento de emociones en el habla es común combinar características pertenecientes a diferentes categorías para representar la señal de voz [1].

A continuación, se presenta el procesamiento hecho en este proyecto para la extracción base de cada característica.

4.3.1 Pitch

4.3.1.1 Pitch vía método de Autocorrelación

Para estimar este parámetro se usó la función de Matlab xcorr que calcula la autocorrelación de una señal discreta en función del *lag* en muestras; de este resultado se consideró solo la mitad de la señal, pues esta es simétrica; a continuación, se identificó el punto máximo de autocorrelación de la señal en las inmediaciones de 2 ms a 20ms, inmediaciones seleccionadas, suponiendo que el *pitch* se encuentra en la región en frecuencia de 50 Hz a 500 Hz [23], finalmente el *pitch* es encontrado como la inversa del número de muestras desde que se comienza el análisis de autocorrelación hasta el primer punto de máxima autocorrelación multiplicado por la frecuencia de muestreo, esta es una la razón principal la función pitch_autoco encontrada en el archivo con este mismo nombre muestra el procesamiento realizado para encontrar el *pitch* en cada trama.

Entender la suposición que el *pitch* se encuentre en la región en frecuencia de 50 Hz hasta 500 Hz es bastante importante, indica que una ventana de 20 ms puede contener periodos completos de *pitch* tan bajos como 50 Hz, esta suposición es la base principal del criterio de selección de ventanas de 20 ms para el proceso de ventaneo en este sistema, explicado más adelante.

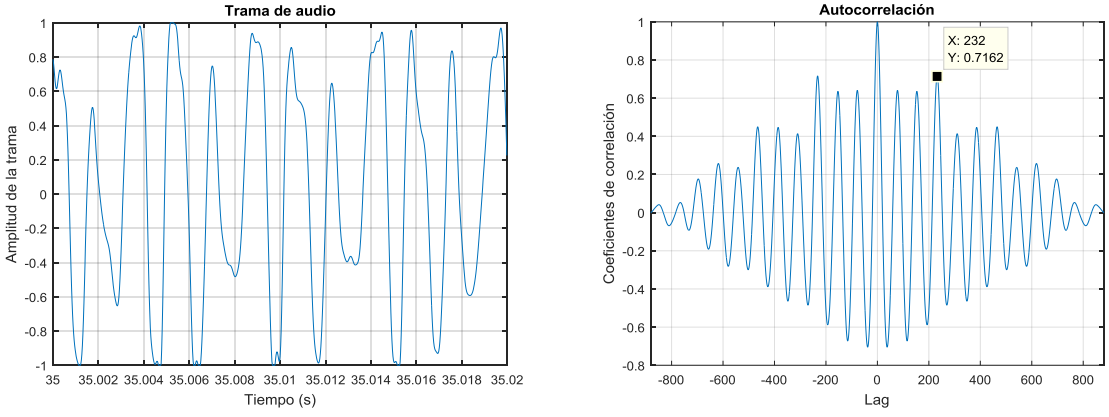


Ilustración 7 Comparación de una trama de audio contra su autocorrelación.

En este ejemplo (visible en la Ilustración 7) el *pitch* encontrado fue de 118,52 Hz proveniente de la voz de una mujer.

4.3.1.2 Pitch vía Método Cepstral

En este método se realizó un procedimiento similar al realizado en el método anteriormente descrito. La función pitch_cepstrum, encontrada en el archivo pitch_cepstrum.m de Matlab, implementa la estimación del *pitch* así: primero se realiza la transformada rápida de Fourier (FFT), seguidamente se encuentra el cepstro como la transformada inversa de Fourier del logaritmo del valor absoluto de la FFT, y finalmente, se busca sobre la región de interés (50 Hz a 500 Hz) el pico máximo; el inverso de este valor será la frecuencia fundamental. Este método fue finalmente descartado y se trabajó únicamente con el método anterior.

4.3.2 MFCC

A continuación, en la Ilustración 8, se presenta y posteriormente se explica el diagrama de bloques del proceso de extracción de los coeficientes MFCC.

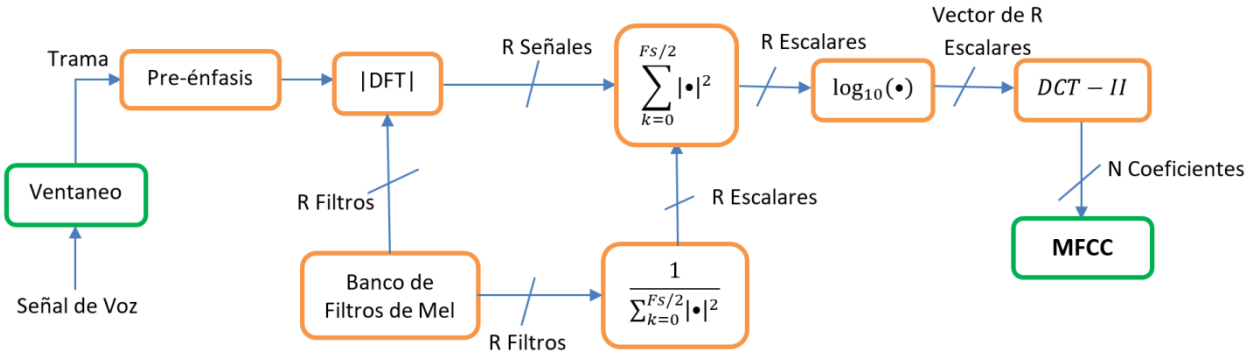


Ilustración 8 Diagrama de bloques para la extracción de los MFCC.

