1. TIPO DE DOCUMENTO: Trabajo de grado para optar por el título de INGENIERO DE SONIDO.

2. TÍTULO: DESARROLLO E IMPLEMENTACION DE UN SISTEMA DE PROCESAMIENTO DE VOCES PARA EL ANÁLISIS DE TRES ESTADOS EMOCIONALES

3. AUTOR: Juan Daniel Morales Piedrahita

4. LUGAR: Bogotá, D.C**.**

5. FECHA: Noviembre de 2013

6. PALABRAS CLAVE: Coeficientes Delta, Matriz de Confusión, MFCC, , Reconocimiento de Emociones, Redes Neuronales Artificiales.

7. DESCRIPCIÓN DEL TRABAJO: El objetivo principal de este proyecto es desarrollar un algoritmo de reconocimiento de voz que permita identificar los estados emocionales de ira, tristeza y alegría. Se realizó una investigación sobre los algoritmos y parámetros existentes para el reconocimiento emocional de la voz y se seleccionaron los parámetros que presentaban mayor porcentaje de reconocimiento para el estudio. Finalmente se realizaron pruebas al sistema implementado y se compararon los resultados con antecedentes institucionales, nacionales e internacionales.

8. LÍNEAS DE INVESTIGACIÓN: Línea de Investigación de la USB: Tecnologías actuales y Sociedad. Sub linea de la Facultad de Ingeniería: Acústica. Campo Temático del Programa: Procesamiento digital de señales.

9. FUENTES CONSULTADAS:

- FRIBERG, Anders*. Digital Audio Emotions - An Overview of Computer Analysis and Synthesis of Emotional Expression in Music.* Paper. Espoo, Finland.September 1-4, 2008.
- MOURJO*POU*LO*S,* J., TSOUK*A*LA*S,* D. *Neural Network Mapping to Subjective Spectra of Music Sounds.* Presented at the 90th AES Convention. Paris. February 19-22, 1991.

- KOSTEK, Bozena. *Application of Learning Algorithms to Musical Sound Analyses.* Presented at the 97th AES Convention. San Francisco. November 10-13, 1994.
- FABBR*I* Richard J*. Neuralphoneme recognition during a cocktail party.* Presented at the 94th AES Convention. Berlin. March 16-19, 1993.
- CHRISTENSEN, Niels Sander; CHRISTENSEN, Karl Ejner; WORM, Henning*. Classification of Music Using Neural Net.* Presented at the 92th AES Convention. Vienna. March 24-27, 1992.
- KOSTEK, Bozena*. Parametric Representation of Musical Phrases.* Presented at the 101th AES Convention. Los Angeles. November 8-11, 1996.
- KOSTEK, Bozena*. Feature Extraction Methods for the intelligent processing of musical signals.* Presented at the 99th AES Convention. New York. October 6-9, 1995.
- PALOMAKI, Kalle, *Neural Network Approach to Analyze Spatial Sound.* Paper 16th AES International Conference. Espoo, Finlandia.
- SCHMIDMER, Keyh *.A combined measurement tool for the objective, perceptual based evaluation of compressed speech and audio signals.* Presented at the 106th AES Convention. Munich, May 8-11, 1999.
- SZCZERBA, Marek. *Recognition and Prediction of Music -A Machine Learning Approach.* Presented at the 106th AES Convention. Munich, May 8-11, 1999.
- OUDEYER, Pierre. *The Production and Recognition of Emotions in Speech: Features and Algorithms".* Paper Released in Science Direct Human Computer Studies, Sony CSL. Paris, Nov 30, 2002.
- PRADIER, Melanie. *Emotion Recognition from Speech Signals and Perception of Music.*Thesis Degree Paper, Stuttgart University.Germany.2011
- SIDOROVA,Julia. *Speech Emotion Recognition.*PhD Paper, Universitat Pompeu Fabra. España. Jul 4, 2007
- TREJOS, H. URIBE C. *Motor Computacional de Reconocimiento de Voz: Principios básicos para su Construcción.* Documento de Grado. Universidad Tecnológica de Pereira.Nov 2007
- ALDANA A. PIÑEROS J.*Desarrollo e Implementación de un Algoritmo de Reconocimiento de Voz que Permita Seleccionar una Imagen a Partir de un Banco de Nueve Fotografías Utilizando Redes Neuronales.* Documento de Grado, Universidad San Buenaventura. Bogotá, Oct. 22, 2010
- PANG, Yixiong. *Speech Emotion Recognition Using Support Vector Machine.* Paper, Jiao Tong University. China. 2012
- MIYARA, Federico. *Pasos del algoritmo k-means.* [En línea]. Curso virtual de Procesamiento Digital de Señales en Voz. Facultad de Ciencias Exactas, Ingeniería y Agrimensura. Argentina. Disponible en la Web: "http://www.fceia.unr.edu.ar/prodivoz/Clustering_bw.pdf"

- DO, Min .*An Automatic Speaker Recognition System.* [En línea]. DSP Mini Project. Universidad de Illinois. USA. Disponible en la Web: "http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/"
- LI, Eldon. *Artificial neural networks and their business applications.* Paper, National Chung Cheng University.China.1994
- DELGADO, Alberto. *Aplicación de las Redes Neuronales en Medicina.*Paper, Universidad Nacional de Colombia.Colombia.1994
- Grupo de Investigación de Redes Neuronales Artificiales. *Perceptrón.* [En línea]. Universidad Carlos III. España. Disponible en la Web: "http://www.lab.inf.uc3m.es/~a0080630/redes-de-neuronas/perceptron-simple.html"
- SERGIU, Ciumac. *Feedforward Artificial Neural Network.* [En línea]. Universidad de Moldavia. Disponible en la Web: "http://www.codeproject.com/Articles/175777/Financial-predictor-via-neural-network#conclusion"
- HUDSON, Martin. *Neural Network Toolbox User's Guide.* Reference.2011
- BURKHARDT, PAESCHKE. *A Database of German Emotional Speech.* Paper, Technical University of Berlin. Alemania. 2000
- RUEDA,E,TORRES,Y. *Identificación de emociones en la voz .*Thesis Degree Paper, Universidad Distrital de Santander. Colombia .2007.
- MORALES, M, ECHEVERRY, J, OROZCO, A. *Reconocimiento de emociones empleando procesamiento digital de la señal de voz.* Documento de Grado, Universidad Tecnológica de Pereira. Colombia.
- MONTERO, GUTIERREZ. *Analysis and modelling of emotional speech in spanish. Documento de Grado*, Universidad Politécnica de Madrid. España.
- SCHULLER, Bjorn, RIGOLL, Gerhard. *Hidden Markov model based speech emotion recognition. Thesis Paper*, Universidad Técnica de Munich. Alemania. 2003
- DARWIN, Charles. *The expression of the Emotions in Man and Animals.* Escocia. 1872
- Moving Pictures Experts Group *. Overview of the MPEG-4 Standard*. Norma. 2002.

10. CONTENIDOS:

- Reconocimiento de voz:

Rama de investigación enmarcada en el aprendizaje computacional, en la que se establecen metodologías de implementación para el desarrollo de sistemas inteligentes que permiten establecer parámetros de análisis de la voz humana.

- Coeficientes Cepstrales en escala Mel

Parámetros espectrales de una señal que muestran la contribución pico de la energía en bloques de tiempo pequeños. Estas contribuciones están acotadas a partir de unos filtros que simulan la respuesta en frecuencia del sistema fonador humano.

- Matriz de Confusión:

Herramienta estadística que permite evaluar la eficiencia de los algoritmos de reconocimiento.

- Redes neuronales artificiales :

Sistemas heurísticos de reconocimiento de patrones que, mediante un aprendizaje supervisado, realizan un ajuste en la linealización de la entrada a la red para encontrar generalizaciones de sistemas lineales y no lineales.

11. METODOLOGÍA: Es de carácter empírico-analítico. Se propone un algoritmo de investigación iterativo para el desarrollo del proyecto.

12. CONCLUSIONES:

- El análisis energético de las señales de audio para la extracción de parámetros es un método eficaz de obtener datos que pueden ser utilizados en el entrenamiento de algoritmos de reconocimiento emocional. El análisis de los coeficientes cepstrales, al igual que de su velocidad y aceleración, validan las similitudes y marcadas diferencias de las emociones estudiadas con respecto a sus componentes de potencia, valencia y activación.

- Las redes neuronales artificiales son algoritmos de reconocimiento que, por medio de aprendizaje computacional, permiten realizar predicciones específicas de estados emocionales al ser entrenadas con parámetros energéticos.

- El uso de las redes neuronales artificiales fue útil para encontrar una generalización en los patrones de reconocimiento, de esta forma se validan los resultados obtenidos en las redes binarias creadas en las primeras iteraciones del proyecto.

- La diferencia en el porcentaje de reconocimiento general de diferentes voces en español es muy baja, por tanto los parámetros emocionales extraídos de la voz son independientes del género, edad y grado de experticia de los locutores, este resultado puede corroborarse en investigaciones futuras que involucren la creación de un corpus de voces con emociones no actuadas.

- Los resultados de las pruebas realizadas entre idiomas se deben a las fuertes diferencias entre las lenguas analizadas (Anglosajona y Romance), al igual que las diferencias entre las condiciones de captura de los dos corpus de audio utilizados (cámara anecóica, preselección de locutores, entre otros.).

- El porcentaje de reconocimiento promedio general de todas las redes neuronales implementadas fue de un 74.03%, un valor mucho más elevado que el porcentaje de reconocimiento emocional alcanzado por sujetos de prueba humanos.

- Los resultados obtenidos en el desarrollo del proyecto son comparables, y en algunos casos superan, los resultados de investigaciones similares a nivel nacional e internacional.

- Al poder realizar un reconocimiento satisfactorio de las emociones en la voz utilizando algoritmos de reconocimiento entrenados con voces foráneas, se comprueba la teoría de parámetros universales propuesta por Charles Darwin [1], para voces Colombianas.

DESARROLLO E IMPLEMENTACIÓN DE UN SISTEMA DE PROCESAMIENTO DE VOCES PARA EL ANÁLISIS DE TRES ESTADOS EMOCIONALES

JUAN DANIEL MORALES PIEDRAHITA

UNIVERSIDAD DE SAN BUENAVENTURA
FACULTAD DE INGENIERÍA
INGENIERÍA DE SONIDO
BOGOTA
2013

DESARROLLO E IMPLEMENTACIÓN DE UN SISTEMA DE PROCESAMIENTO DE VOCES PARA EL ANÁLISIS DE TRES ESTADOS EMOCIONALES

JUAN DANIEL MORALES PIEDRAHITA

Cod: 20093235075

Trabajo presentado como requisito para optar por el título de profesional en Ingeniería de Sonido

Tutor:
Ing. Marcelo Herrera, Ph. D.

UNIVERSIDAD DE SAN BUENAVENTURA
FACULTAD DE INGENIERÍA
INGENIERÍA DE SONIDO
BOGOTA
2013

Nota de Aceptación:

_____
_____
_____
_____
_____
_____

_____

Director de Programa

_____

Jurado

_____

Jurado

Bogotá, Octubre 25 de 2013

Dedicado a mi hijo Juan Carlos, porque su luz ilumina mi trabajo, mi camino y mi vida.

AGRADECIMIENTOS

Por su colaboración y asesoría:
A mi director de tesis Ph.D. Marcelo Herrera Martínez

Por su asesoría:

Ing. Camilo Pino
D.I. Diego Pino

Por su colaboración y apoyo desinteresado:
Auxiliares laboratorio de sonido Universidad de San Buenaventura

Y a todas aquellas personas, familiares y amigos, que de una u otra manera aportaron intelectual, o personalmente, durante el desarrollo del proyecto.

CONTENIDO

LISTA DE TABLAS

LISTA DE ECUACIONES

LISTA DE FIGURAS

INTRODUCCIÓN

El análisis emocional del habla trata del uso de diferentes métodos para analizar el comportamiento vocal como identificador de parámetros característicos en diferentes estados emocionales, enfocando su centro de estudio en los aspectos no verbales del habla.

El procesamiento de señales es uno de los campos más importantes en cuanto a audio se refiere, ya que mediante su análisis se justifican los conceptos en los que se fundamentan las técnicas de grabación, mezcla, comportamientos acústicos, diseño de sistemas digitales y análogos, entre otros. Dentro de los tipos de señales importantes a analizar, cabe resaltar la importancia de la voz como índice caracterizador de información del ser humano, la psicología, dentro de sus múltiples campos de acción, estudia cómo los fonemas, cambios de altura, tono y timbre de un espectro vocal, pueden dar una descripción objetiva del comportamiento y las emociones humanas. Las aplicaciones de este análisis abarcan los campos de la filosofía, las ciencias sociales, humanas, económicas y políticas, estudiando los efectos de los estados emocionales en las diferentes interacciones entre los hombres. Este proyecto de grado propone el diseño de un sistema de análisis espectral para el reconocimiento de tres estados emocionales (“ira”,”felicidad”,”tristeza”), a partir del estudio de parámetros globales de la voz humana, independientes de su origen morfológico.

Se presentará, adicionalmente, un estudio de eficiencia del sistema de reconocimiento, comparando los resultados obtenidos con estudios similares llevados a cabo en la Universidad de San Buenaventura, Sede Bogotá, y en otros proyectos internacionales que involucran el estudio de estados emocionales a partir del análisis espectral.

1. PLANTEAMIENTO DEL PROBLEMA

1.1. ANTECEDENTES

A nivel internacional, se han llevado a cabo diferentes proyectos de investigación sobre reconocimiento emocional de la voz, resaltando el trabajo de Pïerre-Yves Oudeyer, quien en el 2002 realizó una comparación entre los diferentes parámetros y algoritmos existentes para la producción y reconocimiento de emociones en el habla, para su subsecuente aplicación en expresiones robóticas. En el 2007, Julia Sidorova, para su proyecto de grado en el doctorado de "Ciencias Cognitivas del Lenguaje", menciona la importancia del SER (Speech Emotion Recognition), en campos como las interfaces humano-maquina, cambios de interacción en sistemas de ventas y, adicionalmente, propone un sistema de reconocimiento de estados emocionales comprobado con bases de datos de archivos de voces actuadas, con un nivel de eficiencia alto. Otro de los trabajos más importantes realizados es el realizado por Melanie Fernández Pradier en la Universidad de Stuttgard, en donde realiza una descripción profunda de las diferentes teorías de clasificación de estados emocionales a partir de parámetros globales evolutivos inherentes en el ser humano. Adicionalmente, Fernández propone un algoritmo de reconocimiento que evalúa parámetros no convencionales para el reconocimiento de estados emocionales, siendo estos marcadores obtenidos a partir del estudio espectral de contenidos musicales.

A nivel nacional se destaca el trabajo realizado por Hernando Antonio Trejos Posada y Carlos Andrés Uribe Pérez en la Universidad Tecnológica de Pereira en el 2007, en donde se diseño un motor computacional de procesamiento de audio en c++, el cual, mediante un algoritmo de redes neuronales artificiales, realiza una comparación de parámetros formantes en la voz, comparándolos con patrones de dispersión en el campo acústico para realizar el reconocimiento.

A nivel institucional se encuentran dos proyectos de grado, para Ingeniería Electrónica e Ingeniería se Sonido, en el que realizan procesamiento de voces para aplicaciones de reconocimiento y domótica:

En el 2007 se presentó el trabajo de grado "Diseño de un Dispositivo para El Reconocimiento de Caracteres Vocálicos, para Ordenar Comandos al Televisor" por Javier Alfredo Báez, en donde se implementó un algoritmo de reconocimiento automático del habla en el dispositivo de orden de comandos, dicho algoritmo es utilizado usualmente para el reconocimiento de caracteres vocálicos en el lenguaje.

En el 2010 se presentó el trabajo de grado "Desarrollo e Implementación de un Algoritmo de Reconocimiento de Voz que Permita Seleccionar una Imagen a partir de un Banco de Nueve Fotografías utilizando Redes Neuronales" en donde se desarrollo, en primera instancia, un algoritmo de redes neuronales artificiales para el reconocimiento, obteniendo con este una eficiencia del 17%, subiendo dicho valor a 87% implementando un algoritmo de reconocimiento de patrones para el procesamiento de formantes en el habla.

1.2 DESCRIPCIÓN Y FORMULACIÓN DEL PROBLEMA

La producción y reconocimiento emocional en la voz humana es un campo de estudios ampliamente investigado y utilizado en diferentes mecanismos de interacción entre individuos y sistemas, su importancia se deriva de las investigaciones evolutivas desarrolladas por Charles Darwin, el cual propone la existencia de parámetros universales dentro de las especies, independientes de sus ubicaciones geográficas, bajo los cuales se puede hacer una descripción cualitativa de sus estados de ánimo.
Actualmente la investigación para el descubrimiento de nuevos parámetros de procesamiento del habla continúa, sin embargo en Colombia no han existido avances importantes en el tema.
Partiendo del nivel de importancia que posee este campo de estudios, y de la falta de desarrollo del mismo en el país, se formula: *¿Cómo se pueden reconocer de manera eficaz los estados emocionales humanos a partir del habla?*

1.3 JUSTIFICACIÓN

Las emociones tienen un fuerte impacto en las decisiones que toma el ser humano, debido a que el cerebro contiene muchos sistemas relacionados con los estados emocionales, desde procesos de almacenamiento de información, hasta decisiones comerciales y personales.
Los algoritmos de reconocimiento de voz han sido un campo de estudio desde 1992, y en 20 años se han podido realizar avances importantes debido al aumento en la capacidad de procesamiento computacional en la modulación estadística de parámetros, así pues, cada vez más parámetros pueden ser analizados simultáneamente para encontrar una respuesta precisa a las diferentes condiciones humanas identificables a partir de la voz.
Dentro de los diferentes métodos de análisis del habla, el nivel de descripción acústico ofrece las mayores ventajas a nivel de objetividad, economía y facilidad de implementación, puesto que no es necesario ningún tipo de medición ni de dispositivos específicos para su desarrollo.

1.4 OBJETIVOS DE LA INVESTIGACIÓN

1.4.1 OBJETIVO GENERAL

Desarrollar un algoritmo de reconocimiento de voz que permita identificar los estados emocionales de ira, tristeza y alegría.

1.4.2 OBJETIVOS ESPECÍFICOS

- Determinar los parámetros acústicos que se tendrán en cuenta para la identificación de estados emocionales
- Determinar qué tipo de algoritmos serán implementados para la evaluación.
- Entrenar los algoritmos de reconocimiento mediante una base de datos de archivos de audio clasificados por estado emocional.
- Comprobar el funcionamiento del algoritmo con archivos de audio foráneos y colombianos.

1.5 ALCANCES Y LIMITACIONES DEL PROYECTO

1.5.1 ALCANCES

El proyecto está enfocado a generar el sistema de reconocimiento, su comprobación con señales nativas y foráneas y la comparación de los resultados obtenidos en la comprobación con resultados de eficiencia de proyectos nacionales e internacionales.

1.5.2 LIMITACIONES

Este proyecto se limitará al estudio de muestras de audio de voces actuadas, puesto que las bases de datos disponibles para el estudio fueron realizadas por actores pagados, sin embargo las pruebas realizadas en las bases de datos para su uso en el procesamiento de audio son lo suficientemente confiables como para ser utilizadas en el proyecto.

Es necesario considerar los diferentes parámetros de error no evadibles debido a las condiciones presentes en la voz humana, tales como la ambigüedad, el intento consciente de ocultar estados emocionales, la influencia del lenguaje y la sociedad, entre otros. Dichos parámetros de error pueden ser solucionados, en su mayoría, utilizando archivos de audio con voces actuadas, puesto que los interlocutores enfatizan los parámetros emocionales característicos de los estados emocionales para realizar un estudio más limpio.

1.5.3 PROYECCIÓN

Aunque este proyecto se limita al diseño virtual del algoritmo, en el futuro puede estar dirigido a la generación de diferentes tipos de tecnologías que hagan uso del sistema, tales como plugins para reproductores musicales, físicos y virtuales, diseño de sistemas físicos de procesamiento vocal y análisis espectral de voz humana para acústica forense. Adicionalmente, la facilidad de adaptación de los algoritmos de reconocimiento permite la implementación del software en un dispositivo independiente, tal sea el caso de sistemas DSPs, portarretratos digitales adaptados como interfaz de audio, entre otros.

2. MARCO DE REFERENCIA

2.1 MARCO TEÓRICO - CONCEPTUAL

2.1.1. REPRESENTACIÓN DE LA VOZ HUMANA COMO UN SISTEMA DE FILTROS LINEALES

Para poder empezar a realizar un análisis objetivo sobre los parámetros necesarios para el reconocimiento emocional del habla, es necesario hacer una breve descripción de cómo modelan el habla dichos sistemas, siendo el modelo fuente-filtro el más utilizado.
La generación de sonidos mediante la voz humana suele ser modelado como una convolución lineal entre la fuente (frecuencia fundamental producida por las cuerdas vocales) y un filtro (el tracto vocal, el cual debido a su forma, realza o atenúa ciertos armónicos de la frecuencia fundamental debido a efectos de resonancia),este sistema puede ser modelado como una entrada periódica (hablando de sonidos vocálicos periódicos) que pasa por un filtro variable, la salida entonces estará delimitada por la potencia imbuida por los pulmones, los armónicos de la frecuencia fundamental f_0 generada por las cuerdas vocales, y por último las formantes, generadas por el tracto vocal debido a las inflexiones musculares en el conducto.

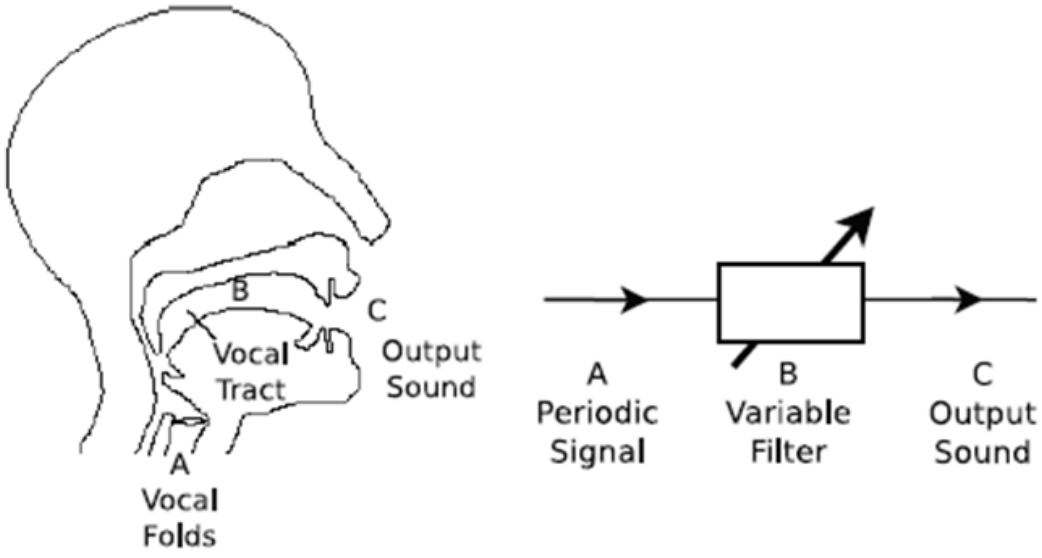


Figura 1**.** Representación del modelo fuente-filtro en el tracto vocal.
Fuente. Epps, J., Smith, J.R. and Wolfe, J. (1997) "A novel instrument to measure acoustic resonances of the vocal tract during speech" Measurement Science and Technology 8, 1112-1121.).

Se puede observar entonces que el sistema fonatorio humano se puede modelar a partir de un sistema de filtros variables, las carácterísticas de estos filtros son dependientes de la linealidad de los parámetros de entrada hacia el sistema. Diversas investigaciones se están llevando a cabo para el descubrimiento de las características que debe poseer el sistema del procesamiento fonatorio, uno de los modelos de estudio que presenta resultados e implementaciones novedosas al estado del arte del proyecto es el procesamiento y simulación computacional actual concerniente al reconocimiento de voz.

2.1.2. CARACTERÍSTICAS EMOCIONALES DE LA VOZ HUMANA

Aunque las emociones son el resultado de experiencias subjetivas inherentes al ser humano, existen dos teorías dentro de la psicología que consideran las emociones como variables discretas y continuas.

La teoría discreta, propuesta por Paul Ekman[2] ,define siete emociones básicas: felicidad, tristeza, enojo, ansiedad, aburrimiento, disgusto y neutral. Combinaciones de las emociones base dan lugar a emociones complejas que conjugan los sentimientos humanos.

La teoría continua, propuesta por Harold Schlosberg, propone cada emoción como una combinación lineal de parámetros de clasificación denominados: valencia (que tan positiva o negativa es la emoción), activación (grado de excitación), y potencia (niveles energéticos de la emoción), usualmente se realiza una correlación entre las teorías de Ekman y Schlosberg para poder ubicar las emociones básicas del plano discreto en el espacio continuo propuesto por la teoría continua.

- [2] PRADIER, Melanie. *Emotion Recognition from Speech Signals and Perception of Music.*Thesis Degree Paper, Stuttgart University.Germany.2011

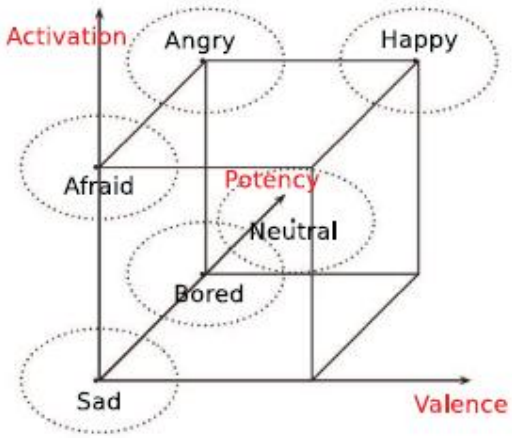


Figura 2**.** Espacio emocional tridimensional y seis emociones básicas.
Fuente: PRADIER, Melanie. *Emotion Recognition from Speech Signals and Perception of Music* .Thesis Degree Paper, Stuttgart University.Germany.2011

La comunicación a partir de la voz humana tiene dos mecanismos principales para la transmisión de información entre individuos, el canal de comunicación lingüístico realiza una transmisión directa de la información, ofreciendo un nivel de entendimiento literal de los datos que se están expresando, por ejemplo, el decir "estoy triste" da una información completa que indica el estado emocional del locutor. El otro canal de comunicación de la voz tiene que ver con el canal paralingüístico del habla, en el que se estudia la forma en la que el sonido y las expresiones están siendo generadas, y no el contenido específico de la información emitida.

Dentro de los diferentes estudios realizables sobre el canal paralingüístico, el análisis de parámetros acústicos y energéticos del audio ocupa un lugar importante en el espectro investigativo. La voz humana tiene dos componentes sonoros: los sonidos vocálicos, producto de las vibraciones del sistema fonador y la caja torácica, y los sonidos no vocálicos producto de silencios y ruidos provenientes de las cavidades de la boca. Diferentes parámetros tradicionales extraídos de la voz humana para las componentes vocálicas y no vocálicas serán descritos en el apartado **[2.1.7]**[3]

[3]PRADIER, Melanie. *Emotion Recognition from Speech Signals and Perception of Music.* Thesis Degree Paper, Stuttgart University.Germany.2011

2.1.3.ALGORITMOS DE RECONOCIMIENTO DE PATRONES

Existen diferentes algoritmos de reconocimiento de patrones, generalmente la selección de un algoritmo está relacionado con los resultados de reconocimiento sobre pruebas de validación.

Todos los algoritmos de reconocimiento deben ser entrenados, o enseñados, dependiente de las características funcionales del sistema, y del conjunto de datos al que se le desean analizar los patrones.

Existen cuatro tipos de algoritmos de reconocimiento, definidos por su forma de aprendizaje:

- Aprendizaje supervisado: el programador delimita el conjunto de datos de entrada y de objetivos para el entrenamiento del algoritmo.
- Aprendizaje no supervisado: el algoritmo intenta reconocer patrones inherentes en los conjuntos de datos, estableciendo núcleos o "clusters" en puntos de concentración de los parámetros de entrada analizados.
- Aprendizaje Semisupervisado: en el caso de poseer grandes cantidades de datos, se hace conveniente establecer una matriz de salida esperada solo para una porción de los datos de entrada, de esta forma el sistema reduce cargas de procesamiento sin sacrificar sustancialmente eficiencia en el reconocimiento.
- Aprendizaje Reforzado: Este tipo de algoritmo es utilizado para optimizar parámetros de control, en el que la salida es retroalimentada hacia la entrada para poder realizar compensaciones necesarias para el ajuste de parámetros específicos.

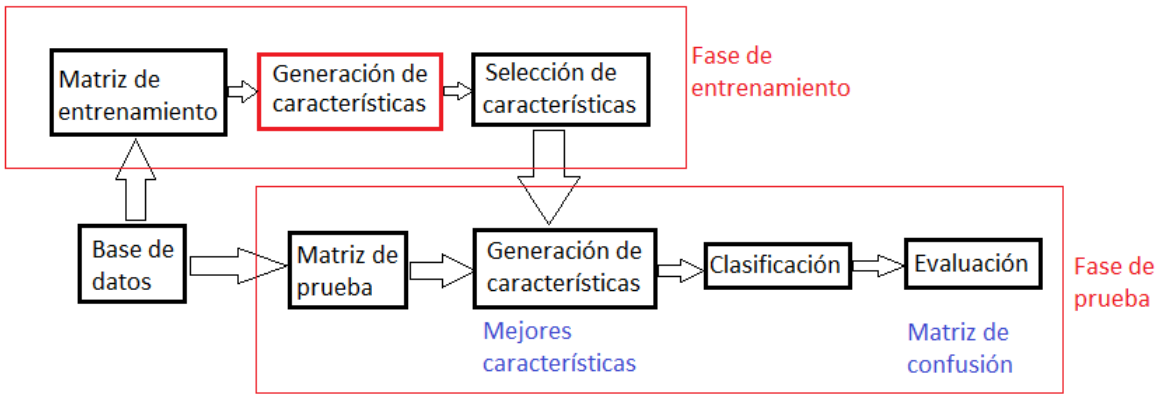


Figura 3.Diagrama de bloques general de un algoritmo de reconocimiento.

Los algoritmos de reconocimiento más utilizados son:

2.1.3.1. SUPPORT VECTOR MACHINE

Un sistema de clasificación SVM es un algoritmo de reconocimiento desarrollado por Vladimir Vapnik en los laboratorios de AT&T, el principio de funcionamiento es similar al de las redes neuronales artificiales: se establece un conjunto de parámetros de entrada y, en el caso del aprendizaje supervisado, un conjunto de salidas deseadas. La diferencia fundamental entre el SVM y las redes neuronales artificiales radica en la cantidad de salidas del algoritmo SVM, puesto que este solo permite una salida binaria para todo el procesamiento.

El principio de funcionamiento de las SVM es el encontrar un hiperplano (conjunto de elementos diferenciadores) que permita la clasificación de los datos en dos categorías, esto es, generar una linealización en el espacio (plano divisorio) entre los conjuntos de datos para que así la mayoría de los datos pertenecientes a una categoría se separen de la mayoría de los datos pertenecientes a la segunda[4].

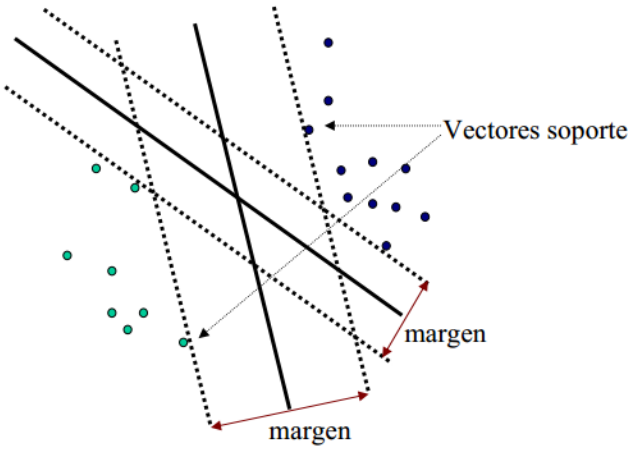


Figura 4. Organización de clase utilizando SVM.

Fuente: PANG, Yixiong. *Speech Emotion Recognition Using Support Vector Machine.* Paper, JiaoTong University. China. 2012

En la gráfica anterior se puede ver la generación de un sistema SVM para un conjunto de datos, en este, los márgenes de liealización establecen el hiperplano divisorio entre las dos categorías presentes en la matriz de entrada.

[4]PANG, Yixiong. *Speech Emotion Recognition Using Support Vector Machine.* Paper, JiaoTong University. China. 2012

2.1.3.2. CUANTIZACIÓN VECTORIAL (VQ)

En aplicaciones en las que el conjunto de parámetros de aprendizaje es muy extenso, se suele realizar una extracción de los vectores característicos de la matriz de entrada, estos vectores característicos son almacenados posteriormente en una matriz denominada "Libro de Código" , reduciendo así la cantidad de parámetros de entrada a un sistema vectorial homogéneo, el cual será revisado por los audios de prueba para encontrar el vector característico que más se ajuste a la salida deseada del algoritmo.

Una característica del reconocimiento utilizando la cuantización vectorial es que el entrenamiento del sistema (la generación del libro de código) se implementa a partir del algoritmo de clasificación denominado "K-means".

Las etapas del algoritmo K-means se describen a continuación:

1. Se eligen arbitrariamente M vectores del conjunto de L vectores de entrenamiento como conjunto inicial de vectores (palabras código) del libro de códigos.
2. Para cada vector de entrenamiento, se busca el vector en el libro de códigos actual que se acerque más al vector de entrenamiento (la distancia entre los vectores de entrenamiento y las palabras código se determina a partir del error cuadrático medio) y se asigna la celda (fila/columna) contigua a la palabra código para almacenar el vector de entrenamiento.
3. Se actualiza el vector característico a cada palabra código definiendo nuevamente los centroides de los vectores de entrenamiento almacenados en el libro de códigos.
4. Se repiten los pasos 2 y 3 hasta que el total de vectores de entrenamiento haya sido almacenado en el libro de códigos (o el error cuadrático medio haya alcanzado un valor mínimo definido)[5].

La resultante del algoritmo de K-means es la organización en el libro de códigos de los vectores característicos a los datos a reconocer, estableciendo núcleos (clusters) similares a los ubicados en el algoritmo SVM.

[5]MIYARA, Federico. *Pasos del algoritmo k-means.* [En línea]. Curso virtual de Procesamiento Digital de Señales en Voz. Facultad de Ciencias Exactas, Ingeniería y Agrimensura. Argentina. Disponible en la Web: "http://www.fceia.unr.edu.ar/prodivoz/Clustering_bw.pdf"

En el siguiente gráfico se puede observar el principio de creación del libro de códigos, la distorsión (ECM) se puede observar como la distancia entre las muestras de entrenamiento y los núcleos o clusters que se van generando en la matriz a medida que van siendo almacenados los vectores característicos.

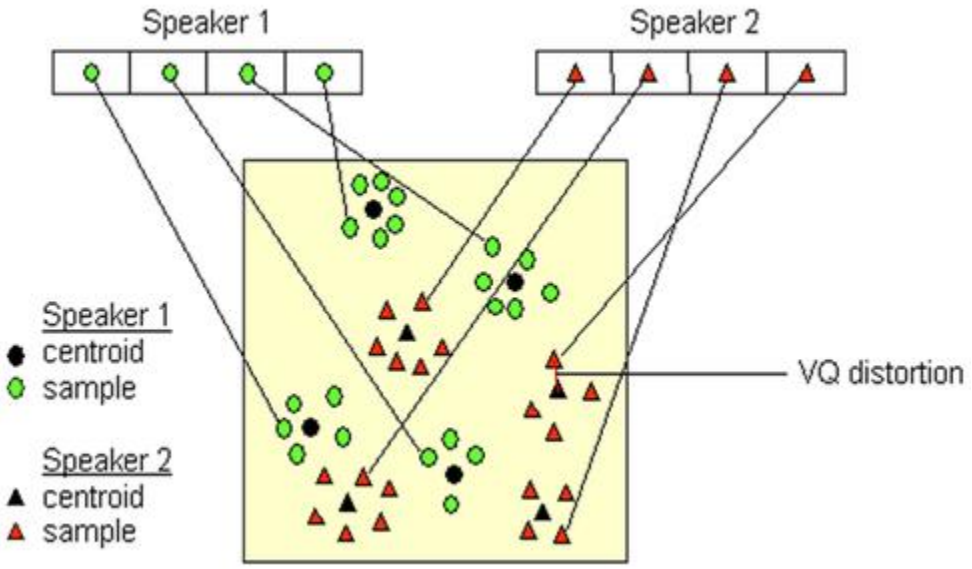


Figura 5.Creación de libro de códigos para la cuantización vectorial.

Fuente: DO, Min .*An Automatic Speaker Recognition System.*[En línea]. DSP Mini Project. Universidad de Illinois. USA. Disponible en la Web: "http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/"

El sistema de cuantización vectorial es comúnmente utilizado en sistemas de reconocimiento no supervisado, puesto que la generación del libro de códigos permite extrapolar patrones intrínsecos en la base de datos sin la necesidad de una matriz de objetivos.

2.1.3.3. REDES NEURONALES ARTIFICIALES (ANN)

Las redes neuronales artificiales son los sistemas de organización, clasificación y reconocimiento de datos más utilizados en los desarrollos académicos y profesionales que involucran cantidades masivas de información vinculadas a parámetros (o clases) no intuitivas. Las aplicaciones de las redes neuronales

artificiales varían desde clasificación de clientes en entidades financieras[6] hasta la interpretación de sintomatologías en pacientes para predecir diagnósticos[7].
El principio de funcionamiento de una red neuronal artificial parte desde su unidad más básica: el perceptrón.

El perceptrón puede ser definido como una unidad sistémica capaz de procesar una entrada de datos, y mediante una discriminación lineal de la entrada, clasificar dichas entradas en un grupo, clase o dimensión, ya sea definido por el programador en la matriz de objetivos o desde un conjunto de núcleos en el caso del aprendizaje no supervisado.

El diagrama general de un perceptrón es el siguiente:

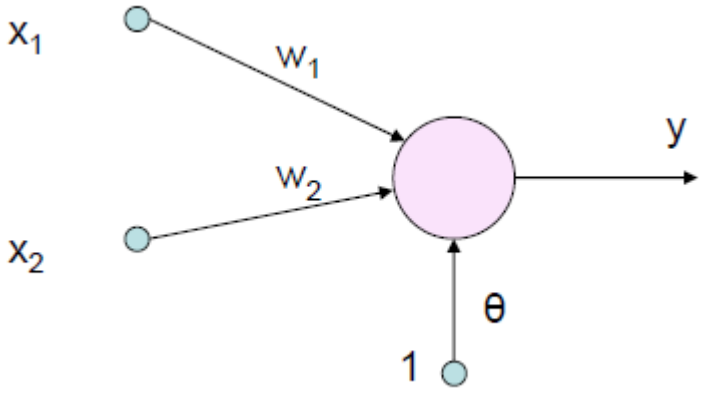


Figura 6.Modelo básico de un perceptrón.

Fuente: Grupo de Investigación de Redes Neuronales Artificiales. *Perceptrón.* [En línea]. Universidad Carlos III. España. Disponible en la Web: "http://www.lab.inf.uc3m.es/~a0080630/redes-de-neuronas/perceptron-simple.html".

Las partes principales correspondientes a un perceptrón son:

- N variables de entrada $[X_1, X_n]$ que serán procesadas por el perceptrón.
- N pesos de entrada $[w_1, w_n]$ para cada una de las entradas al perceptrón.
- 1 variable de ajuste denominada "bias" , simbolizada como "b" o "ɵ".
- una salida $[y]$ determinada por el procesamiento del perceptrón.

---

[6] LI, Eldon. *Artificial neural networks and their business applications.* Paper, NationalChungChengUniversity.China.1994.

[7]DELGADO, Alberto. *Aplicación de las Redes Neuronales en Medicina.* Paper, Universidad Nacional de Colombia.Colombia.1999.

La función de transferencia del perceptrón está dada por la ecuación:

$$\overrightarrow{\boldsymbol{Y}}(\overrightarrow{\boldsymbol{X}}) = \overrightarrow{\boldsymbol{X}} * \overrightarrow{\boldsymbol{w}} + \boldsymbol{b} \qquad \text{Fórmula 1}$$

La función de transferencia del perceptrón implica la linealización de los parámetros de entrada, de esta forma se realiza la clasificación o discriminación de parámetros hacia el vector de salida $\vec{Y}$.

Para entender el funcionamiento de una red neuronal más avanzada, se toma el ejemplo del comportamiento de una compuerta lógica XOR, se quiere implementar una red neuronal artificial que reciba dos entradas [X,Y] y mediante el ajuste de los pesos y las bias se obtenga la salida [Z] correspondiente al resultado de la operación $Z = X(+)Y$.

X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	0

Tabla 1. Tabla de verdad de la compuerta XOR

Por lo tanto se sabe que el rango de las dos entradas a la red neuronal será [0 1], al igual que la salida.

El siguiente paso para la generación de la red, en el caso de un aprendizaje supervisado, es la identificación y creación de las matrices de entrada y salida esperada que serán utilizados como parámetros de entrenamiento para la red.

La matriz de entrada es el conjunto de valores y datos completos o parciales de las entradas:

0	1	0	1
0	0	1	1

Tabla 2. Matriz de entrenamiento de la red neuronal

La matriz o vector de salida esperada para esta red será entonces:

0	1	1	0

Tabla 3. Salida esperada de la red neuronal

La estructura de una red neuronal básica para la emulación de una compuerta básica se representa en la siguiente imagen:

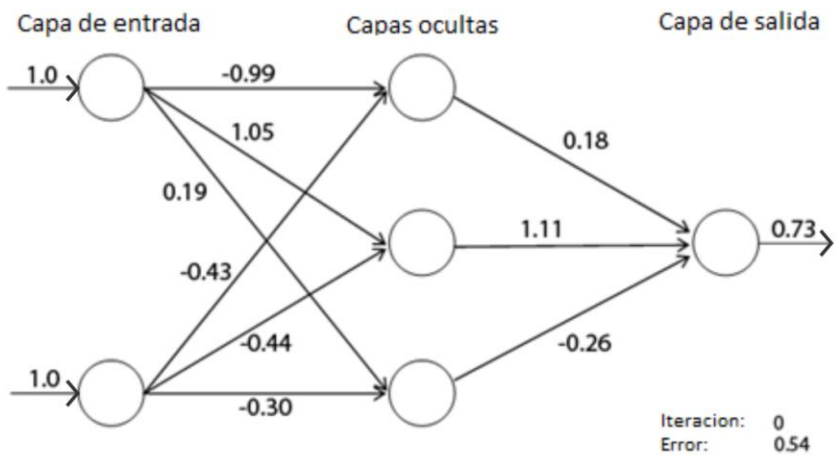


Figura 7. Ejemplo de red neuronal feedforward antes del entrenamiento.