4.3.2.1 Ventaneo

Dado que las señales de voz no son estacionarias, inclusive en un sentido amplio, es común en aplicaciones de procesamiento del habla el dividir una señal de voz en pequeños segmentos llamados tramas, en donde cada trama se considera que es aproximadamente estacionaria [21]. A las características extraídas sobre estas tramas se les entiende por características locales [1].

Ventanas de 20 ms fueron utilizadas sobre cada uno de los segmentos pre-procesados (de mínimo 100 ms) para la extracción de características locales, el segundo motivo para la selección de este parámetro es que, tanto para la extracción de características relacionadas con el *pitch*, como para la extracción de características relacionadas con los MFCC, 20 ms son suficientes para obtener la información de interés acerca de los cambios en el *pitch* producto de la presencia, o no, de ira en el locutor, más adelante será más claro esta idea al aplicar la Transformada Discreta de Fourier.

En orden de realizar una transición más suave y recoger algo más de información es usual traslapar estas ventanas, adicionalmente, para reducir *ripples* en el espectro, producto del ventaneo (efecto de convolución en frecuencia), muy frecuentemente cada trama es multiplicada por una ventana tipo Hamming [1]. Es por esto que en este trabajo de grado estas ventanas han sido traslapadas al 50% y se han usado del tipo Hamming.

Función responsable de esta tarea encontrada en el archivo ventaneo_hamming.m.

Finalmente, en la Ilustración 9, se observa el ventaneo tanto rectangular (usado para calcular el *pitch*), como el ventaneo utilizando ventanas tipo Hamming (usado para calcular los MFCC).

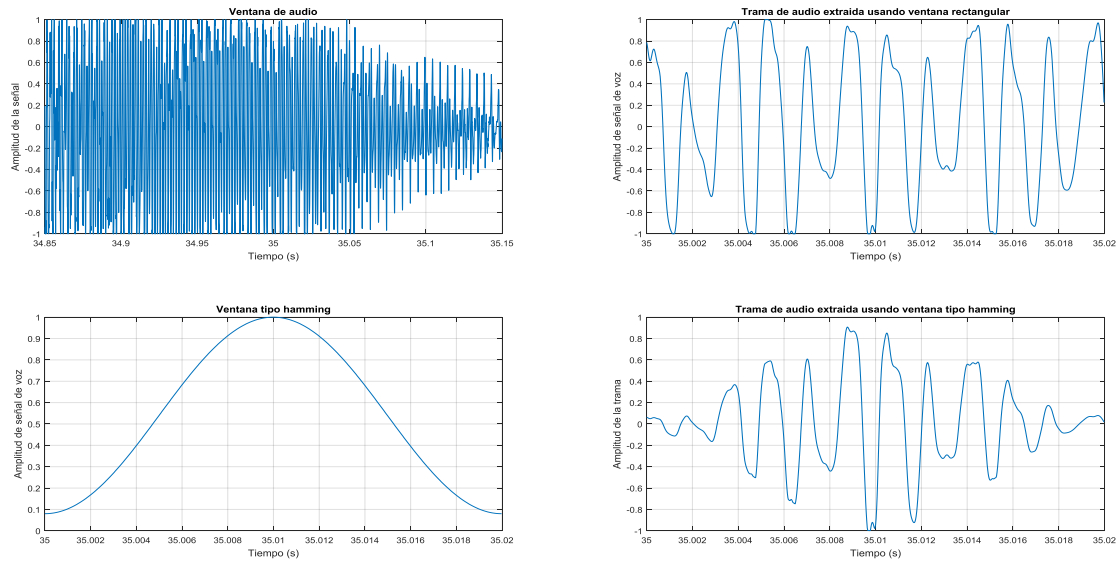


Ilustración 9 Ejemplificación de ventaneo utilizando ventana tipo Hamming.

4.3.2.2 Pre-énfasis

Para el filtro de pre-énfasis, con función de transferencia $H(z) = 1 - \alpha z^{-1}$, el valor de α recomendado para esta aplicación se encuentra en el intervalo [0,95 0,98] [7], [21], 0,95 ha sido el valor de trabajo, otras implementaciones utilizan 0,97, exclusivamente [1], [21].

Función responsable de esta tarea encontrada en el archivo pre_enfasis.m.

En la Ilustración 10 se encuentra gráficamente el resultado de la aplicación de este filtro sobre una trama de audio de 20 ms.