FUENTE: SERGIU, Ciumac. *Feedforward Artificial Neural Network.*[En línea]. Universidad de Moldavia. Disponible en la Web: "Http://www.codeproject.com/Articles/175777/Financial-predictor-via-neural-network#conclusion"

La gráfica representa la primera iteración del entrenamiento. La red se compone de tres capas, las cuales son bloques de perceptrones independientes de la red. La capa de entrada contiene dos perceptrones (uno por cada entrada) con pesos constantes de valor 1. La capa oculta contiene tres perceptrones, la cantidad de perceptrones para las capas ocultas pueden variar dependiendo de la necesidad de procesamiento.
Para el caso de la primera iteración los pesos de entrada de la capa oculta se seleccionan aleatoriamente y la salida de esta capa es la entrada a la capa final de salida, una vez en ella se realiza el procesamiento por el perceptrón de decisión y entrega un valor entre 0 y 1 correspondiente al

procesamiento de las capas anteriores. Para el caso de la primera iteración, al seleccionar aleatoriamente los valores de los pesos de entrada para la capa oculta y la capa de salida, se va a obtener un valor aleatorio de salida, este valor será comparado con el valor de salida esperado (0 para el caso de entradas X=1 Y=1) y se calculará una medida estadística de error (usualmente el error cuadrático medio) para revisar la desviación entre el valor de salida calculado y la salida esperada. Una vez establecido el valor de desviación se compara con un valor de error mínimo admisible, y en caso de que el error calculado sea mayor al error mínimo, se realizará un reajuste a los pesos de las capas ocultas y final, para compensar el porcentaje de error y acercarse más a la salida esperada.

El siguiente gráfico muestra los valores finales de los pesos una vez alcanzado un valor de salida de la red neuronal cuyo error cuadrático medio es igual o inferior al mínimo establecido:

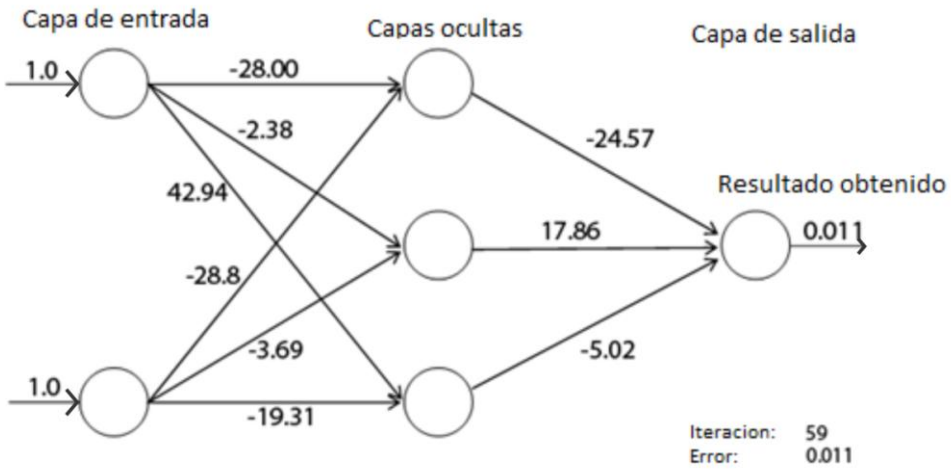


Figura 8.Ejemplo de red neuronal feedforward después del entrenamiento.

FUENTE: SERGIU, Ciumac. *Feedforward Artificial Neural Network.*[En línea]. Universidad de Moldavia. Disponible en la Web: "Http://www.codeproject.com/Articles/175777/Financial-predictor-via-neural-network#conclusion"

Cincuenta y nueve iteraciones de reajuste de pesos fueron necesarias en este ejemplo para alcanzar un valor de salida de 0.011, el error con respecto al valor esperado es, de igual forma, 0.011, para este ejemplo fue considerado suficiente este valor de error para terminar el proceso de entrenamiento y asumir las variables de peso w, fijas.

Este tipo de red neuronal artificial se denomina "Red neuronal Feedforward de retropropagación" y es utilizada como media estándar para el reconocimiento de patrones en bases de información. Su nombre se deriva del proceso iterativo "a la inversa" en el que el reajuste de pesos se lleva a cabo desde la última capa hacia atrás. Para este tipo de red en específico

las bias siempre toman valor de 1 en todas las capas, en otras implementaciones de redes neuronales, las bias pueden entrenarse para identificación de núcleos.
Una vez finalizado el entrenamiento (para el ejemplo de la compuerta XOR el ajuste de pesos se realiza utilizando la información de toda la matriz de entrada) los pesos w se asumen fijos para las etapas de validación y pruebas, en las cuales se introducen matrices de entrada distintas a la matriz de entrenamiento para comprobar los valores de salida obtenidos a partir del procesamiento de la red.

Es importante resaltar la diferencia entre este tipo específico de red neuronal (denominada red "Feedforward de retropropagación") en la cual el ajuste de los pesos es dependiente del vector de salidas esperadas para el sistema, este es un tipo de aprendizaje sistémico denominado "aprendizaje supervisado". Otros tipos de redes neuronales (como las redes de función radial y las redes de recurrencia) son utilizadas para la clasificación de datos y procesamiento secuencial de información, en los que no se tiene una salida especifica deseada, y por lo tanto el entrenamiento de estas redes se lleva a cabo para descubrir patrones inherentes en los datos (aprendizaje no supervisado).

2.1.3.4. REDES NEURONALES ARTIFICIALES MULTICLASE

En el ejemplo de implementación de una red neuronal para la simulación de una compuerta XOR, la salida final del sistema era un valor entre 0 y 1, correspondiente a una sola salida. Estas redes binarias tienen las mismas cualidades de eficiencia que un algoritmo SVN, sin embargo, en ocasiones es necesario realizar un reconocimiento de patrones que involucre mas de una salida, es en esta caso en el que se suele utilizar una red neuronal multiclase.

El principio de implementación de la red multiclase es similar al de las redes binarias, la capa de salida en este caso tendrá una cantidad de perceptrones igual a la cantidad de salidas esperadas, debido a esto, en ocasiones es necesario implementar este tipo de redes con una cantidad mayor de capas ocultas para el procesamiento y linealización. Como se ve en la figura 7, cada una de las salidas de los perceptrones de una capa son transmitidas a cada una de los perceptrones de la capa siguiente, modificadas por los pesos correspondientes a la capa de entrada. es por eso que es necesario más procesamiento computacional para el entrenamiento de una red neuronal multiclase, ya que en cada una de las capas, se está realizando una linealización y correlación entre la salida de cada uno de los perceptrones.

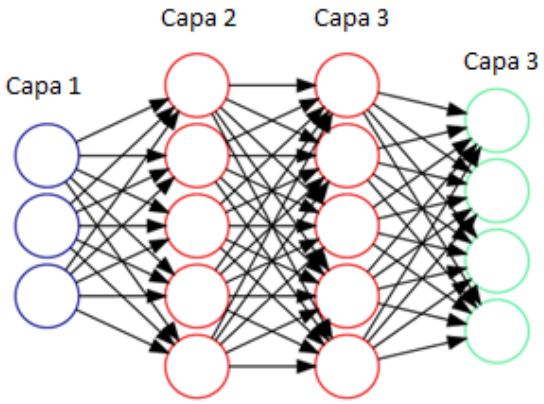


Figura 9.Estructura de una red neuronal multicapa.

La gráfica anterior muestra la configuración de una red neuronal multiclase con tres entradas, cuatro salidas y dos capas ocultas de cinco perceptrones.

2.1.4. APRENDIZAJE DE ALGORITMOS DE RECONOCIMIENTO

Adicionalmente a la forma y estructura de los algoritmos de reconocimiento, existen funciones adicionales a las funciones de transferencia para optimizar el proceso de aprendizaje.

Tomando como ejemplo el caso del perceptrón, a continuación se muestra la configuración de un sistema de reconocimiento con implementación de función de aprendizaje:

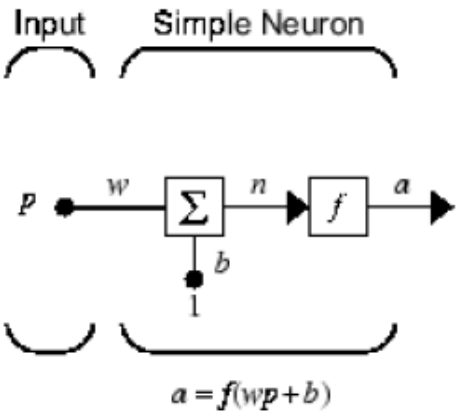


Figura 10. Diagrama de bloques de un perceptrón con función de entrenamiento.

FUENTE: HUDSON, Martin. *Neural Network Toolbox User's Guide.*Reference.2011

Adicionalmente al procesamiento de linealización de las entradas p de la red, (w*p+b), la señal es procesada por una función de transferencia de aprendizaje.

A continuación se mencionan dos funciones de transferencia de aprendizaje, utilizadas usualmente en el reconocimiento de patrones:

- Función de transferencia lineal: Los perceptrones entrenados con esta función de transferencia son usualmente utilizados en la capa de salida, de redes multicapa implementadas como aproximadores de datos, la entrada a esta función de transferencia estará procesada por la siguiente forma:

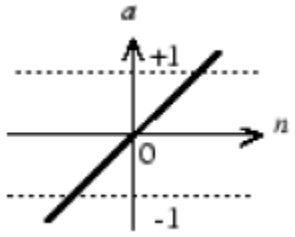


Figura 11.Función de transferencia lineal.

FUENTE: HUDSON, Martin. *Neural Network Toolbox User's Guide.*Reference.2011

- Función de transferencia logarítmica sigmoide: Esta función de transferencia toma entradas de cualquier rango entre más y menos infinito, acotando sus valores a un rango de procesamiento entre 0 y 1:

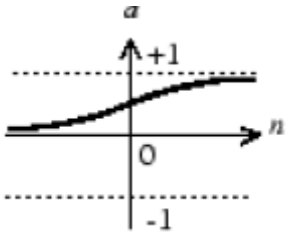


Figura 12: Función de transferencia logarítmica sigmoide.

FUENTE: HUDSON, Martin. *Neural Network Toolbox User's Guide.*Reference.2011.

Esta función de transferencia es utilizada usualmente en capas ocultas ya que, debido a su dominio y continuidad, permite la derivabilidad de los datos procesados por ella.

2.1.5. VALIDACION DE ALGORITMOS DE RECONOCIMIENTO

Uno de los elementos estadísticos más utilizados para comprobar la eficiencia de un algoritmo de reconocimiento es mediante la herramienta de visualización llamada Matriz de Confusión.

La matriz de confusión, propuesta por Kohavi y Provost en 1998, contiene información sobre los índices de clasificación realizado por un sistema de reconocimiento. Para poder realizar esta visualización, es necesario considerar:

- El número de predicciones positivas (correctas) sobre una variable negativa(incorrecta) (w).
- El número de predicciones negativas (incorrectas) sobre una variable positiva(correcta)(x).
- El número de predicciones negativas (incorrectas) sobre una variable negativa(incorrecta)(y).
- El número de predicciones positivas (correctas) sobre una variable positivas (correcta)(z).

Una vez obtenidos estos valores, la matriz de confusión se construye:

		Valores Esperados	
		Negativos	Positivos
Valores Reales	Negativos	w	x
	Positivos	y	z

Tabla 4. Presentación de una matriz de confusión

Se pueden calcular parámetros de eficiencia sobre la matriz de confusión para validar su funcionamiento:

- La precisión de la red neuronal, definida como la proporción de casos positivos que fueron reconocidos correctamente, se calcula a partir de la siguiente expresión:

$$\boldsymbol{P} = \frac{\boldsymbol{z}}{\boldsymbol{x}+\boldsymbol{z}}$$ Fórmula 2

- La exactitud de la red neuronal, definida como la proporción del total de número de predicciones que fueron detectadas correctamente, está definida como:

$$\boldsymbol{E = \frac{W+Z}{W+X+Y+Z}}$$ Fórmula 3

- El error cuadrático medio es calculado para la matriz de confusión a partir de la ecuación

$$\boldsymbol{ECM = \frac{1}{N}\sum_{i=1}^{N}(P_i - R_i)^2}$$ Fórmula 4

Donde N es el número de muestras, P es el vector de valores esperados y R el vector de valores reales.

Analizando los porcentajes de precisión, exactitud y el error cuadrático medio de un algoritmo de reconocimiento, se pueden establecer valores mínimos de funcionamiento para las aplicaciones que se deseen implementar con esta clase de sistemas.

2.1.6. PARÁMETROS DE ANÁLISIS

Dentro de los diferentes parámetros objetivos utilizados generalmente para el reconocimiento del habla, se hacen referencia a parámetros extraíbles del espectro temporal y energético de la señal de audio.

2.1.6.1. PARÁMETROS TEMPORALES

En cuanto a los parámetros temporales, se deben mencionar:

- ***Duración:*** Este tipo de parámetro ofrece propiedades temporales sobre segmentos vocálicos y no vocálicos del habla, dichos parámetros operan directamente sobre la señal en el tiempo.

- ***Energía:*** La energía de una señal x acotada por una función ventana de N muestras está dada por:

$$\boldsymbol{En = \sum_{n=1}^{N} x(n).x'(n)}$$ Fórmula 5

donde x' corresponde a la matriz/vector de entrada conjugado.

- ***Tasa de Cruces por Cero (ZCR):*** Este parámetro considera cuantas veces la señal cambia de signo en el tiempo:

$$\boldsymbol{ZCR = \frac{1}{2} * \sum_{n=1}^{N} |sgn(x_n) - sgn(x_{n+1})|}$$ Fórmula 6

- ***Afinación:*** Siendo este uno de los parámetros más utilizados para el reconocimiento del habla, la frecuencia fundamental F0 variante en el tiempo puede ser identificada mediante una diversa cantidad de algoritmos y métodos (autocorrelación, RAPT, etc.).

- ***Formantes:*** Estos parámetros capturan información temporal sobre la posición y la envolvente de las formantes, son utilizados en el reconocimiento automático del habla por su velocidad de procesamiento.

Los parámetros espectrales energéticos han sido utilizados debido a su eficiencia superior con respecto a los parámetros temporales, ya que estos últimos, en el caso de la detección emocional, tienen una variabilidad y vulnerabilidad ante los parámetros de error no evadibles, y por tanto el nivel de procesamiento y la cantidad de muestras necesarias para el entrenamiento de un algoritmo de reconocimiento son demasiado extensas para un desarrollo específico.

2.1.6.2. PARÁMETROS ESPECTRALES

- **Mel-Frequency Cepstrum Coefficients (MFCC)**

Estos coeficientes resultan de la transformación al espacio cepstral, para así obtener información sobre la envolvente espectral de la señal en instantes de tiempo determinados. La transformación al dominio cepstral se puede obtener aplicando la transformada de Fourier al logaritmo en base diez de la señal en el tiempo, de esta manera se realiza una caracterización de la señal en frecuencia, independizando la componente espectral de variación lenta (contribución del filtro) de la rápida (contribución de la fuente).

$$\boldsymbol{powerCepstrum} = \left|\boldsymbol{FT}\{\mathbf{log}(|\boldsymbol{FT}\{\boldsymbol{x}(\boldsymbol{n})\}|^{\mathbf{2}})\}\right|^{\mathbf{2}} \qquad \text{Fórmula 7}$$

La implementación de la función para extraer los coeficientes cepstrales de la señal de audio se puede representar en un diagrama de bloques como el siguiente:

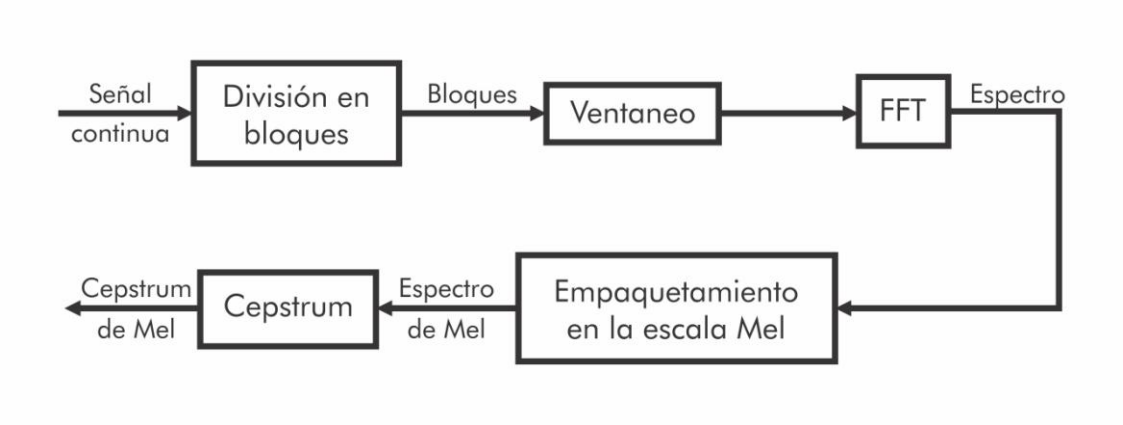


Figura 13. Diagrama de bloques para la extracción de los MFCC

En donde cada uno de los bloques realiza una función específica en el proceso de extracción de características:

- División en Bloques: La señal de audio es almacenada en bloques de N muestras con un solapamiento de M muestras, esto se lleva a cabo para poder analizar el comportamiento espectral de los coeficientes en intervalos de tiempo definidos.
- Ventaneo: A cada bloque se le aplica una función de ventaneo para suavizar los valores en los extremos de cada vector, usualmente se emplea la ventana Hamming presentada a continuación.

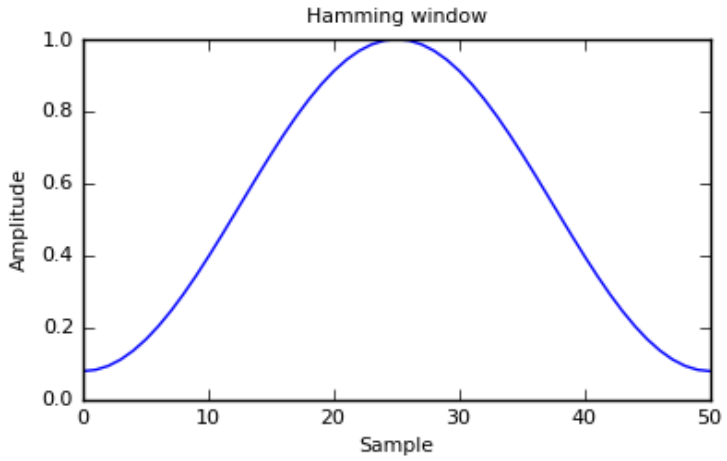


Figura 14. Forma de onda de una ventana Hamming

- FFT: La transformada de cada bloque ventaneado al dominio frecuencial es necesaria para el cálculo de los valores espectrales de las envolventes de la siguiente etapa.

- Empaquetamiento en escala Mel: para poder realizar la extracción de parámetros cepstrales, es necesario procesar los bloques espectrales de la voz para simular el proceso de escucha humano. Como se mencionó en **[2.1.1],** El proceso de producción sonora del ser humano puede ser modelado mediante un sistema de filtros variables, de igual forma el proceso de escucha humana tiene un comportamiento similar y por tanto, tiene un carácter no lineal a la percepción frecuencial. Es por esto que se debe realizar un ajuste de los valores de frecuencia en Hz a una afinación subjetiva dependiente de la forma de escucha, este ajuste se realiza a partir de la escala frecuencial Mel, la cual realiza un espaciamiento frecuencial de forma lineal para frecuencias inferiores a 1000Hz, y logarítmica para frecuencias superiores. Este espectro subjetivo suele ser simulado a partir de un banco de filtros triangulares pasa banda, en donde la frecuencia fundamental para cada filtro esta espaciada utilizando el escalamiento Mel.

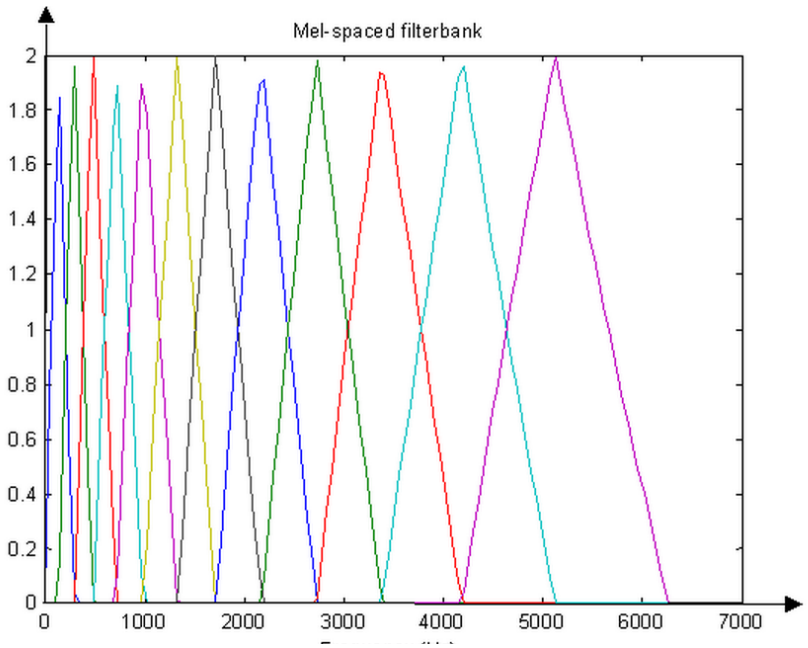


Figura 15. Banco de filtros triangulares en escala Mel

FUENTE: DO, Min *.An Automatic Speaker Recognition System.*[En línea]. DSP Mini Project. Universidad de Illinois. USA. Disponible en la Web: "http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/"

Una vez generado el banco, cada bloque es multiplicado por cada filtro triangular, para así modificar los valores de amplitud y de contribución energética en cada uno de los bloques.