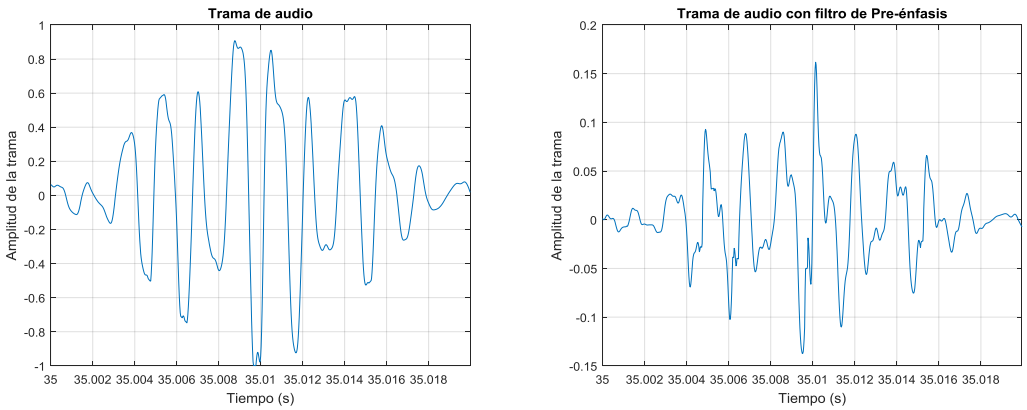


Ilustración 10 Ejemplificación de proceso de pre-énfasis.

4.3.2.3 DFT

El valor absoluto de la DFT se ha computado usando el comando abs() y fft() de Matlab para N puntos, N=L donde L es la longitud de cada ventana y el resultado de este procedimiento puede observarse en la Ilustración 11.

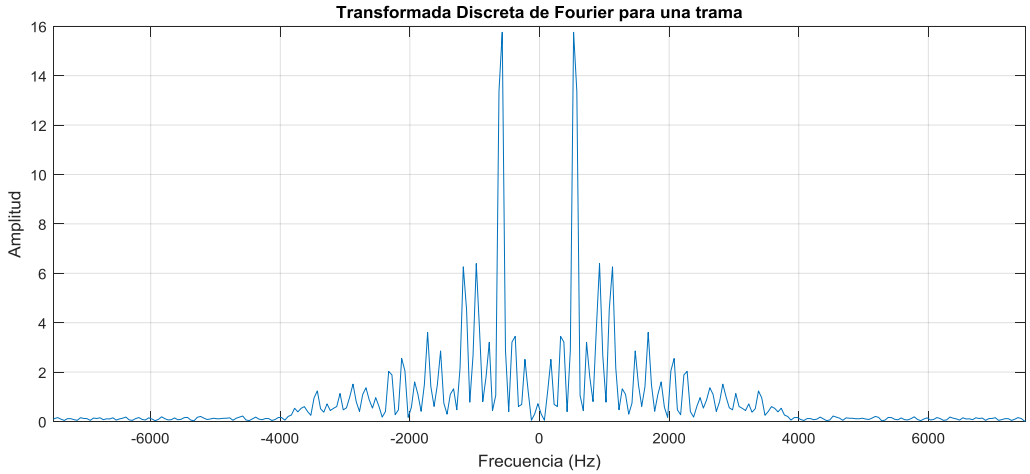


Ilustración 11 Respuesta en frecuencia para una trama de análisis.

4.3.2.4 Banco de Filtros de Mel

Se usaron 10 filtros linealmente espaciados desde 0 Hz hasta 1 kHz y 11 logarítmicamente espaciados según la escala de Mel (hasta aproximadamente 7.3 kHz), previamente comentado, dando prioridad a las frecuencias de 0 Hz a 1 kHz, como se muestra en la Ilustración 12.

Función responsable de esta tarea encontrada en el archivo banco_de_filtros.m

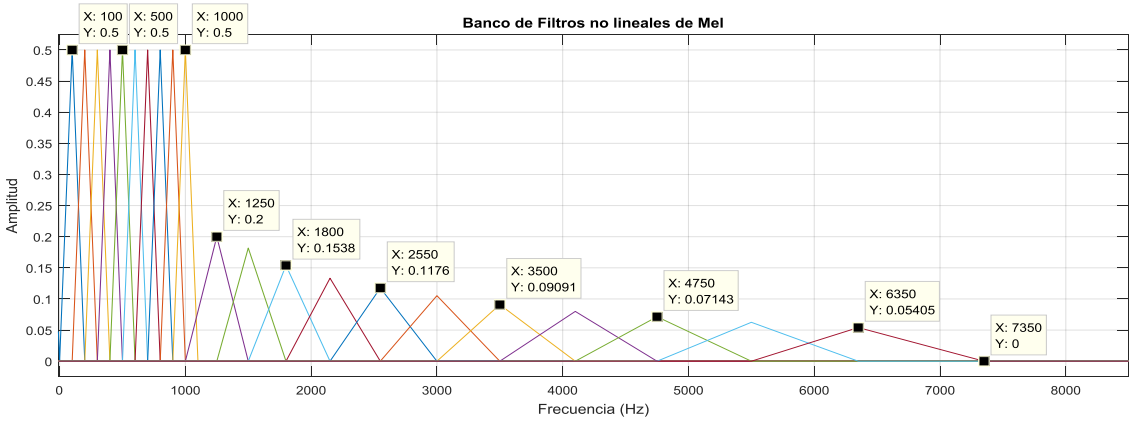


Ilustración 12 Banco de Filtros de Mel utilizado.

4.3.2.5 Filtrado

La señal en tratamiento se filtra con cada uno de los 21 filtros anteriormente presentados, conduciendo a 21 nuevas señales, cada una de estas 21 señales se multiplica por el inverso de la energía de cada filtro respectivo, posteriormente se calcula la energía de cada una de estas últimas 21 señales y, finalmente, se calcula el logaritmo de las anteriores 21 energías encontradas, concluyendo en un vector de 21 escalares.

Función responsable de esta tarea encontrada en el archivo filtrado_MFCC.m y log_energia.m.

En la Ilustración 13 se observa el resultado de la aplicación de algunos filtros sobre una señal espectral, utilizada en esa sección a modo de ejemplo.

En la Ilustración 14 puede comprobarse el correcto comportamiento del cepstrum de Mel, este debe ser una medida de la envolvente de la señal espectral [10].

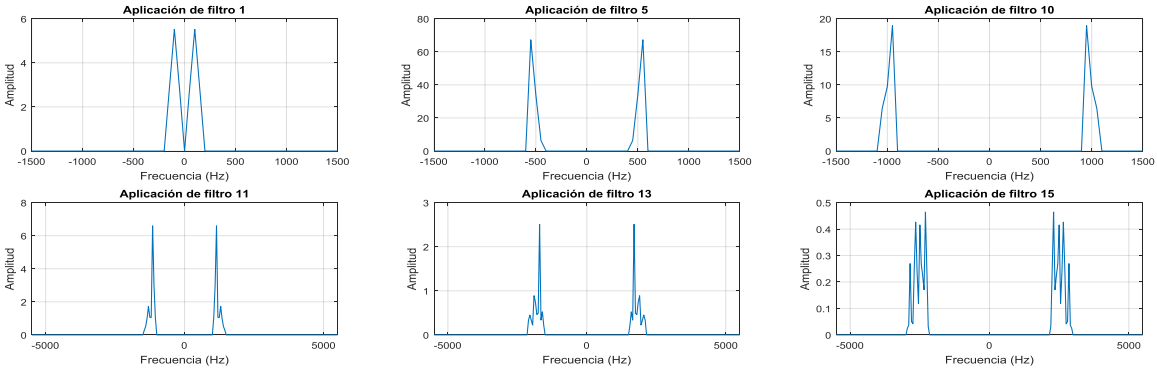


Ilustración 13 Algunos resultados del filtrado sobre la respuesta en frecuencia de la trama.

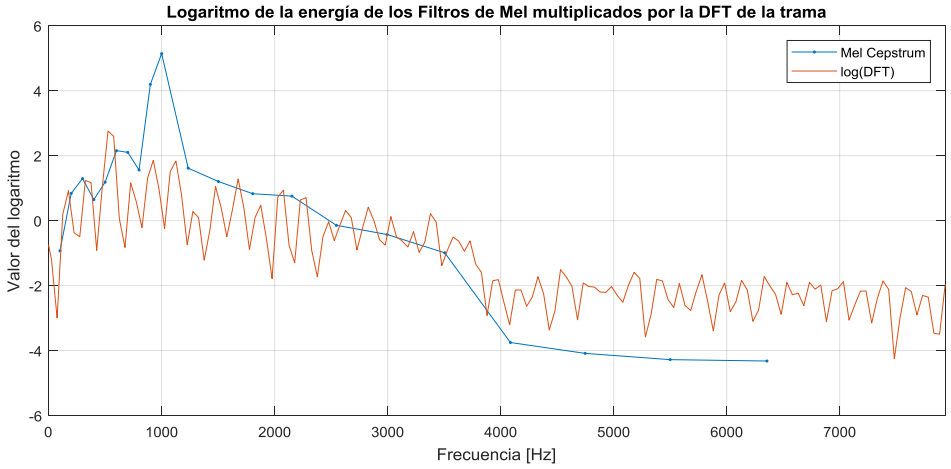


Ilustración 14 Comportamiento del cepstrum de Mel.

4.3.2.6 DCT-II

Finalmente, se encuentran 15 Coeficientes de Mel aplicando la transformada Discreta de Fourier al vector del paso anterior, este resultado se resume en la Tabla 4.

MFCC1	MFCC2	MFCC3	MFCC4	MFCC5	MFCC6	MFCC7	MFCC8	MFCC9	MFCC10	MFCC11	MFCC12	MFCC13
-0.0645	-0.0068	-0.0474	-0.0297	-0.0010	0.05017	0.01153	-0.0242	0.02185	0.06602	0	-0.06602	-0.02185

Tabla 4 Ejemplificación de coeficientes de Mel encontrados.

4.3.3 Características Derivadas

4.3.3.1 Características Derivadas del Pitch

Sobre cada ventana traslapada se ha calculado el *pitch* de la manera anteriormente descrita, método de autocorrelación, a partir de este parámetro se derivan características como lo son: *pitch* promedio, *pitch* máximo, *pitch* mínimo, *pitch* máximo menos *pitch* mínimo y varianza del *pitch*. Estas nuevas caracteristicas son encontradas para cada una de las detecciones de voz continua mayor a 100ms (segmentos válidos) hechas en el pre procesamiento, previamente explicado.

4.3.3.2 Características Derivadas de los MFCC

En este apartado el promedio de coeficientes MFCC para una ventana válida ha sido considerado como las características derivadas de los MFCC, experimentos realizados en [24] dan muestra de su desempeño y capacidad de decision.