- Cepstrum: En esta última etapa, se calculan los valores de energía media de la señal temporal para cada uno de los filtros de cada bloque, la

resultante son K valores energéticos, donde K es el número de filtros. Finalmente, a la matriz de parámetros se le aplica la transformada discreta del coseno, de esta forma, los K parámetros resultantes corresponden a los coeficientes cepstrales, los valores máximos de la envolvente espectral escalada en frecuencia Mel. Debido a que se desea aplicar la transformada discreta del coseno a los valores energéticos de cada uno de los filtros, para así obtener una representación temporal de las envolventes espectrales, el método de la transformada se calcula a partir de la siguiente ecuación:

$$\widetilde{c}_n = \sum_{k=1}^{K} (\log \widetilde{S}_k) \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], \qquad n = 0,1,\ldots,\ K\text{-}1$$

Fórmula 8

Donde $\tilde{c}_n$ corresponde al enésimo coeficiente cepstral en la escala mel , $\tilde{S}_k$ es el coeficiente correspondiente al k-ésimo filtro triangular y K la cantidad de filtros aplicados.

- Parámetros de variación: Coeficientes Delta y DeltaDelta

  Una serie de parámetros, adicionales a la extracción de los MFCC, son los llamados coeficientes Delta y coeficientes DeltaDelta, también llamados coeficientes de Velocidad y Aceleración. La fórmula de cálculo de estos parámetros es:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}$$

Fórmula 9

Donde d_t corresponde al coeficiente delta en un tiempo t, N es el número de bloques, y c_{t+n} y c_{t-n} corresponden a los coeficientes cepstrales en las muestras t+n y t-n respectivamente. Para el cálculo de los coeficientes DeltaDelta, la ecuación es la misma, solo que la resta dentro de la sumat oria se realiza sobre los coeficientes Delta y no sobre los MFCC.

Esta fórmula de regresión es análoga a realizar la derivada de los parámetros MFCC, por lo tanto los coeficientes Delta contienen información sobre qué tanto varían los coeficientes en el tiempo, la rapidez de esta variación, es decir la aceleración, está descrita en los coeficientes DeltaDelta.

Los parámetros descritos pueden ser definidos como “parámetros intermedios” en el proceso de identificación emocional, puesto que para poder hacer un análisis de estos parámetros, es necesario medir y determinar variables estadísticas de más bajo nivel que puedan describir su comportamiento: Media, máximos, mínimos, y la diferencia entre estos, la varianza, la mediana, entre otros.
De igual manera estos parámetros intermedios son utilizados para describir los parámetros objetivos directamente relacionados con los estados emocionales: Activación, Potencia, y Valencia.

MARCO NORMATIVO

El estándar ISO/IEC MPEG-4[8]fue desarrollado por el Moving Pictures Experts Group para regular los procesos relacionados a la codificación de información para difusión en medios digitales (Radio, Televisión, Internet, etc.).

En su apartado 11.5, denominado "Objetos Sintéticos", se establecen los estándares para la síntesis de emociones en avatares virtuales, definiendo como norma siete emociones discretas: Ira, disgusto, miedo, alegría, tristeza, sorpresa y neutralidad.
Para el desarrollo de este proyecto, se seleccionaron las emociones de felicidad, tristeza y alegría no solo por sus marcadas diferencias en el espectro dimensional de activación, potencia y valencia **[2.1.2.]**. Teniendo en cuenta las especificaciones emocionales recomendadas por la norma, la implementación del algoritmo desarrollado en el proyecto cumple los estándares internacionales para ser utilizado en las aplicaciones propuestas en **[1.5.3.]**.

[8]-Moving Pictures Experts Group.*Overview of the MPEG-4 Standard*. Norma. 2002.

3. METODOLOGÍA

3.1.ENFOQUE DE LA INVESTIGACIÓN

Puesto que, dentro del desarrollo del proyecto se encuentra el desarrollo de un producto tecnológico a modo de prueba empírica de teorías sustentadas en modelos matemáticos y algoritmos mencionados en el marco teórico, el enfoque de la investigación es Empírico-analítico, se propone una estructura de bloques definida claramente por los objetivos específicos del proyecto, y presentada en la siguiente imagen:

Figura 16. Algoritmo propuesto para el desarrollo del proyecto.

3.2.LÍNEA DE INVESTIGACIÓN DE LA UNIVERSIDAD DE SAN BUENAVENTURA/ SUB-LÍNEA DE INVESTIGACIÓN/CAMPO DE INVESTIGACIÓN

Este proyecto tiene un componente elevado de procesamiento de señales, por lo tanto esa sería su sublínea de facultad, sin embargo el campo de investigación en el que se suscribe el proyecto dentro de la carrera es en el área de Acústica, puesto que los conceptos requeridos para realizar el reconocimiento, nacen desde la Psicoacústica y la respuesta del oído humano.
La línea institucional universitaria es, entonces, “Tecnologías Actuales y Sociedad”.

3.3.TÉCNICAS DE RECOLECCIÓN DE INFORMACIÓN

Para los procesos de entrenamiento y prueba de los algoritmos de reconocimiento que se van a implementar en el desarrollo del proyecto, es necesario contar con bases de archivos de audio para la extracción de sus parámetros.

3.3.1. BASE DE AUDIOS EN ALEMÁN.

La base de archivos de audio "Berlin Emotional Speech Database" (Emo-DB), fue creada por W. Sendlmeier, A. Paeschke, M. Kienast, F. Burkhardt y B. Weiss en la Universidad Técnica de Berlín.

El corpus está conformado por 536 archivos de audio, dentro de sus características específicas cabe mencionar:

- Clasificación emocional: acorde a las emociones discretas de la norma MPEG-4: Ira, disgusto, miedo, alegría, tristeza, sorpresa y neutralidad.
- Los diez locutores seleccionados para la creación del corpus fueron filtrados de una preselección de cuarenta voluntarios:

- Hombre, 31 años
- Mujer, 34 años
- Mujer, 21 años
- Hombre, 32 años
- Hombre, 26 años
- Hombre, 30 años
- Mujer, 32 años
- Mujer, 35 años
- Hombre, 25 años
- Mujer, 31 años

- Diez frases en alemán fueron implementadas con distintas intenciones para el desarrollo del corpus.

- La grabación del corpus se llevó a cabo en la cámara anecóica de la Universidad Técnica de Berlín, se utilizó un micrófono Sennheiser MKH 40 P 48 y una grabadora DAT portátil TASCAM DA-PI.

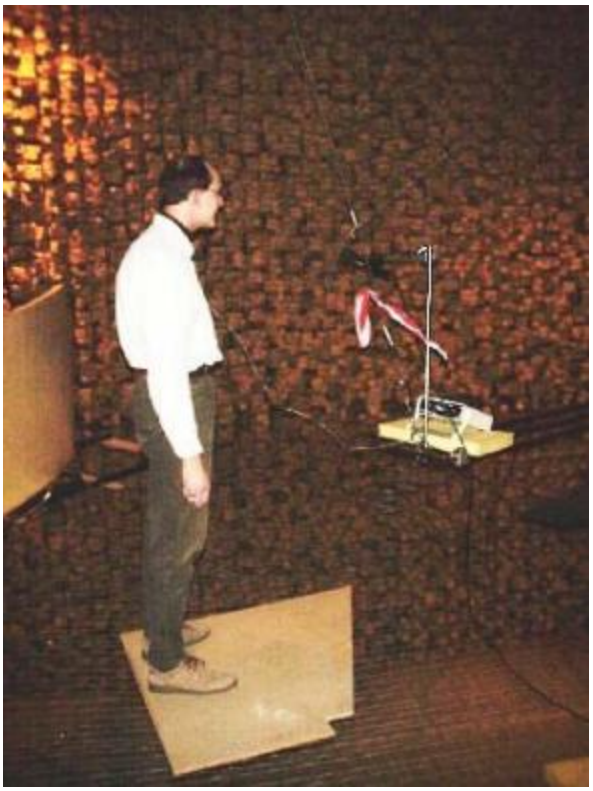


Figura 17. Grabación del corpus Emo-DB

FUENTE: BURKHARDT, PAESCHKE. *A Database of German Emotional Speech.* Paper, Technical University of Berlin. Alemania. 2000

- Para el desarrollo del proyecto, se clasificaron los audios de emociones de felicidad (71), tristeza (62) y enojo (128),para un total de 261 audios.
- Los audios fueron codificados y almacenados con una frecuencia de muestreo de 16000Hz y una resolución de 16 bits.

El corpus Emo-DB ha sido utilizado satisfactoriamente en diferentes proyectos relacionados con el reconocimiento emocional[9], algunas de las cuales se llevaron a cabo en Colombia[10] y sus resultados serán comparados con los resultados del proyecto.

[9] PRADIER, Melanie. *Emotion Recognition from Speech Signals and Perception of Music.* Thesis Degree Paper, Stuttgart University.Germany.2011

[10] RUEDA,E,TORRES,Y. *Identificación de emociones en la voz* .Thesis Degree Paper, Universidad Distrital de Santander.Colombia .2007.

3.3.2. CAPTURA DE LA BASE DE AUDIOS EN ESPAÑOL

Ya que se pretende realizar la comparación de los porcentajes de reconocimiento entre voces foráneas y nativas **[1.4.2]**, es necesaria la captura de señales de audio actuadas en español.

El proceso de captura y generación de la base de audios en español para ser utilizados en el desarrollo del proyecto tiene las siguientes características:

- La grabación se llevó a cabo el día Viernes 18 de Octubre de 2013 en el estudio Digital de la Universidad de San Buenaventura.

- La grabación de los audios fue realizada con los siguientes elementos:

  - Micrófono dinámico SHURE SM-58.
  - Preamplificador Manley VOXBOX..
  - Interfaces AD/DA Apogee ADX 116.
  - Protools 8.0.5 HD.

- Aunque en la creación del corpus Emo-DB, se realizó una preselección de locutores sin tener en cuenta su experiencia en este campo, para el caso de la creación de la base de audios en español se desean comparar los resultados del reconocimiento de las voces de individuos de diferente género, edad, y grado de experticia en la actuación, por lo tanto Los locutores seleccionados para la captura de audio fueron:

  - Juan Daniel Morales Piedrahita: Estudiante de Ingeniería de Sonido, 24 años.
  - Felipe Estrada Gómez: Ingeniero de Sonido, 25 años.
  - Luz Elvira Piedrahita Galeano: Médico General, 50 años.
  - Nathalia Acosta Diagama: Actriz y Bailarina profesional, 26 años.

- Se realizó la captura de dos fragmentos de texto de la escena 4 del cortometraje "El Admitido Nefasto", escrito por Daniela Catalina Acosta Moyano:

Fragmento 1:

"'¡Ay, no te digas mentiras Raquel!, yo sé que jamás olvidaste a Santiago"

Fragmento 2:

"¡Raquel! ¡Yo sé que tú lo harías! ¡Hazlo por mí! ¡Hazlo por Santiago!"

- Los fragmentos fueron seleccionados por su corta duración y facilidad al ser interpretados por los cuatro locutores en las tres emociones necesarias para la investigación.
- Los fragmentos fueron grabados por los cuatro locutores, con las tres intenciones emocionales (felicidad, tristeza, enojo), con tres repeticiones por emoción.
- Los audios fueron editados en Protools, recortando los silencios y aplicando una compuerta de ruido para minimizar el ruido de fondo presente en la grabación.
- Se realizó una normalización de los archivos de audio correspondiente a los niveles establecidos por el corpus alemán Emo-DB (valor RMS = 0dBFS).

El resultado del proceso de captura fue un corpus de 72 audios de entre tres y cuatro segundos, 24 audios por estado emocional.

3.4.HIPÓTESIS

El sistema implementado tendrá el nivel de eficiencia suficiente para comprobar la teoría de los parámetros universales.

3.5.VARIABLES

3.5.1.VARIABLES INDEPENDIENTES

Las variables independientes utilizadas para el análisis son:

- Media
- Máximos
- Mínimos
- Error cuadrático medio

3.5.2. VARIABLES DEPENDIENTES

Las principales variables dependientes de este estudio corresponden a los estados emocionales que se pretenden detectar:

- Felicidad.
- Tristeza.
- Enojo.

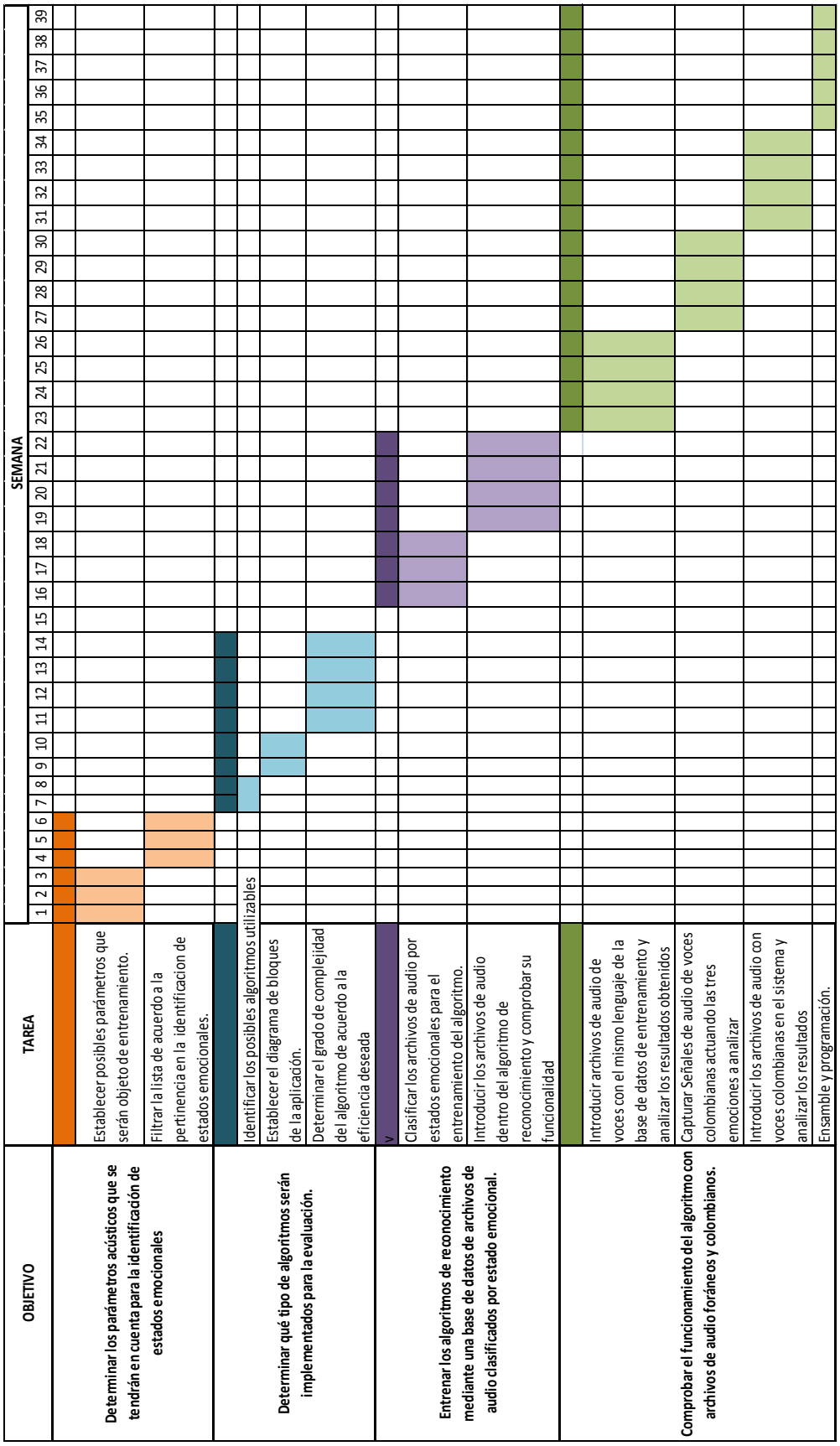
Los parámetros dependientes objetivos utilizados para la detección de los estados emocionales son:

- MFCC.
- Energía media.
- Coeficientes Delta.
- Coeficientes Delta Delta.
- Tasa de Cruces por Cero.

Las variables dependientes propuestas desde la psicología **[2.1.3]** en las se enmarcan los estados emocionales continuos son:

- Activación.
- Potencia.
- Valencia.

4. CRONOGRAMA



5. PRESUPUESTO

PRESUPUESTO						
					APORTE	
	DESCRIPCIÓN	CANTIDAD	VALOR UNITARIO	UNIVERSIDAD	PROPIOS	ADQUIRIDOS
RECURSOS HUMANOS						
Ingeniero	Investigación	140 Hora	$ 25.000		$ 3.500.000	
	Diseño y Programación	250 Hora	$ 35.000		$ 8.750.000	
Actores	Requeridos para captura de voces colombianas	8 Hora	$12.000			$ 96.000
	TOTAL			$ 0	$ 12.250.000	$ 96.000
EQUIPAMIENTO						
Computador	Computador con capacidad y rendimiento suficiente para desarrollo investigativo y programación	1 Unidad	$ 1.200.000		$ 1.200.000	
Impresora Multifuncional		1 Unidad	$ 240.000			$ 240.000
Memoria USB	Almacenamiento y Backup	1 Unidad	$ 40.000		$40.000	
Estudio de Grabacion	Captura de Voces	4 Horas	$ 80.000	$ 320.000		
Matlab (Licencia Universitaria)	Programa de desarrollo	1 Licencia	$ 1.400.000	$ 1.400.000		
	TOTAL			$ 1.720.000	$ 1.240.000	$ 240.000
PAPELERIA						
Resma		2 Unidad	$ 9.900			$ 19.800
Cartucho de Tinta		4 Unidad	$ 75.000			$ 300.000
Internet		9,75 Mes	$ 66.500		$ 648.375	
Papeleria General						$ 100.000
	TOTAL			$ 0	$ 648.375	$ 419.800
CAFETERIA - TRANSPORTE						
Transporte	Transporte de 39 semanas	390 Trayecto	$ 5.000		$ 1.950.000	
	Transporte Actores	8 Trayecto	$ 5.000			$ 40.000
Alimentos	Almuerzos durante 39 semanas	195 Unidad	$ 6.000		$ 1.170.000	
	Refrigerio Actores	8 Unidad	$ 3.500			$ 28.000
	TOTAL			$ 0	$ 3.120.000	$ 68.000
	IMPREVISTOS		**10%**			$ 82.380
	SUBTOTAL			$ 1.720.000	$ 17.258.375	$ 906.180
			GENERAL	$ 19.884.555		
	TOTAL		**REAL**	$ 906.180		

6. DESARROLLO INGENIERIL

Durante la investigación realizada sobre los diferentes algoritmos de reconocimiento existentes para el reconocimiento y la clasificación de patrones, tres sistemas fueron analizados **[2.1.3].** Entre estos tres sistemas de reconocimiento, las redes neuronales artificiales fueron seleccionadas sobre la cuantización vectorial (K-means) y los support vector machines:

- En estudios comparativos de reconocimiento realizados entre las redes neuronales artificiales feedforward y el algoritmo K-means (base de la cuantización vectorial) se encontró que el algoritmo K-means es menos eficiente, y en ciertas ocasiones falla en realizar un reconocimiento de patrones con datos obtenidos de experimentos reales. Ya que este proyecto realizará una extracción de parámetros producto de archivos de audio con voces humanas, es necesario seleccionar un sistema que permita realizar un reconocimiento eficiente de patrones no lineales. La cuantización vectorial es un sistema óptimo para el reconocimiento de patrones de variables lineales, dependientes de teorías matemáticas continuas, sin embargo al no ser este el caso del estudio, las redes neuronales y los sistemas de support vector machine son más útiles para el presente estudio[11].
- Aunque las redes neuronales de retropropagación y los sistemas de support vector machine son ampliamente utilizados para el reconocimiento de patrones, el ajuste de los vectores de soporte requiere un mayor tiempo de procesamiento que el ajuste de los pesos de las redes neuronales.
- El reconocimiento de patrones utilizando redes neuronales artificiales es un proceso estocástico, lo que significa que el procesamiento en su entrenamiento involucra variables determinísticas (número de perceptrones, valor mínimo del error cuadrático medio para el entrenamiento, función de transferencia) y variables aleatorias (los valores de inicialización de los pesos en la primera iteración del entrenamiento). Por otra parte, la configuración y el entrenamiento de un sistema de support vector machine involucra el ajuste de parámetros no intuitivos para su correcto funcionamiento (número de vectores de soporte, el número de hiperplanos, la configuración de los hiperplanos, el valor de margen mínimo, entre otros).[12]Puesto que el centro de estudio de presente proyecto es el estudio tanto de los parámetros extraídos de la voz humana como la selección de un algoritmo con una implementación robusta, confiable y de configuración comprensible, las redes neuronales artificiales ofrecen la versatilidad de poderse re-entrenar, además de ser más rápidas en el procesamiento que los sistemas de support vector machine, esto es importante puesto que la

---

11 HRUSCHKA,Harald, *"Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation"*- Thesis. University of Regensburg. Alemania. 1997.