4.4 Entrenamiento y Evaluación

Para el entrenamiento y la evaluación de cada uno de los clasificadores seleccionados para este trabajo de grado no se ha utilizado la totalidad de la base de datos, esto, debido al alto sesgo que se crearía sobre la etiqueta furia; precisiones del 90% serían alcanzables sin ningún problema. En cambio, se implementado una función de nombre selección_patrones, esta hace una selección aleatoria de los patrones, maximizando las muestras de la base de datos de trabajo, de manera que los patrones de esta se distribuyan 50-50, así: primero, la función evalúa cuál es la etiqueta con menos cantidad de muestras sobre la base de datos total, en seguida, selecciona cada uno de estos patrones, y finalmente, hace una selección aleatoria sobre los patrones restantes, esta última selección será del mismo tamaño de la selección hecha para los patrones de igual etiqueta, pero de menor cantidad sobre la base de datos total.

Implementado el método de validación cruzada [14] o *cross-validation*, se ha evaluado y probado cada uno de los tipos de clasificadores seleccionados en este trabajo de grado.

La base de datos de trabajo, que estará compuesta de 71 muestras para la clase furia y 71 muestras para la clase de ausencia de esta, ha sido indexada con los números del uno al tres de forma aleatoria; una distribución 50-50 de la base de datos de trabajo garantiza que en este indexado existan muestras de ambas etiquetas para procesar en etapa de entrenamiento (como mínimo existirá un 26.66% de una clase frente a un 50% de la otra, referidos a la base de datos de trabajo). Una vez hecha esta indexación, se ha entrenado y evaluado para cada uno de los índices, así: el 33.33% de los índices con numeral “1” serán seleccionados como muestra de prueba y el 66.66% restante se seleccionará como muestra de entrenamiento, de esta manera, se encuentran los indicadores de precisión (alternativamente, indicadores de eficiencia o desempeño) de entrenamiento y de prueba para este indexado, análogamente, se realiza

este procedimiento para los índices “2” y “3”. Estos indicadores de precisión son acumulados y posteriormente promediados con otros 19 procesos de indexados diferentes; un total de 60 indicadores de precisión de entrenamiento y prueba son encontrados y finalmente promediados. La motivación de este procedimiento yace en obtener una estimación más válida de estos resultados, comparados con indicadores producto de un único entrenamiento y una única prueba.

Aunque uno de los objetivos de este trabajo involucraba el uso de redes neuronales, se descartó la posibilidad de usarlas por varias razones. Varios tipos de clasificadores se usan en la tarea de reconocimiento de emociones, HMM, GMM, SVM, redes Neuronales (ANN), KNN, entre muchos otros, a pesar de esto, no existe un consenso, demostración o conclusión sobre cual clasificadores es más adecuado para clasificar emociones [1]. Si bien, las redes neuronales son usadas en múltiples aplicaciones de reconocimiento de patrones y se conocen por realizar efectivos mapeos de modelos no lineales, particularmente, la distribución sobre el hiperespacio de características de las emociones es desconocida, adicionalmente, las redes neuronales poseen el inconveniente de tener muchos criterios de diseño, como lo son la forma de la función de activación de las neuronas, el número de capas ocultas y el número de neuronas por capa, que usualmente son seleccionados a manera *ad hoc* [1], hecho que entorpece los resultados del trabajo realizado puesto que las redes entrenadas no son generalizables; finalmente, las redes neuronales son difíciles de analizar, y, según la documentación presentada en [1] estas presentan una precisión de clasificación bastante baja en comparación con otros clasificadores.

4.4.1 KNN

Mediante el comando de Matlab knnclassify() se ha entrenado y evaluado sistemas KNN para diferentes valores de k. En el script protocolo_knn.m se puede evidenciar tal procedimiento.

4.4.2 SVM

Mediante el comando de Matlab svmtrain() y svmclassify() se ha entrenado y evaluado sistemas SVM para diferentes diseños en función de los *kernel*. En el script protocolo_svm.m se puede evidenciar tal procedimiento.

4.5 Consolidación del Sistema

Luego de consolidar una base de datos de audios de locuciones en español con su respectiva etiqueta, compuesta, además, de características usadas en reconocimiento de emociones por habla, se establecieron pruebas para diferentes sistemas de reconocimiento de patrones como los SVM y los KNN, estos fueron entrenados y evaluados, así como también analizados. Se llegó al resultado de que, clasificadores tipo KNN mostraron un mejor desempeño en términos generales en la tarea de clasificación, es por esto que un clasificador de este tipo, para $k = 4$ fue seleccionado como sistema definitivo. Este clasificador puede encontrarse en el archivo con nombre sistema_final. Las características de este sistema son: 85,26% de precisión en entrenamiento (error de 14,74%) y 78,72% de precisión en evaluación (error de 21,28%).

Finalmente, y a modo de estimación, se ha elaborado un clasificador donde los insumos son audios enteros de voces, de este modo, este clasificador utiliza el concepto de mayoría simple (ver Ilustración 15) como hipótesis de clasificación, esto es, un audio es procesado extrayendo en todos sus segmentos válidos sus respectivas características junto con sus etiquetas, etiquetas discriminadas por el clasificador tipo KNN para k=4 previamente descrito, seguidamente, este clasificador evaluara que etiqueta es la de mayor repetición para el audio en evaluación, con base en esto, esta será la etiqueta propuesta para este audio.