12 ANTKOWIAK, Michal."Artificial Neural Networks vs. Support Vector Machines for Skind Siseases Recognition". Master's Thesis. Umea University. Suecia.

cantidad de variables de entrenamiento es considerable (entre 3000 y 7000 datos por archivo de audio).

De acuerdo al algoritmo de investigación propuesto para este proyecto **[3.1]**, se realizaron delimitaciones adicionales a los procesos a analizar:

- Dentro de la selección y delimitación de parámetros, el estudio se enfocará a las variables espectrales de las señales de audio, ya que estas son las más utilizadas en el reconocimiento emocional de la voz[13] ,las variables temporales (inflexiones, formantes) no serán consideradas puesto que estas son confiables para estudios en reconocimiento de hablantes, mas no en reconocimiento emocional.
- En la selección de algoritmos de clasificación, la construcción de Redes Neuronales Artificiales permiten enfocar el estudio a la generación de matrices de parámetros previamente calculados. Al seleccionar un apropiado número de capas de procesamiento y una base de datos adecuada, la Red Neuronal Artificial realiza un reconocimiento adecuado para los parámetros acústicos extraídos de la voz.
- La metodología de entrenamiento de las redes será aprendizaje supervisado, ya que al tener el corpus Emo-DB debidamente clasificado por estado emocional, es posible realizar la construcción de la matriz objetivo para el proceso de reconocimiento, de esta manera el entrenamiento de las redes neuronales de retropropagación podrán realizar el ajuste de los pesos entre capas hasta que el error cuadrático medio entre el vector objetivo y la salida obtenido sea igual, o inferior, a un valor determinado en la configuración de la red.

6.1.PRIMERA ITERACIÓN

6.1.1. SELECCIÓN Y EXTRACCIÓN DE PARÁMETROS

Dentro de los diferentes tipos de parámetros temporales y espectrales que pueden ser extraídos de la voz humana, los coeficientes cepstrales juegan un papel fundamental en el reconocimiento de voz en general. Esto debido a múltiples razones, entre ellas sobresalen su facilidad de extracción, sus buenos resultados documentados en diversas investigaciones[14][15][16], su independencia a parámetros

---

[13] ITTICHAICHAREON, Chadawan. *"Speech Recognition using MFCC",* Paper, International Conference on Computer Graphics, Simulation and Modelling. Tailandia. 2012

[14] DSP Mini Project: An Automatic Speaker Recognition System. Universidad de Illinois http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/

[15] SRIVASTAVA, Nidhi. *"Speech Recognition using MFCC and Neural Networks".* Ambúlika Institute of Technology. India

de error como el ruido de fondo, las inflexiones temporales, los sistemas de captura, los canales de transmisión, entre otros. Es por esto que el estudio de los coeficientes MFC serán de gran importancia en el desarrollo del proyecto, sin embargo, y para comparar los resultados del reconocimiento emocional de la voz entre parámetros temporales, en una siguiente iteración del algoritmo investigativo se compararan los resultados de los MFCC con los resultados de reconocimiento utilizando como entrada la tasa de cruces por cero.

Para un primer acercamiento a la selección y análisis de los parámetros acústicos adecuados para el reconocimiento emocional de la voz humana, los coeficientes Cepstrales en escala Frecuencial Mel (MFCC), fueron calculados para la verificación de los picos espectrales de las señales de audio a analizar.

Inicialmente se realizó la extracción de este parámetro aplicando la función "kannumfcc.m", la cual recibe como entradas:

- El número de coeficientes cepstrales a calcular (K).
- La señal de audio.
- La frecuencia de muestreo de la señal de audio.

Esta función realiza todas las operaciones necesarias para el cálculo de los coeficientes cepstrales**[2.1.6.2]** en bloques definidos en intervalos de 10ms, adicionalmente el algoritmo realiza operaciones de preprocesamiento a los bloques , implementando una compuerta de ruido rudimentaria al aproximar valores inferiores a 10^-22 a dicho valor (kannumfcc.m, línea 36) y realiza un filtrado de pre énfasis en altas frecuencias (para así compensar la componente en alta frecuencia suprimida durante el proceso de producción sonora humana)[17].
Dentro del estado del arte investigado, se propone el análisis de los primeros trece coeficientes para el reconocimiento de locutores, puesto que el análisis emocional en las primeras etapas requiere un análisis detallado de la contribución energética para las bandas en escala Mel, se delimitó el análisis para esta primera etapa a los primero 26 coeficientes MFCC, posteriormente se analizará la contribución de los coeficientes al reconocimiento, y se reducirán o aumentarán la cantidad de parámetros calculados de acuerdo a los resultados de la primera iteración.

---

16 PRADIER, Melanie. *"Emotion Recognition from Speech Signals and Perception of Music".* Thesis. Stuttgart University. Germany.2011
17 JANG, Roger. *Audio Signal Processing and Recognition.* Capítulo 12-2.

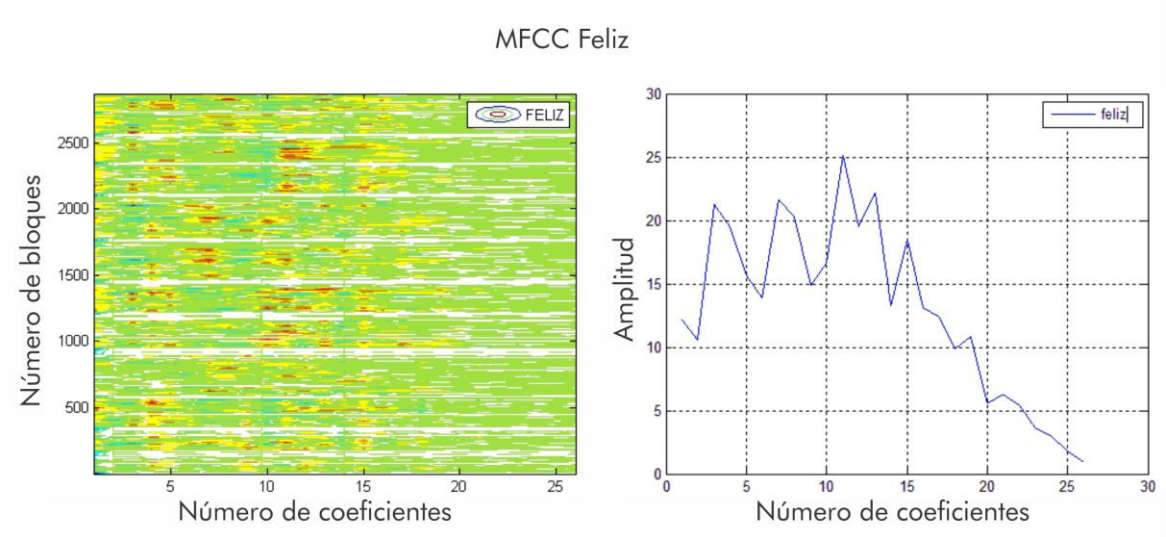


Figura 18.Espectro y valores máximos de MFCC para emoción feliz.

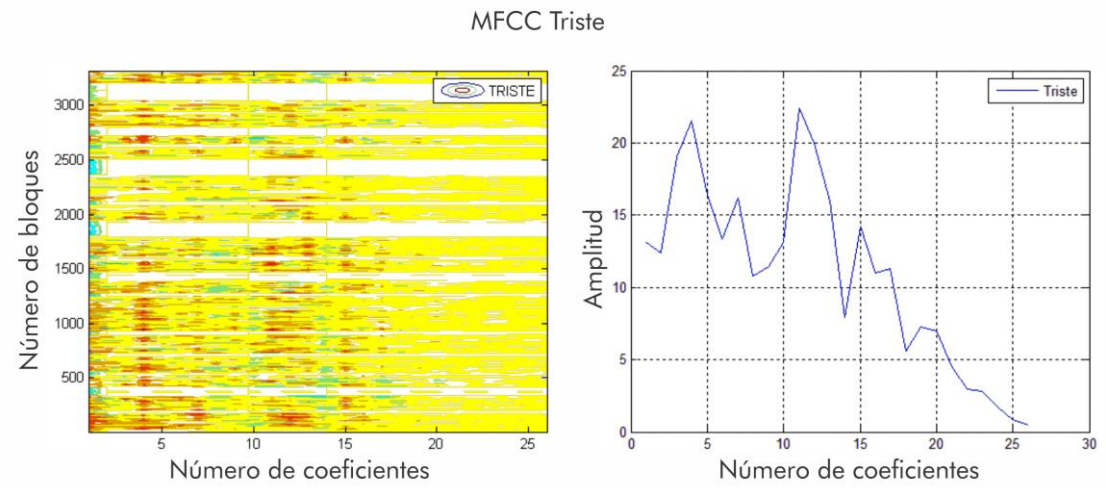


Figura 19. Espectro y valores máximos de MFCC para emoción triste.

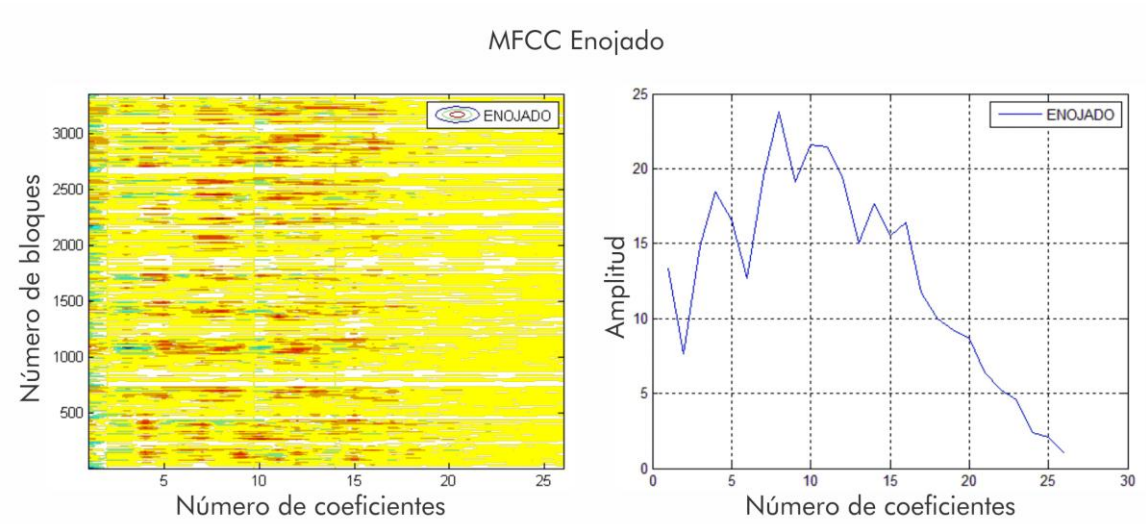


Figura 20. Espectro y valores máximos de MFCC para emoción enojado.

En las imágenes anteriores se puede observar el comportamiento espectral de los estados emocionales utilizando la función Kannumfcc, se procesaron diez audios aleatoriamente de la base de datos Emo-DB para cada una de las emociones objetivo. Las imágenes de la izquierda corresponden a la gráfica de contornos, en donde el eje horizontal es el número de coeficientes cepstrales, el eje vertical el número de bloques, y el color el valor del coeficiente [en un rango entre -25 y 25], en estos gráficos se puede apreciar una diferencia en los valores de intensidad espectral de los parámetros cepstrales de los audios con emoción de felicidad, presentan valores medios más elevados que los audios de enojo y tristeza, los picos de los audios de enojo son más pronunciados y reflejan la diferencia de esta emoción en el dominio de la valencia (que tan positiva o negativa es la emoción) con respecto al enojo y a la tristeza (valencia negativa).

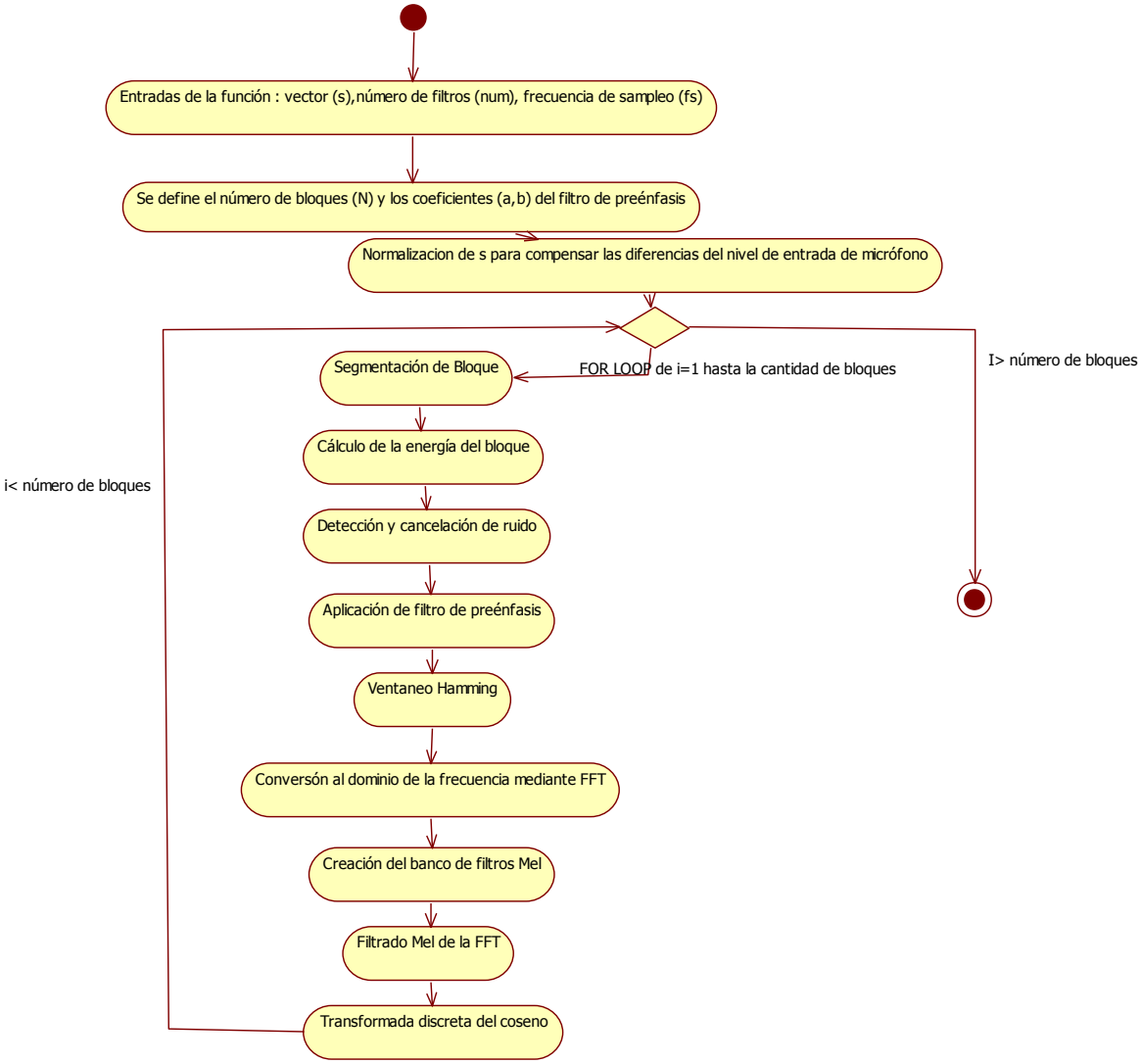


Figura 21. Diagrama de flujo de la función "kannumfcc"

Las gráficas de la derecha corresponden a los valores máximos de los parámetros cepstrales, en los que se puede apreciar la contribución de cada uno de los filtros triangulares en escala de Mel, así, se puede ver un incremento importante en los valores máximos para los coeficientes [5-10] de la emoción enojo y [10-15] para felicidad y tristeza, estos valores están relacionados con la activación (grado de excitación) de la emoción, al estar más bajo el ancho de banda al que pertenece el valor máximo, más cercano están los audios a emociones como la tristeza y el aburrimiento.

Otra representación importante para el análisis corresponde a los valores promedio de los parámetros cepstrales en cada emoción (este promedio fue calculado a partir de los valores de los coeficientes para cada uno de los bloques obtenidos de los audios de entrenamiento:

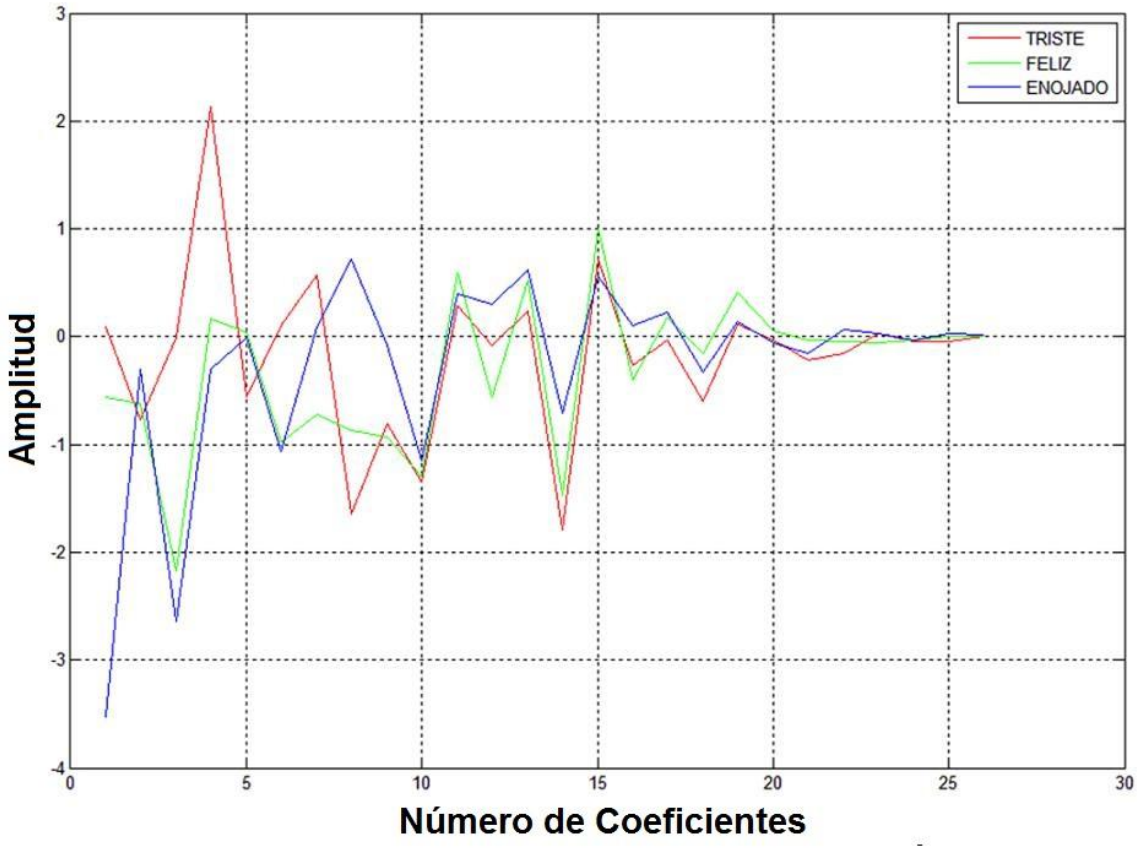


Figura 22. Representación espectral promedio para 26 Coeficientes Cepstrales

Al analizar los valores promedio se observa que la característica de los parámetros para los filtros triangulares de alta frecuencia es común para los tres estados emocionales, específicamente en el valor de amplitud del coeficiente entre los filtros 15 al 20, con una mayor diferenciación en los filtros de baja frecuencia. Se puede observar un nivel energético promedio pico en los

parámetros [1-5] para la tristeza, rango en el cual ocurre un valor mínimo en la felicidad y el enojo, esto es consecuente con la potencia energética de la emoción, en este caso la diferencia entre los picos máximo y mínimo son interpretadas como cuadrantes de potencia diferentes en el modelo de Schlosberg[18], en los cuales la felicidad y el enojo comparten la misma dimensión en potencia energética.

6.1.2. SELECCIÓN DE ALGORITMOS DE CLASIFICACIÓN

Teniendo en cuenta los diferentes tipos de algoritmos de clasificación existentes, y la delimitación inicial, se seleccionó una red neuronal "feedforward" con retropropagación, puesto que sus características de aprendizaje son adecuadas para el reconocimiento de patrones en conjuntos de datos como los parámetros espectrales extraídos de la voz humana[19].

Un primer acercamiento a la generación de la red neuronal necesaria para el reconocimiento emocional, es considerar cada emoción como variables independientes entre sí, de esta forma, se propone la creación de tres redes neuronales, a cada una se le ingresarán los parámetros espectrales extraídos de los archivos de audio en la etapa anterior, la salida de cada red será entonces un valor que se acercará a la unidad cuando exista un reconocimiento de la emoción determinada.

Las redes neuronales, entonces, deben tener las siguientes propiedades para su funcionamiento:

- 20-26 perceptrones en su capa de entrada.

Al existir una relación directa entre el número de parámetros de entrada a la red y la cantidad de perceptrones de la misma en su capa de entrada, el número de perceptrones será proporcional a la cantidad de coeficientes MFC calculados. Para verificar la influencia de los filtros en escala mel en el reconocimiento, se realizarán pruebas con 20 y 26 filtros.

- 1-3 perceptrones de salida.

En una primera etapa en el desarrollo, se pretende implementar un sistema de redes neuronales binarias, como se observó en **[2.1.3.3].** Este tipo de redes

---

[18] PRADIER, Melanie. *Emotion Recognition from Speech Signals and Perception of Music.*Thesis Degree Paper, Stuttgart University.Germany.2011

[19] CRUZ, Luis, ACEVEDO, Marco. *Aplicación del Reconocimiento de Voz de un Hablante Mediante una Red Neuronal Backpropagation y Coeficientes LPC sobre un Canal Telefónico.*Thesis Degree Paper, Instituto Politécnico Nacional.Mexico. 2008

poseen, en su capa de salida, solo un perceptrón correspondiente a la salida de reconocimiento de una emoción. Posteriormente se implementarán redes neuronales multiclase, en las que mediante solo una base de datos de entrenamiento se obtendrán ponderaciones para los tres estados emocionales, y por tanto poseerán tres perceptrones de salida.

- Función de transferencia logarítmica sigmoide

Este tipo de función de transferencia, implementada a la salida de cada perceptrón, realiza un entrenamiento lento en las primeras iteraciones, y rápido en las últimas, adecuado para el reconocimiento de patrones.

- El rendimiento de la red se medirá a partir del error cuadrático medio (Matriz de Confusión).

Puesto a que mediante una herramienta como la matriz de confusión permite observar y analizar el comportamiento de las redes neuronales artificiales en sus procesos de entrenamiento, validación y prueba, la cantidad de muestras reconocidas correcta e incorrectamente, y el porcentaje de reconocimiento por estado emocional.

- 1-30 capas ocultas.

Las capas ocultas en una red neuronal permiten un ajuste más preciso de los pesos de la red, realizando la linealización de los parámetros de entrada para alcanzar una generalización (reconocimiento de patrones) y no una memorización (sistemas de clasificación). Las redes neuronales binarias son construidas solo con una capa oculta, ya que la linealización de los parámetros debe tender solo a una salida. En el caso de las redes neuronales multicapa este ajuste lineal requiere de más procesamiento computacional, por lo tanto ser verificara la influencia de la cantidad de capas ocultas de las redes multicapa en el reconocimiento, variando su cantidad de entre 10 a 30 capas.

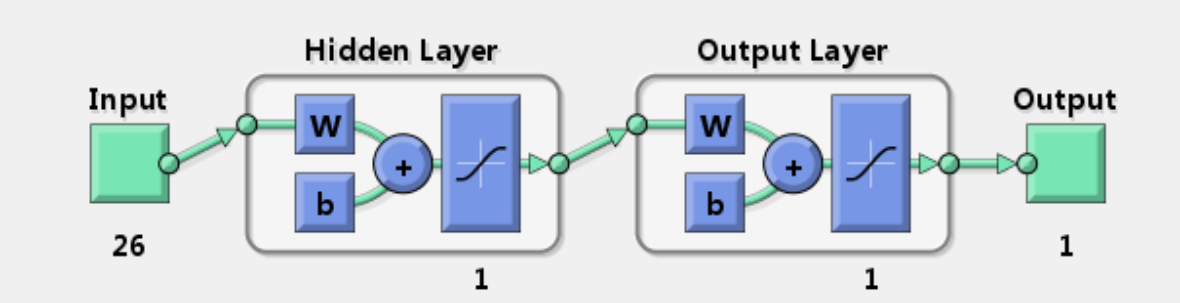


figura23. Diagrama de bloques de la red neuronal

La creación de las redes neuronales se realizará utilizando en principio el Neural Network Toolbox de Matlab:

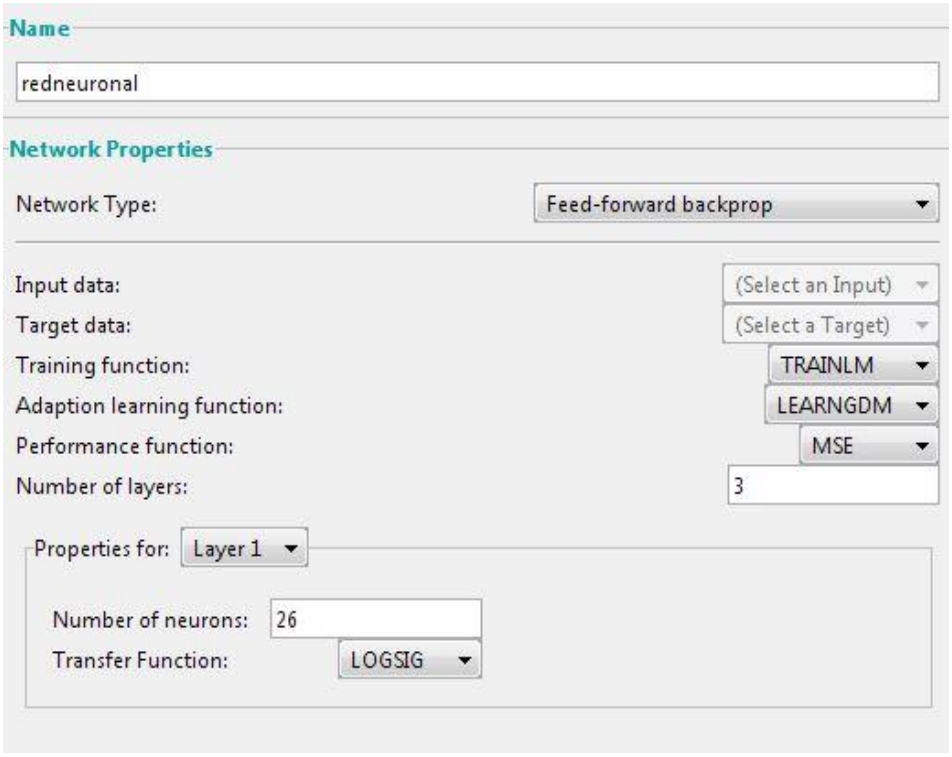


Figura 24. Creación de las redes neuronales binarias

Las redes finalmente fueron exportadas al workspace de Matlab y almacenadas en archivos .mat como "FELIZNET1", "TRISTENET1" y "ENOJADONET1".

6.1.3. ENTRENAMIENTO

El corpus sonoro Emo-DB fue utilizado para el entrenamiento de las redes neuronales. Mediante la implementación de las rutinas "TRAININGCREATE.m" se realizó un proceso iterativo, recorriendo las carpetas en donde estaban alojados los archivos de audio y extrayendo los parámetros MFCC, almacenándolos en una matriz principal de entrenamiento. Simultáneamente se genera el vector de objetivos para que las redes realicen el ajuste a los pesos y bias correspondiente a la retropropagación.

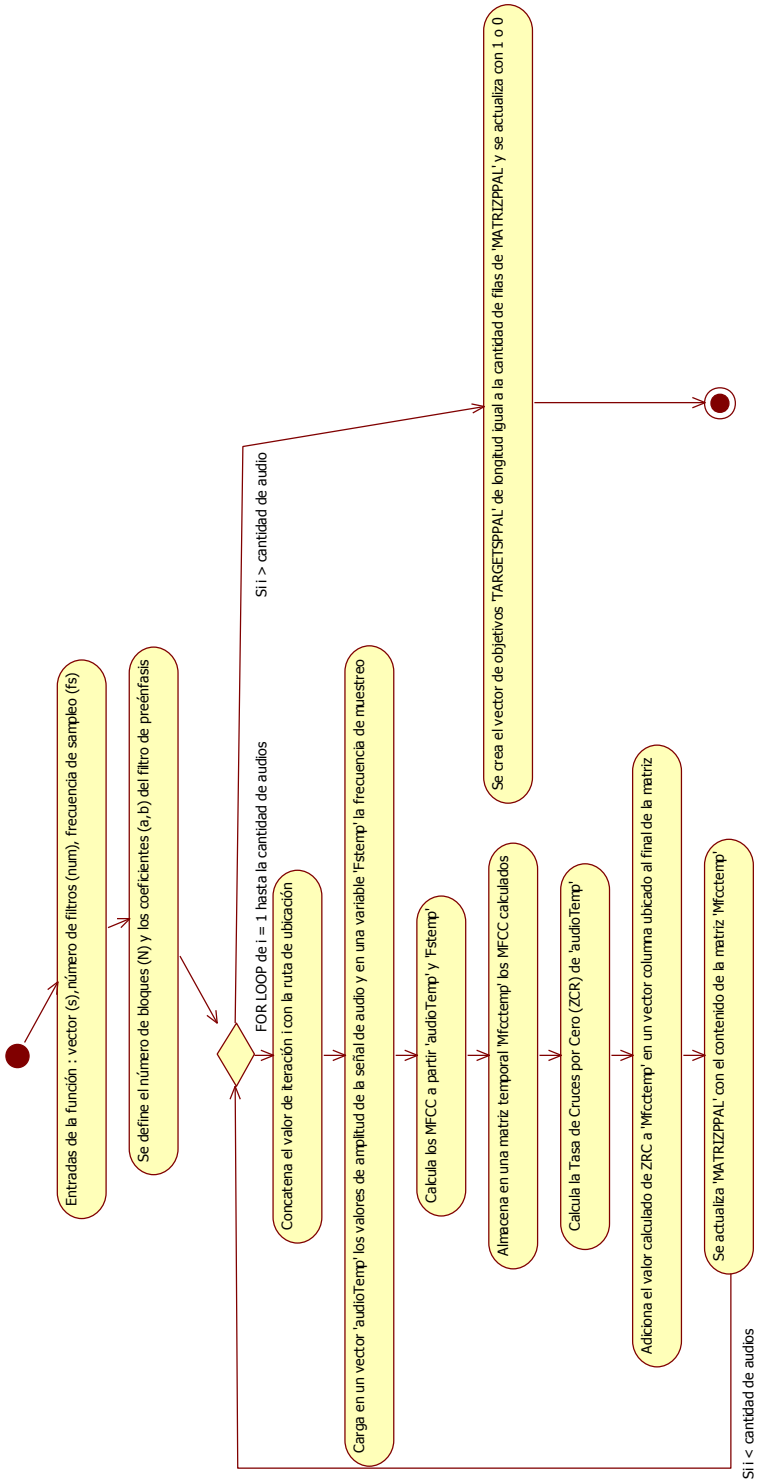


Figura 25. Diagrama de flujo de la función "TRAININGCREATE.m"

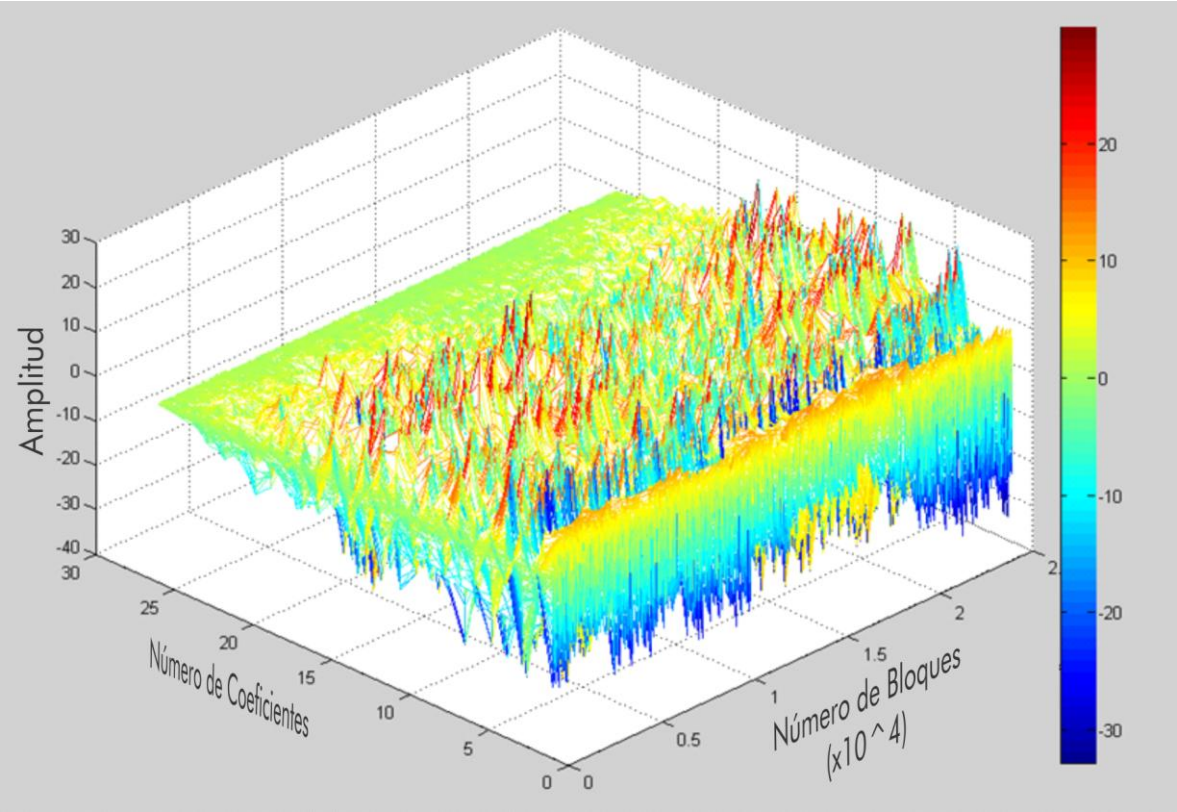


Figura 26.: Espectrograma de la base de entrenamiento

La base de entrenamiento se generó a partir de 46 audios positivos (emoción detectada) y 46 negativos (23 de la segunda emoción y 23 de la tercera), todos los audios fueron seleccionados de forma aleatoria sobre la base Emo-DB clasificada previamente por estado emocional.

Para el entrenamiento de cada una de las redes neuronales implementadas, se generó un vector de L elementos correspondientes a las salidas esperadas de la red, donde L es el número total de bloques de todos los audios a los que se les fueron extraídos los MFCC. Durante el proceso iterativo de la creación de la matriz de entrenamiento, se fue generando el vector de objetivos asignando un valor de 1 a los bloques provenientes de audios positivos
(emoción correcta) y 0 para los bloques de audios negativos(audio incorrecto), este proceso de creación se realiza para los vectores objetivo de las tres subredes de salida binaria especializados en el reconocimiento de cada emoción.

```
for i=1:46

    temp=strcat(direccion, int2str(i), direccion2); %RUTA DEL AUDIO
    [audiotemp Fstemp]=wavread(temp);               %LECTURA DEL AUDIO
    Mfcctemp=kannumfcc(20,audiotemp,Fstemp);        %CÁLCULO DE MFCC

    MATRIZPPAL=[MATRIZPPAL;Mfcctemp];               %CREACION DE MATRIZ DE
                                                    %ENTRENAMIENTO


end

% CREACION TARGET
TARGETSPPAL(1:length(MATRIZPPAL))=ones; %ASIGNACION DE 1 A POSITIVOS
TARGETSPPAL=TARGETSPPAL';
```

Figura 27. Creación de Matrices Objetivo y Vector de entrenamiento "TRAININGCREATE.m"

6.1.4. PRUEBA

Para la prueba de las redes neuronales, se generaron matrices de prueba conteniendo los parámetros MFCC de 20 audios con la emoción que se deseaba detectar y 20 audios seleccionados aleatoriamente de las emociones que no pertenecían al conjunto de positivos. Adicionalmente, se generaron vectores objetivo esperados para poder realizar la matriz de confusión de cada una de las redes. Para la creación de esta matriz fue implementada la función "TESTCREATE.m" con un algoritmo similar a la función de entrenamiento.

La evaluación de rendimiento de las redes neuronales creadas se realiza mediante el análisis de las matrices de confusión producto de las pruebas realizadas sobre cada red. La función "plotconfusion" de Matlab organiza los resultados de falsos positivos, falsos negativos, correctos positivos y correctos negativos necesarios para la construcción de la matriz **[2.1.5.].** Las entradas de esta función son los vectores de salidas obtenidas en la red, y los vectores objetivo construidos en la etapa de entrenamiento. Adicionalmente a los porcentajes de reconocimiento para positivos y negativos, la matriz de confusión presenta la cantidad de muestras reconocidas (correcta e incorrectamente), los porcentajes de reconocimiento globales con respecto a la salida esperada (tercera columna), y los porcentajes de reconocimiento globales con respecto a la salida obtenida (tercera fila).

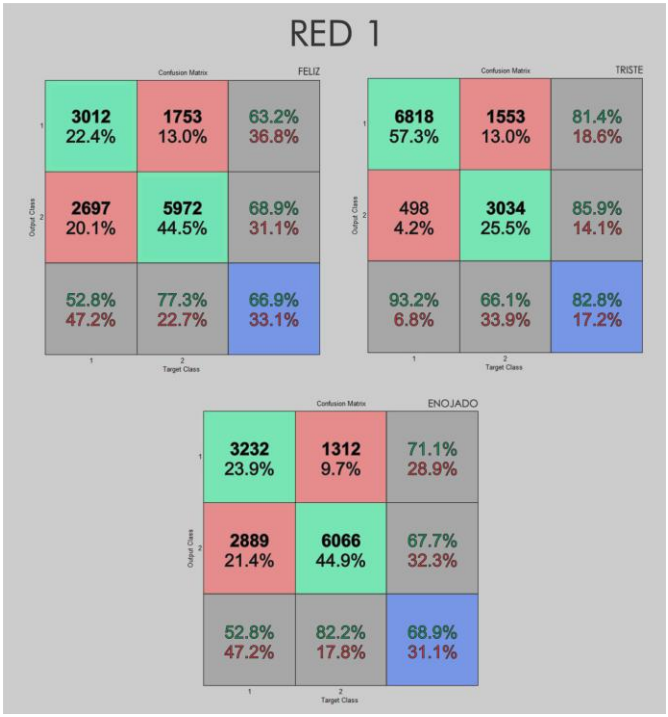


Figura 28. Matrices de confusión (pruebas en RED1).

De acuerdo a las matrices de confusión calculadas para cada una de las redes neuronales, se observa una eficiencia de 66.9% para la red feliznet1, 82.8% para la red tristenet1 y 68.9% para enojadonet1.

Aunque este conjunto de redes presenta índices altos de reconocimiento, se desea optimizar la respuesta del algoritmo lo más posible para posteriormente realizar las pruebas con las voces colombianas, por lo tanto se procede a la generación de nuevas redes neuronales y al análisis de nuevos parámetros extraídos de la voz humana.

6.2. SEGUNDA ITERACIÓN

6.2.1. SELECCIÓN Y EXTRACCIÓN DE PARÁMETROS

Aunque el estudio de este proyecto se centra en el análisis de los parámetros espectrales de la voz humana para el reconocimiento emocional, se desea revisar el efecto de parámetros temporales en el factor de eficiencia de las redes neuronales, es por esto que se pretende implementar una función de cálculo de la tasa de cruces por cero (ZCR), para así determinar la contribución de las altas frecuencias en el reconocimiento.

el valor del tasa de cruces por cero será entonces un 27avo parámetro ingresado a la matriz de entrenamiento y matriz de pruebas de la red neuronal, no se calculará el ZCR sobre los bloques de cálculo de los coeficientes cepstrales puesto que poseen un solapamiento M necesario para el cálculo apropiado de los componentes espectrales, y por tanto, no serían valores fiables de cambio de signo, por lo tanto se le asignará el tasa de cruces por cero de una señal de audio a cada uno de los vectores (bloques) correspondientes a dicha señal.

Se recuerda que la ecuación para calcular el ZCR está dada por:

$$ZCR = 0.5 \sum_{n=1}^{N-1} |sgn(x_n) - sgn(x_{n+1})| \qquad \text{Fórmula 6.}$$

Donde sgn representa el signo de una variable (-1,1), X_n el valor del vector en la muestra n, X_{n+1}, el valor del vector en la muestra n+1 y N el número total de muestras en el vector.

La implementación de la ecuación para el cálculo de la tasa de cruces por cero fue realizada en un archivo .m independiente llamado "ZCRMoral.m":

```
function [a]=ZCRMoral(audiovector)
a=0;
for i=1:length(audiovector)-1
    a=a+abs(sign(audiovector(i))-sign(audiovector(i+1)));
end
a=a/2;

end
```

Figura 29. Implementación de la ecuación de cálculo de la tasa de cruces por cero

En el código se puede observar que la implementación de la Formula 6 se realiza mediante un ciclo que recorre el vector de entrada, censando la cantidad de cambios de signo presente en el vector, asignándole un valor de +1 a cada cambio de signo y un +0 cuando el signo no cambia (de - a - y de + a +).

6.2.2. SELECCIÓN DE ALGORITMOS DE CLASIFICACIÓN

Debido a los resultados preliminares satisfactorios de la red neuronal implementada en **[6.1]**, entrenada con los parámetros MFCC, se propone reducir la cantidad de filtros en escala Mel de 26 a 20, puesto que la contribución de la componente en alta frecuencia para la voz humana es baja comparada a los valores máximos y medios de la componente en frecuencias bajas y medias.

Para analizar los efectos de adicionar una nueva clase temporal dentro de la matriz de entrada, la generación de la red neuronal se llevó a cabo mediante el Neural Pattern Recognition Tool de Matlab, en el cual se dividen las muestras de la matriz de entrada en porcentajes de entrenamiento, validación y prueba, y de esta manera tener un diagnóstico de la red durante su creación y posterior entrenamiento.