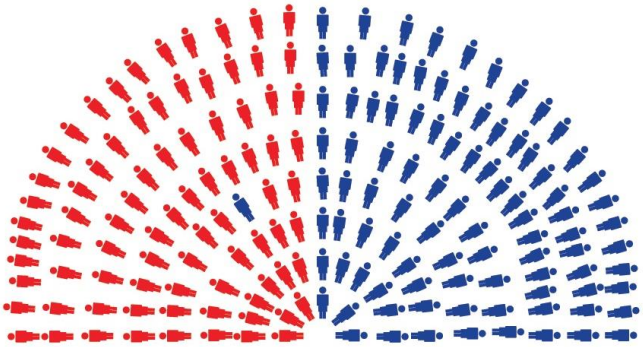


Ilustración 15 Ejemplo gráfico, mayoría simple.

5 PROTOCOLO DE PRUEBAS

Para la consolidación de la base de datos de locuciones en español se ha creado un archivo con nombre base_locuciones.m, en este se procesan cada uno de los 4 audios de la base de datos ('Base_44k.mat'), Furia_44k, NF1_44k, NF2_44h y NF3_44k, correspondientes a un audio de etiquetas de furia a una frecuencia de muestreo de 44.1 kHz, y 3 audios de etiquetas de No furia a una frecuencia de muestreo de 44.1 kHz, respectivamente. En el archivo base_audios_locuciones_caracteristicas.mat se encuentra el resultado de correr script base_locuciones.m -en mención-, este archivo contiene la frecuencia de muestreo Fs 44100 Hz, las señales discretas de audio ya mencionadas, las características para cada audio y su respectivo vector de etiquetas. Previo a correr el script base_locuciones.m es necesario cargar el archivo Base_44k.mat mediante la instrucción load('Base_44k.mat'); alternativamente existe Base_8k una base de datos análoga a Base_44k.mat, la diferencia reside en la frecuencia de muestreo que entonces es de 8 kHz. Esta última base de datos no ha sido tratada pues los filtros de Mel van de 0 Hz hasta más de 7 kHz, el resultado es que la frecuencia máxima de los filtros es mayor que la frecuencia máxima de las señales de audio.

Se ha decidido trabajar con los clasificadores KNN y SVM únicamente, dado sus sencillos métodos de entrenamiento y clasificación, pues uno de los objetivos más razonables consiste en la creación de un prototipo funcional pero también motivado por la simplicidad, clasificadores de tipo redes neuronales son difíciles de analizar, así como es difícil seleccionar su topología, hecho que provocó su interrupción en este proyecto y posteriormente su omisión, como previamente se mencionó.

Utilizando los archivos encontrados en base_final_44k.mat, junto con los archivos protocolo_knn.m y protocolo_svm.m, se puede realizar la comprobación de resultados y procedimiento hecho en entrenamiento y prueba de estos dos sistemas (KNN Y SVM), el archivo base_final_44k.mat contiene la matriz base_carac y el vector base_eti, base_carac es el conjunto de características para toda la base de datos, en sus columnas se encuentran 20 características para cada uno de los segmentos analizados, base_eti es un vector de etiquetas de 1 y -1, sus columnas están directamente relacionadas con las columnas de base_carac. Las pruebas de los sistemas se hacen utilizando estos datos y utilizando los scripts previamente mencionados, protocolo_knn.m y protocolo_svm.m.

El archivo con nombre bases_sistema_final.mat contiene la información necesaria para clasificar un patrón de audio presente en la base de datos de trabajo. El archivo con nombre sistema_final usa estos insumos para clasificar estos patrones.

Previamente, en la Ilustración 14 se ha mostrado el comportamiento de los vectores de coeficientes previos al bloque DTF-II, se puede evidenciar claramente la trayectoria de los mismos, y como estos se guían por la envolvente del logaritmo de la señal en frecuencia, resultado que evidencia el correcto y esperado funcionamiento de estos coeficientes [10].

En el archivo con nombre sistema_clasifica_completos.m se encuentra implementado el clasificador que utiliza la mayoría simple como hipótesis.

6 ANÁLISIS DE RESULTADOS

A continuación, se presenta una tabla ilustrativa (ver Tabla 5) de los resultados obtenidos para algunos valores de k para los clasificadores tipo KNN realizados (usando *cross-validation*), así como también una gráfica (ver Ilustración 16*Ilustración 16*) del comportamiento de la precisión de los clasificadores en función de k, el número de vecinos cercanos.

k	Precisión-Entrenamiento (%)	Precisión-Prueba (%)
1	100	55
2	100	57,54
3	76,65	55,46
4	73,43	56,56
5	82,05	55,02
6	82,99	58,24
7	71,11	61,84
8	78,26	56,54
9	67,84	57,74
10	75,03	56,19
11	69,17	62,79
12	73,68	55,72
13	65,28	55,76
14	72,94	57,15
15	64,6	56,8

Tabla 5 Resultados para realizaciones de KNN.