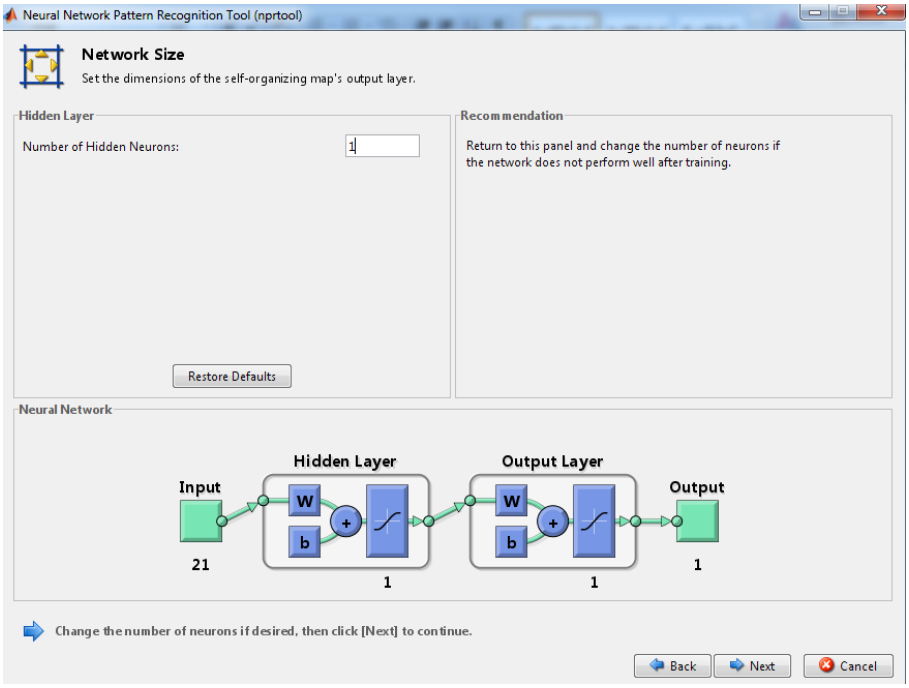


Figura 30.Creación de la segunda red neuronal.

Las redes fueron almacenadas en los archivos "FELIZNET2", "TRISTENET2" y "ENOJADONET2".

6.2.3. ENTRENAMIENTO

Los mismos 92 archivos de audio del corpus Emo-DB fueron utilizados para la creación de la matriz principal de entrenamiento, se implementó la función "ZCRMoral.m " dentro de la iteración del "TRAININGCREATE.m" para que la matriz resultante contuviera el parámetro temporal adicional, se modificó la entrada de la función kannumfcc para que realizara el cálculo de 20 coeficientes cepstrales, de acuerdo a la delimitación establecida en el apartado **6.2.2.**

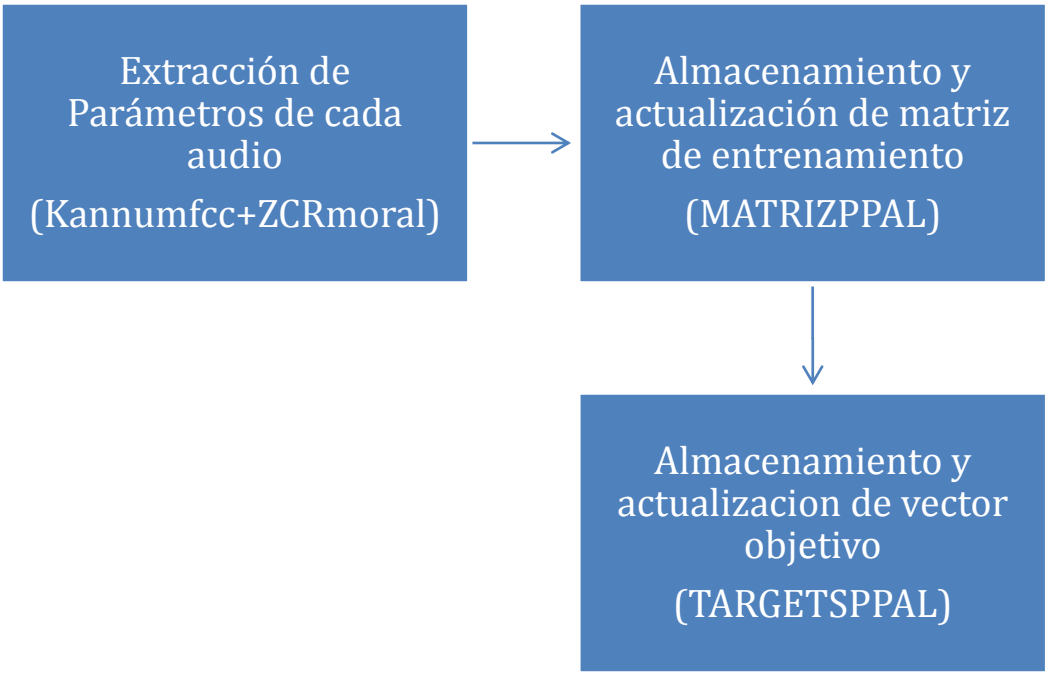


Figura 31. Diagrama de Bloques de la función TRANINGCREATE.m

La creación del vector de targets se conservó sin cambios.

Dentro de la configuración previa a la creación de cada red neuronal, se definió una división del porcentaje de muestras de entrenamiento, validación, y prueba, en 80%, 10% y 10% respectivamente.

La herramienta utilizada para la creación de esta red (y las redes multiclase implementadas posteriormente) fue el Neural Pattern Recognition Toolbox, el cual, una vez realizado el entrenamiento de las redes neuronales, realiza un diagnóstico evaluativo del funcionamiento del sistema de reconocimiento, tomando un porcentaje de la base de datos de entrada para realizar la evaluación. En el caso de las redes neuronales feedforward, el proceso de validación y prueba consisten en la verificación del funcionamiento de la red entrenada. Como estas redes serán probadas con una matriz independiente específicamente para la evaluación, se selecciona un porcentaje bajo de la base de entrenamiento para el diagnostico de

validación y pruebas (20%) y se asigna el 80% restante para el entrenamiento de la red.

A continuación se presentan las matrices de confusión para cada uno de los diagnósticos calculados posteriores al entrenamiento:

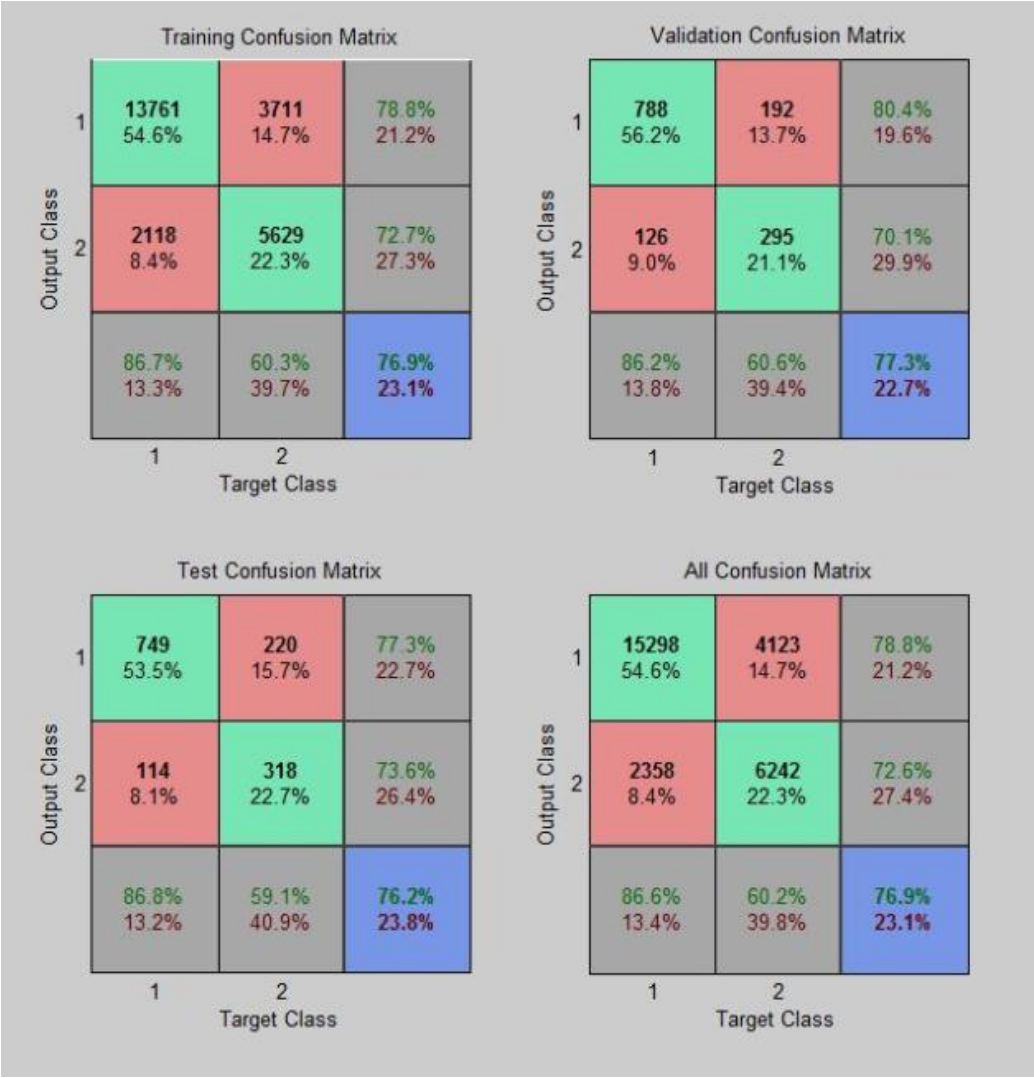


Figura 32. Matriz de confusión entrenamiento nprtool triste

En las gráficas de confusión validadas desde el Neural Pattern RecognitionTool, se observa una disminución en el porcentaje de eficiencia de las tres redes neuronales implementadas, esto puede deberse al porcentaje de muestras de validación y prueba especificadas para la etapa de creación de las redes neuronales.

6.2.4. PRUEBA

La prueba del algoritmo de reconocimiento para esta iteración se llevó a cabo con la misma matriz de prueba que en **[6.1]**, una matriz de 40 audios positivos y negativos fue generada con 21 parámetros para validar la respuesta de la red.

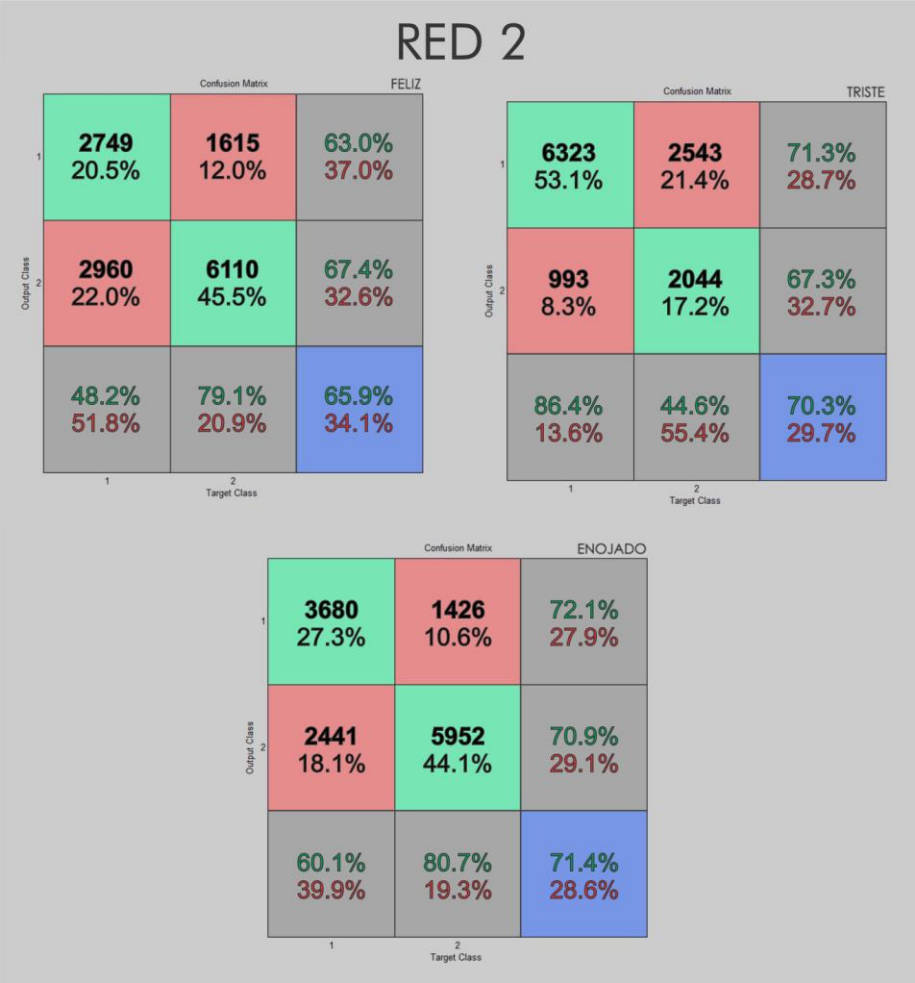


Figura 33. Matrices de confusión (pruebas en RED2).

Se observan valores de eficiencia de 65.9%, 70.3% y 71.4% para las redes de reconocimiento de felicidad, tristeza y enojo respectivamente, aunque hubo un aumento en el porcentaje de reconocimiento de 2.5% para la red de enojo, existe un decremento en el porcentaje de reconocimiento global de emociones de un 3.67%, los resultados de introducir un parámetro temporal a la matriz espectral no son satisfactorios para la eficiencia del algoritmo.

6.3. TERCERA ITERACIÓN

6.3.1. SELECCIÓN Y EXTRACCIÓN DE PARÁMETROS

Para validar la hipótesis sobre los efectos de parámetros temporales en la matriz de entrada propuesta en **[6.2.4.]**, se propone el entrenamiento de tres nuevas redes neuronales, se analizarán los mismos parámetros de **[6.2]**, modificando las condiciones de creación de las bases de reconocimiento.

6.3.2. SELECCIÓN DE ALGORITMOS DE CLASIFICACIÓN

Una posible razón al decaimiento en la eficiencia de las redes neuronales, es la disminución de las muestras de entrenamiento, debido a la división realizada por el Neural Pattern Recognition Toolbox a la matriz de entrada, por lo tanto se realizará la creación de las redes a partir del Neural Network Toolbox utilizado para la creación de las redes en **[6.1]**.

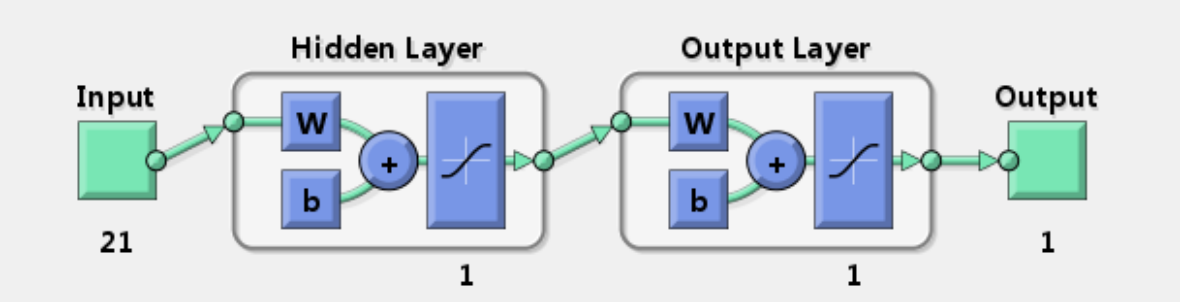


Figura 34. Estructura de la tercera red neuronal.

Las redes fueron almacenadas en los archivos "FELIZNET3", "TRISTENET3" y "ENOJADONET3".

6.3.3. ENTRENAMIENTO

Al utilizar las matrices de entrada implementadas en **[6.2]** en el Neural Network Toolbox, se asegura que el 100% de las muestras de audio utilizadas para la extracción de parámetros, será usada para el entrenamiento de las redes neuronales, se espera un incremento en la eficiencia del reconocimiento al aumentar el número de muestras de entrenamiento.

6.3.4. PRUEBA

Nuevamente se prueban las redes neuronales creadas con la matriz de entrada creada para las pruebas en **[6.2].**

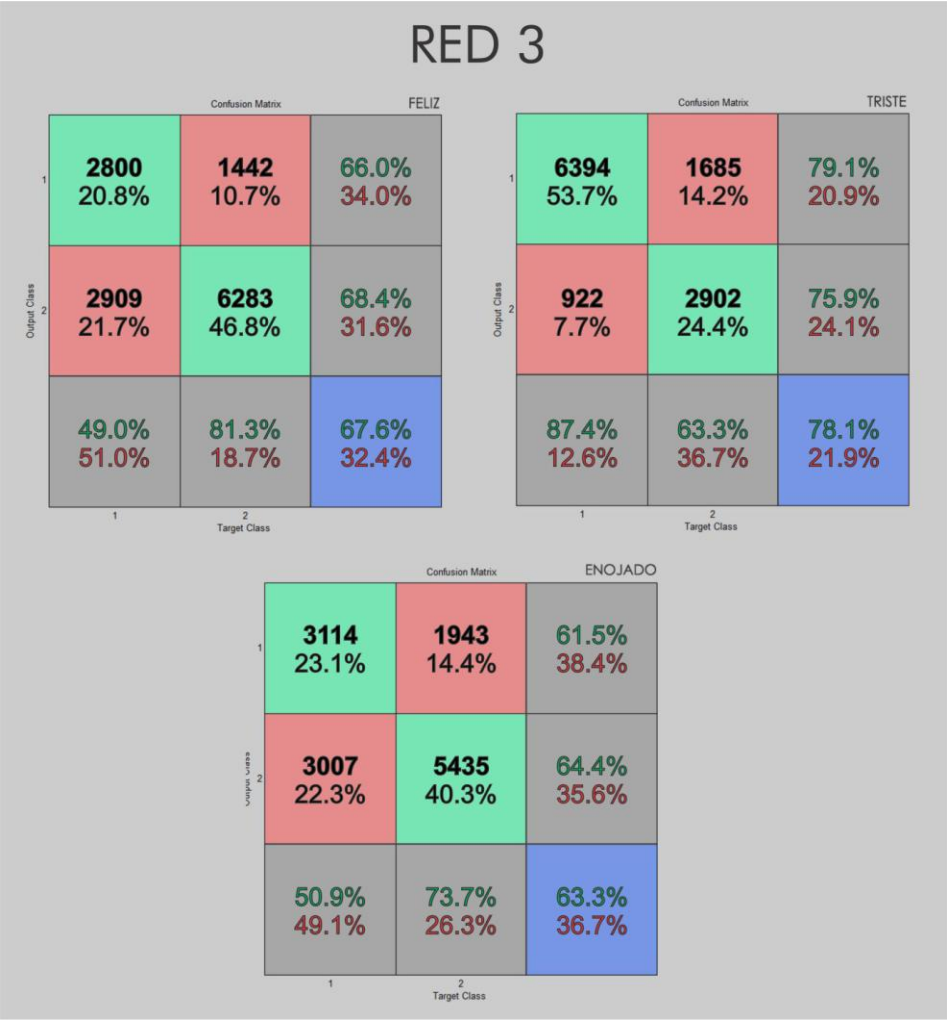


Figura 35. Matrices de confusión (pruebas en RED3).

Efectivamente, se observa un aumento en el porcentaje de reconocimiento de las redes neuronales con respecto a las redes creadas utilizando el Neural Pattern Recognition Toolbox, sin embargo los resultados siguen siendo inferiores a los obtenidos analizando sólo los coeficientes cepstrales, por lo tanto, el enfoque de la investigación se centra ahora en el análisis de estos parámetros, y en la extracción de información adicional de las envolventes energéticas que a su vez se convertirán en nuevas variables de entrenamiento.

6.4. CUARTA ITERACIÓN

6.4.1. SELECCIÓN Y EXTRACCIÓN DE PARÁMETROS

Teniendo en cuenta los resultados obtenidos con los parámetros espectrales con respecto a los temporales, se propone la extracción de información adicional de dichos parámetros.

Una forma de aumentar el porcentaje de reconocimiento emocional a partir de los coeficientes cepstrales hasta en un 20%, es analizar variables derivadas de dichos parámetros tales como:

- La velocidad, o variabilidad, de los coeficientes cepstrales.
- La aceleración, o cambios en la variabilidad, de los coeficientes cepstrales.
- La energía media en escala logarítmica de la señal de audio.
- El 0-ésimo coeficiente cepstral, correspondiente a la energía media de la envolvente.

La velocidad y aceleración de los coeficientes cepstrales son conocidos como las variables delta y delta delta**[2.1.7.2]**, puesto que no se puede realizar una derivada parcial directamente a un vector de parámetros sin realizar la aproximación de los parámetros a una función, los parámetros delta son calculados para obtener una cantidad de parámetros proporcional a la cantidad de coeficientes cepstrales calculados.

Se calculan, adicionalmente, los valores máximo, mínimo, y promedio de la energía media de cada bloque en escala logarítmica y energía media de la envolvente.

Pruebas adicionales sugirieron que incorporar nuevamente los 26 parámetros cepstrales, en efecto aumentaba ligeramente el porcentaje de reconocimiento, por tanto el cálculo de las variables cepstrales volverá a ser con 26 filtros triangulares.

En total se analizaran, entonces, 84 parámetros: 78 correspondientes a las características de envolvente energética de la señal (26 MFCC, 26 delta, 26 deltadelta) y 6 variables relacionadas con la energía media de la señal (3 0th MFCC, 3 Emid).

Para la extracción de estas características se hará uso de la librería VOICEBOX de Matlab, específicamente su función melcepst.m.