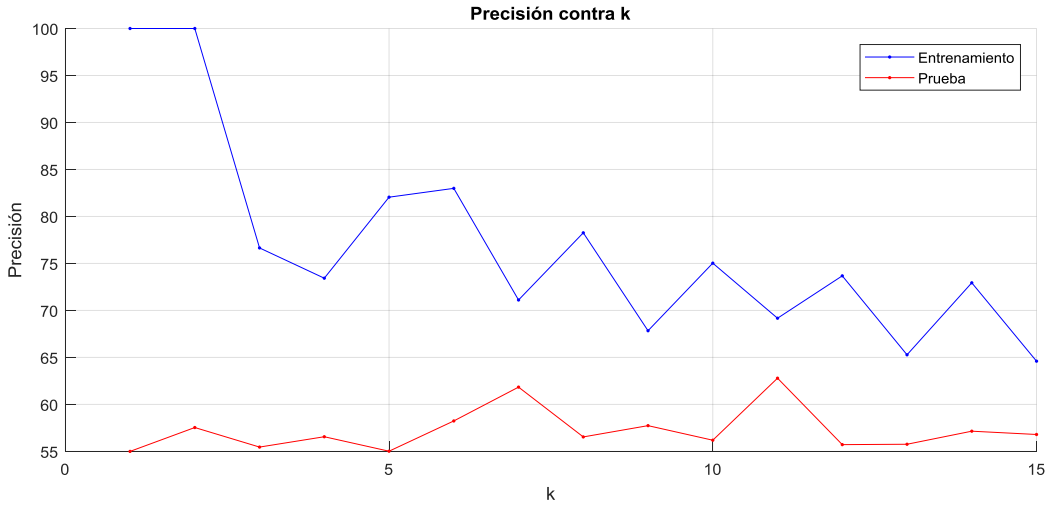


Ilustración 16 Comportamiento de precisión contra k para KNN.

Bajo el supuesto de que la base de datos está bien acondicionada, es decir, las etiquetas propuestas son acordes a los sonidos grabados, se realiza el siguiente análisis:

Es evidente el pobre desempeño expuesto por los clasificadores tipo KNN, puede verse que la diferencia entre precisión de entrenamiento y de prueba son muy amplias y que a medida que k aumenta el indicador de precisión para entrenamiento disminuye, por su parte la precisión de prueba es tan pobre que parece ser independiente del número k.

El hecho de que existan altos resultados de precisión para entrenamiento para valores de k pequeños, y al mismo tiempo valores bajos de precisión para estos mismos valores de k, da indicios de que las representaciones de los patrones son muy independientes entre sí (están aleatoriamente ubicados en el hiperespacio de características), inclusive entre la misma clase, y, por tanto, las características que los representan no parecen tener una correlación considerable, resultando inútiles en esta tarea (para este clasificador). Además, que la precisión de entrenamiento disminuya a medida que k aumenta, induce automáticamente a la sospecha de que los patrones de diferentes etiquetas se encuentran muy cercanos entre sí en el hiperespacio, y, en consecuencia, una transformación de estas características a otra dimensión puede ser una opción válida.

A continuación, se presenta en la Tabla 6 los resultados de indicadores de precisión para diferentes clasificadores tipo SVM entrenados:

Kernel-SVM	Precisión-Entrenamiento (%)	Precisión-Prueba (%)
Lineal	72	57,94
Polinomial, orden 2	99,98	52,9
Polinomial, orden 3	100	52
Polinomial, orden 4	100	54
Polinomial. orden 5	99,08	55,13
Polinomial, orden 6	84,63	51,86
Polinomial, orden 7	66,78	51,41
Polinomial, orden 8	61,44	55,23
Polinomial, orden 9	56,06	52,56
Polinomial, orden 10	52,16	49,69
RBF	100	57,4

Tabla 6 Resultados para realizaciones de SVM.

Nuevamente, puede observarse que las precisiones de entrenamiento son muy superiores a las de prueba, inclusive para *kernels* no-lineales, donde las características son transformadas a otra dimensión, en donde un hiperplano actúa como discriminador, con esto se descarta la propuesta previamente planteada. Esto conduce a la suposición de que o bien, las características no comprimen adecuadamente la información emocional presente en la voz, o bien los clasificadores seleccionado no son suficientemente apropiados en la tarea de reconocimiento de ira en señales de voz.

En el caso donde la base de datos este corrupta, es entendible, los resultados esperados, sin embargo, este es un factor con el que se debe lidiar siempre, que, entre otras cosas, agrega validez a los

análisis hechos sobre esta base, esto, en cuanto a que bases de datos de audios controlados no son el entorno real de un típico *call center*.

El sistema definitivo ha sido determinado iterativamente, es decir, se ha propuesto un método que mediante iteraciones obtuviera el mejor sistema posible, para eso múltiples entrenamientos y pruebas se hicieron para clasificadores KNN con k=4, donde se variaron las muestras de entrenamiento y de prueba, los indicadores para este sistema encontrado son un error de entrenamiento de 14,74%, mientras que error de prueba es de 21,27%, un experimento más se ha realizado sobre este sistema, se clasificó el total de la base de datos utilizando este clasificador, el error encontrado fue de 38,88%, dando por cerrado el análisis de resultados.

Finalmente, el clasificador de hipótesis propuesta como mayoría simple arroja resultados satisfactorios, discriminando los audios adecuadamente, tal y como se presenta en la Tabla 7.

	Audio 1	Audio 2	Audio 3	Audio 4
Etiqueta	Furia	No Furia	No Furia	No Furia
Clasificación	Furia	No Furia	No Furia	No Furia

Tabla 7 Clasificación por mayoría simple.

Una comparación con trabajos anteriores es inválida una vez que este trabajo es únicamente aceptado para la base de datos utilizada, contrastar los resultados acá obtenidos con los obtenidos en investigaciones aisladas a las condiciones de este trabajo es ilógico. Puede anotarse que en general el resultado de esta investigación no es satisfactorio en su totalidad, sin embargo, ciertas conclusiones pueden ser obtenidas a partir de este.

7 CONCLUSIONES Y RECOMENDACIONES

- Las emociones, como un concepto subjetivo, son difíciles de clasificar, existe el error de interpretación del humano, así como también una interpretación personal de las señales de voz, la no existencia de consenso en esta área dificulta su tratamiento, y, en consecuencia, la validez de las etiquetas propuestas para esta base de datos no es indiscutible.
- Para la base de datos trabajada en este proyecto, los métodos de clasificación implementados junto con las características seleccionadas no proporcionan buenos resultados, se sugiere cambiar la base de datos para análisis sobre las características.
- Las características utilizadas, sin embargo, siguen teniendo valor, y, en consecuencia, se sugiere ser utilizadas en contextos diferentes para evaluar la capacidad de compresión de información de señales de voz. Ejemplo: verificación o reconocimiento de locutor, donde las etiquetas están bien definidas.
- Se sugiere explorar la detección de furia para esta base de datos utilizando características diferentes.
- El presente trabajo es una aproximación a un problema con pocas restricciones sobre la base de datos, esperando que un contexto real envolviera el trabajo como tal, se sugiere para trabajos futuros el uso de una base de datos diferente.
- El sistema final seleccionado contiene información importante, presenta una capacidad de clasificación superior a la mayoría de sistemas realizados, y, por ende, un estudio de los patrones que se usan como base para este merecen un análisis sugerido.
- El clasificador con hipótesis de mayoría simple es eficaz en la tarea de detección de furia. Una base de datos más amplia es sugerida para un análisis más profundo de este.

8 BIBLIOGRAFÍA

[1] M. El Ayadi, M. S. Kamel, y F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognit.*, vol. 44, núm. 3, pp. 572–587, 2011.

[2] R. Chayo-dichy, A. Elvira, V. García, N. A. García, G. Castillo-Parra, y F. Ostrosky-Solis, "Valencia, activación, dominancia y contenido moral, ante estímulos visuales con contenido emocional y moral: un estudio en población mexicana", *Rev. Española Neuropsicol.*, vol. 225, pp. 213–225, 2003.

[3] H. Pérez y C. A. Reyes, "Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo", 2010.

[4] M. Chóliz., "Psicología de la emoción: el proceso emocional", *Psicol. la emoción*, pp. 1–34, 2005.

[5] C. E. Williams y K. N. Stevens, "Vocal correlates of emotional states", *Speech Eval. psychiatry*, pp. 189–220, 1981.

[6] P. K. Mongia y R. K. Sharma, "Estimation and statistical analysis of human voice parameters to investigate the influence of psychological stress and to determine the vocal tract transfer function of an individual", *J. Comput. Networks Commun.*, vol. 2014, 2014.

[7] D. Bonomo Laynez, "Sistemas De Verificación Automática De Locutor", pp. 23–33, 2012.

[8] N. Seo, "ENEE632 Project4 Part I: Pitch Detection", 2008.

[9] S. B. Davis y P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", núm. 4, 1980.

[10] L. R. Rabiner y R. W. Schafer, *Introduction to digital speech processing*, vol. 1, núm. 1. 2007.

[11] D. Reig, "Implementación de algoritmos para la extracción de patrones característicos en Sistemas de Reconocimiento De Voz en Matlab", 2014.

[12] R. O. Duda, P. E. Hart, y D. G. Stork, *Pattern Classification*, Segunda Ed. .

[13] S. Theodoridis y K. Koutroumbas, *Pattern Recognition*, 2a ed. Elsevier.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer.

[15] C. E. Williams y K. N. Stevens, "Emotions and Speech: Some Acoustical Correlates", *J. Acoust. Soc. Am.*, vol. 52, núm. 4B, pp. 1238–1250, 1972.

[16] D. Bonomo Laynez, "Sistemas De Verificación Automática De Locutor", 2012.

[17] N. Sato y Y. Obuchi, "Emotion Recognition using Mel-Frequency Cepstral Coefficients", *Inf. Media Technol.*, vol. 2, núm. 3, pp. 835–848, 2007.

[18] O. E. Korkmaz y A. Atasoy, "Emotion Recognition from Speech Signal Using Mel-Frequency Cepstral Coefficients", pp. 1254–1257, 2015.

[19] S. Bedoya-Jaramillo, E. Belalcazar-Bolanos, T. Villa-Canas, J. R. Orozco-Arroyave, J. D. Arias-Londono, y J. F. Vargas-Bonilla, "Automatic emotion detection in speech using mel frequency cesptral coefficients", *STSIVA 2012 - 17th Symp. Image, Signal Process. Artif. Vis.*, pp. 62–65,

2012.

[20] C. M. Lee *et al.*, “Emotion Recognition based on Phoneme Classes”, *Database*, núm. 1, pp. 889–892, 2004.

[21] L. R. Rabiner y R. W. Schafer, *Digital Processing of Speech Signals*. Pearson, 1978.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, y B. Weiss, “A Database of German Emotional Speech”, *Interspeech*, pp. 1517--1520, 2005.

[23] J. Benesty, M. M. Sondhi, Y. Huang, y S. Greenberg, *Springer Handbook of Speech Processing*. 2009.

[24] E. Rueda y Y. Torres, “Identificación de Emociones en la Voz”, vol. 40, núm. 1, 2008.

9 ANEXOS

ANEXO 1