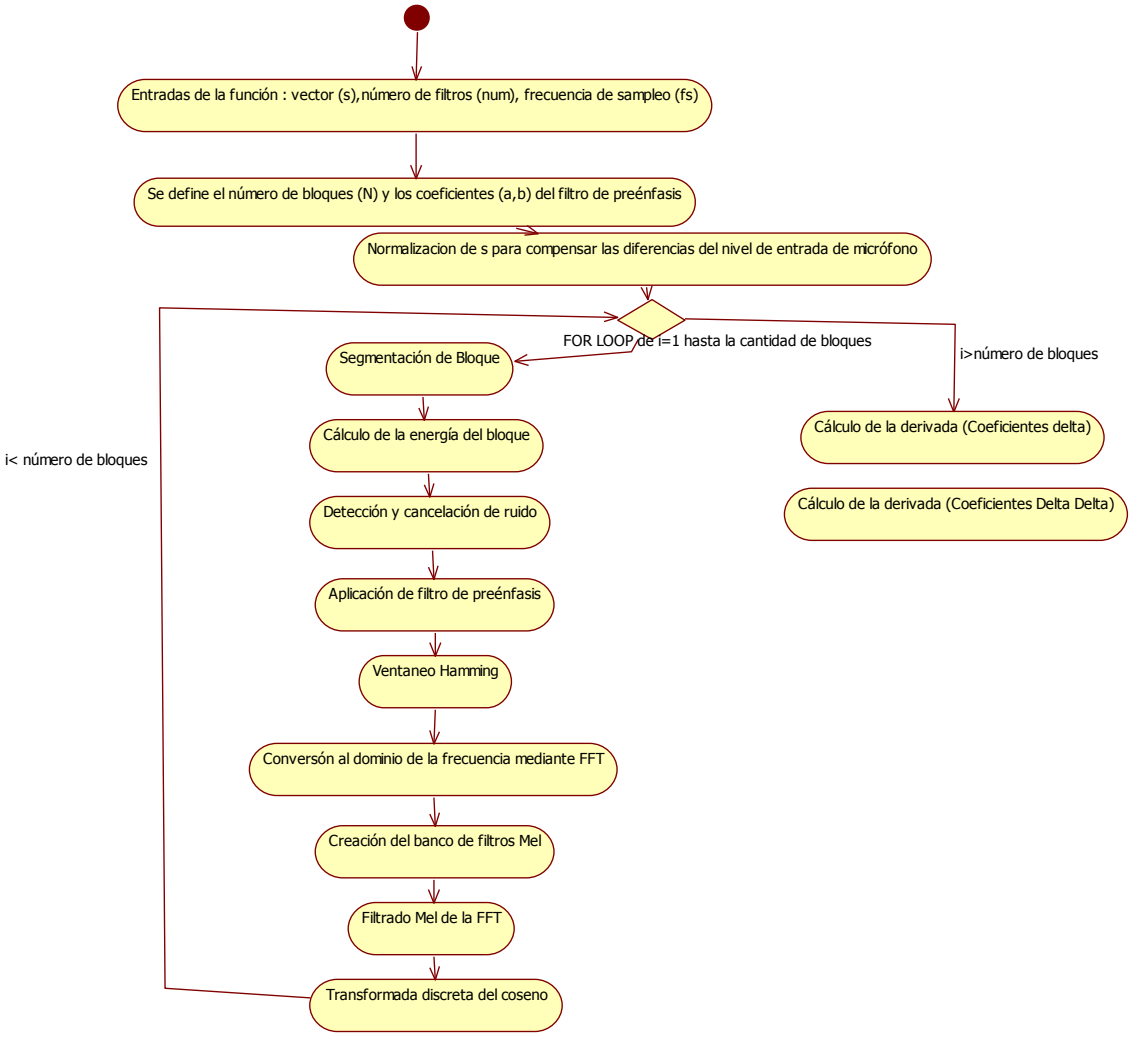


Figura 36. Diagrama de flujo para la función "melcepst.m"

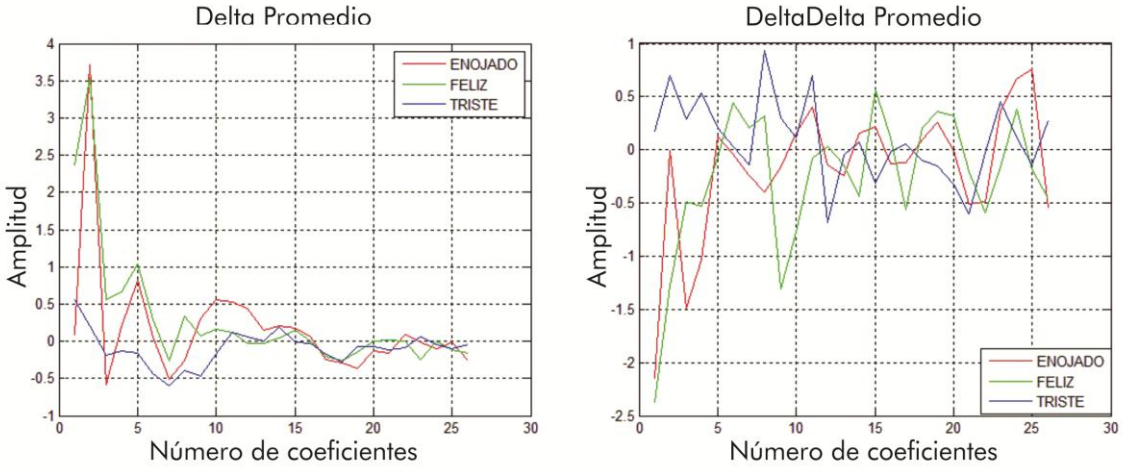


Figura 37. Valores promedio para los coeficientes delta y deltadelta

Se puede observar una notoria diferencia entre los valores promedio de los estados emocionales para ambos conjuntos de parámetros, se puede ver que la rapidez en el cambio de los coeficientes cepstrales es muy baja para los archivos de audio de tristeza, no solo los niveles energéticos son bajos en esta emoción, se ve que el cambio de nivel entre coeficientes cepstrales es lento a comparación de los cambios rápidos a baja frecuencia de las emociones de felicidad y enojo.

En los coeficientes Delta-Delta, por otra parte, se puede observar que las variaciones en la velocidad de cambio de los coeficientes cepstrales para la tristeza son mucho más pronunciadas que en la felicidad o el enojo para frecuencias bajas, al ir aumentando la frecuencia, la aceleración de las emociones de mayor activación aumenta, mientras que la de la tristeza disminuye.

6.4.2. SELECCIÓN DE ALGORITMOS DE CLASIFICACIÓN

Ya que se ha comprobado mediante las anteriores iteraciones la necesidad de maximizar la cantidad de muestras utilizadas para el entrenamiento, se continuará utilizando el Neural Network Toolbox para la generación de las tres redes neuronales, los parámetros de creación de las redes permanecen invariables al igual que en las iteraciones anteriores, solo se modifica el número de perceptrones en la capa de entrada (84).

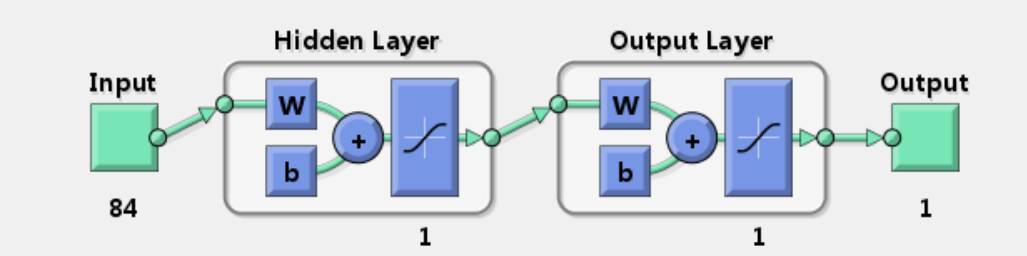


Figura 38.Estructura de la cuarta red neuronal.

Las redes fueron almacenadas en los archivos "FELIZNET4", "TRISTENET4" y "ENOJADONET4" , con sus correspondientes matrices de entrenamiento, prueba, y vectores target.

6.4.3. ENTRENAMIENTO

La base de audios utilizada para el entrenamiento en este caso fue la misma utilizada en **[6.1]**,en este caso la función melcepst.m fue implementada dentro del proceso iterativo de extracción de parámetros por archivo de audio, de esta manera se obtiene una matriz de 84 columnas, cada una correspondiente a los parámetros elegidos para optimizar el reconocimiento.

6.4.4. PRUEBA

Los resultados de reconocimiento para los tres estados emocionales, utilizando la matriz de prueba usada en **[6.1]** implementando la función melcepst.m para la extracción de parámetros son los siguientes:

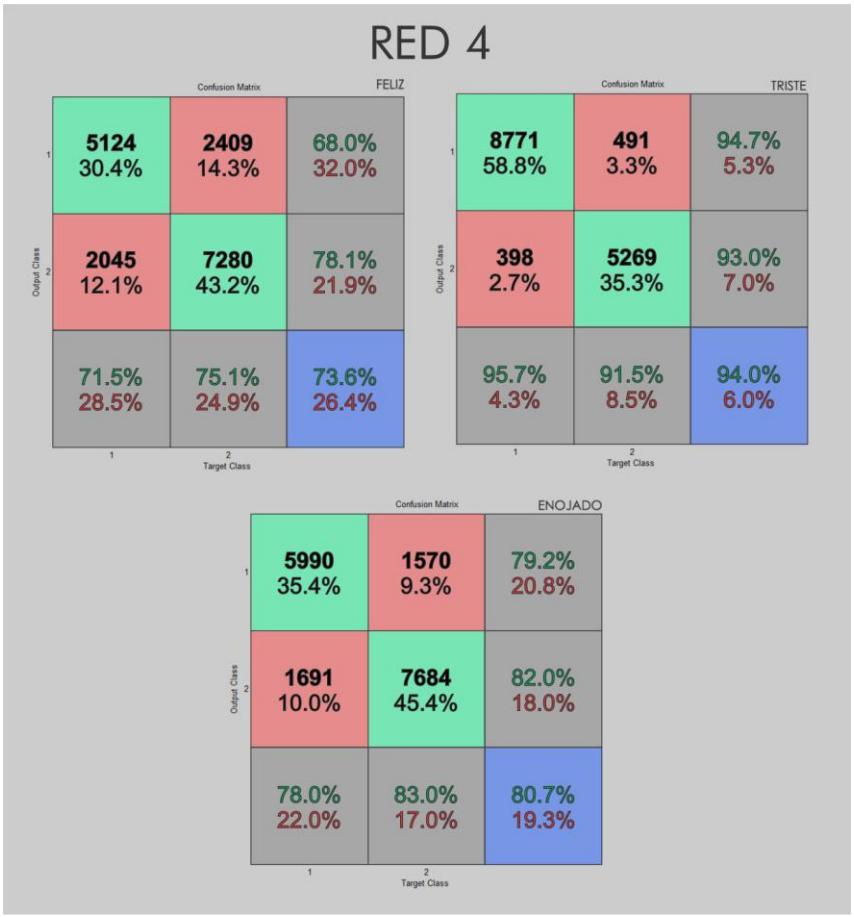


Figura 39. Matrices de confusión (pruebas en RED4).

Se puede observar por los resultados de la matriz de confusión de los tres estados emocionales un aumento de la eficiencia de reconocimiento hasta de un 11.8% con respecto a los resultados de **[6.1]**, obteniendo un valor de reconocimiento máximo de 94% para la red neuronal de reconocimiento de tristeza. El sistema de redes neuronales entrenadas a partir de los parámetros espectrales envolventes, se considera adecuada para el reconocimiento de emociones foráneas al idioma del corpus de entrenamiento (alemán). Se procede entonces a realizar pruebas adicionales a la red neuronal con voces en español.

6.4.5. PRUEBA DEL CORPUS EN ESPAÑOL CON LAS REDES NEURONALES

Se realizaron varias pruebas con matrices de parámetros espectrales correspondientes a los audios presentes en el corpus, la organización y prueba de estas matrices se clasificó así:

- Pruebas por locutor.
- Pruebas por fragmento de texto.
- Pruebas globales.

Para cada una de estas pruebas, se seleccionó el total de archivos de audio correspondiente a su clase (locutor/fragmento/emoción). Los resultados de todas las pruebas realizadas se presenta en el apartado **[7.1.1].**

Es necesario profundizar el análisis sobre los sistemas de reconocimiento de patrones para poder realizar la comparación de resultados entre las voces nativas y foráneas en cuanto a la expresividad de sus emociones a través de la voz.

6.5. REDES NEURONALES MULTICLASE

Para poder realizar la comparación objetiva de los resultados de reconocimiento de las redes neuronales artificiales con voces en español y alemán, es necesario optimizar las redes para la tarea de reconocimiento de patrones, de igual forma, se hace conveniente poder tener una red neuronal entrenada con el corpus de audios en español, y posteriormente probada con una matriz de parámetros espectrales extraídos del corpus alemán. Las redes neuronales multiclase no solo permiten realizar estas dos tareas, también permiten establecer una capa de salida binaria de más de un perceptrón, lo cual es adecuado para las salidas necesarias para el proyecto.

6.5.1. CREACIÓN Y ENTRENAMIENTO DE REDES NEURONALES MULTICLASE

Mediante el uso del Neural Pattern Recognition Toolbox, se procedió a la creación de cinco redes neuronales multiclase, las cuales tienen el siguiente diagrama de bloques básico:

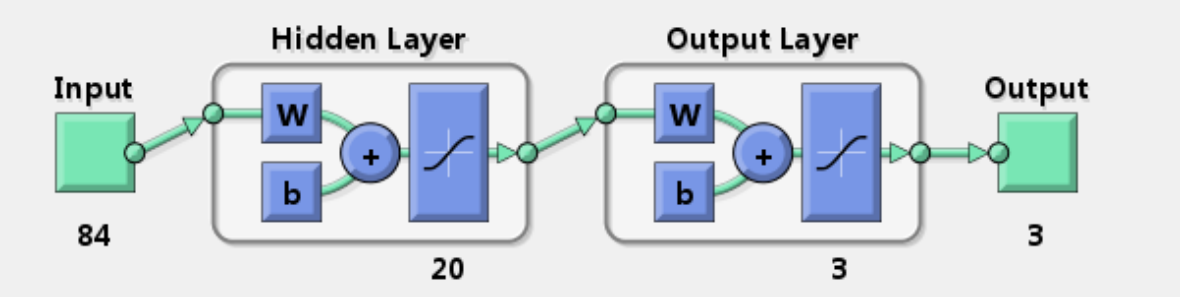


Figura 40.Estructura de la primera red multicapa.

Debido a la cantidad limitada de audios en español para el entrenamiento, fue necesario darle mayor grado de complejidad a la red neuronal, aumentando sus capas ocultas para así poder realizar un aprendizaje más preciso de los patrones emocionales presentes en los 72 audios en español.

Una diferencia fundamental entre la creación de las redes de reconocimiento multiclase y las redes binarias creadas en las iteraciones iníciales es la generación de la matriz de targets. Para el caso de las redes multiclase es necesario incorporar la información binaria de los tres estados emocionales en una matriz. Puesto que las teorías de la emoción en el espacio continuo, asumen los estados emocionales independientes entre sí (siendo dependientes de las variables de activación, potencia y valencia)**[2.1.2]**, a cada emoción se le asignará una ponderación en su resultado entre 0 y 1, siendo 0 la inexistencia de la emoción en la voz, y 1 una total presencia de la emoción en la voz. Lo anterior indica que para cada columna de la matriz de objetivos se define un estado emocional (valor entre 0 y 1). Se organizó esta matriz de targets para que su primera columna contuviera la salida ideal para la emoción felicidad, la segunda para tristeza y la tercera para enojo.

```
%MATRIZ TARGETS POSITIVO: FELIZ
TARGETSPPAL(1:length(MATRIZPPAL),1:3)=zeros;
TARGETSPPAL(1:length(MATRIZPPAL),1)=ones;
%MATRIZ TARGETS POSITIVO: TRISTE
TARGETSN1(1:length(MATRIZN1),1:3)=zeros;
TARGETSN1(1:length(MATRIZN1),2)=ones;
TARGETSPPAL=[TARGETSPPAL';TARGETSN1'];
%MATRIZ TARGETS POSITIVO: ENOJADO
TARGETSN2(1:length(MATRIZN2),1:3)=zeros;
TARGETSN2(1:length(MATRIZN2),3)=ones;
TARGETSPPAL=[TARGETSPPAL';TARGETSN2'];
```

Figura 41. Creación de matriz de objetivos para las redes multiclase "TRAININGMULTICLASS.m"

Las características de las cinco redes neuronales implementadas son las siguientes:

6.5.1.1. MULTICLASS COLOMBIANA 1

La primera red neuronal multiclase implementada, fue entrenada con el corpus completo de voces en español (72 audios) , se extrajeron los 84 parámetros de análisis para cada bloque de cada audio, esta matriz principal fue dividida en un 80% para entrenamiento, 10% para validación y 10% para pruebas, mediante las funciones TRAININGMULTICLASSCREATE.m se generó dicha matriz, al igual que la matriz target multiclase necesaria para el entrenamiento.

De igual forma a lo presentado en la segunda iteración, la selección de los porcentajes de entrenamiento, validación y prueba se debe a la necesidad de tener una cantidad de datos más robusta para el entrenamiento, ya que posteriormente se realizarán pruebas con una matriz creada específicamente para evaluar el funcionamiento de la red.

Figura 42. Matrices de confusión para la primera red multiclase.

Como se puede observar, el porcentaje de reconocimiento global para la red multiclase es de un 70.5%, sin embargo el porcentaje de reconocimiento para felicidad (56.2%) debe ser optimizado para poder realizar pruebas sobre las voces en alemán.

6.5.1.2. MULTICLASS COLOMBIANA 2

Para mejorar la eficiencia de la red neuronal multiclase entrenada con voces colombianas, se procedió a aumentar el número de capas ocultas a 30, por otra parte, el porcentaje de parámetros de entrenamiento se incrementó de 80% a 90%, dejando el 10% restante para validación y pruebas:

Figura 43. Matrices de confusión para la segunda red multiclase

Tanto el porcentaje de reconocimiento para la emoción de felicidad (+3.9%) como el reconocimiento general (+2.2%) aumentaron lo suficiente para poder realizar pruebas con voces alemanas.

La matriz de prueba de voces alemanas se generó con 40 audios aleatorios del corpus, los resultados de las pruebas se presentan en la siguiente matriz de confusión:

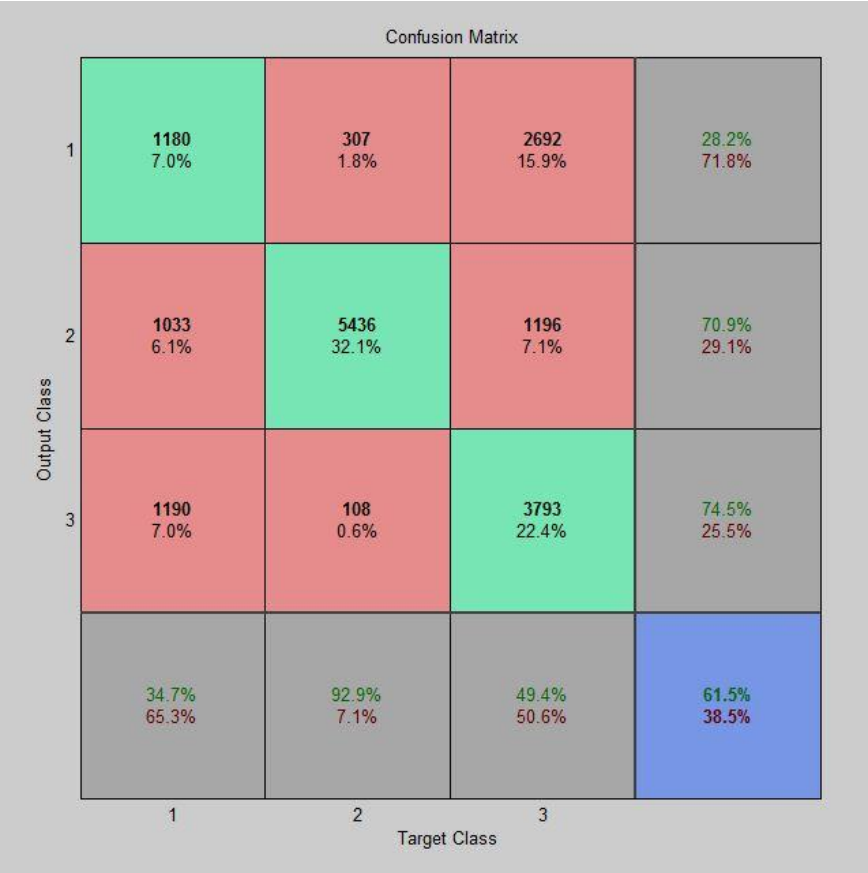


Figura 44.Matriz de confusión para pruebas alemanas.

Aunque el porcentaje de reconocimiento de las voces en alemán es bajo, es suficiente para sacar conclusiones de los componentes diferenciadores de reconocimiento entre ambos idiomas.

6.5.1.3. MULTICLASS ALEMANA 1

Adicionalmente a la creación de las redes multiclase entrenadas con voces en español, se genera una matriz multiclase entrenada con voces en alemán, para poder comparar la eficiencia de las redes binarias creadas anteriormente con respecto a las redes multiclase.

La matriz de entrenamiento fue implementada utilizando 62 archivos de audio del corpus Emo-DB por estado emocional, para un total de 186 audios procesados, la división de las muestras fue 90% entrenamiento, 5% validación y 5% pruebas, se generó la red de reconocimiento de patrones con diez capas ocultas.

Figura 45. Matrices de confusión para la tercera red multiclase.

Aunque el porcentaje de reconocimiento de tristeza es mayor que el obtenido por cualquiera de las redes binarias implementada en las iteraciones del primer acercamiento investigativo, se desea que el reconocimiento para las emociones de felicidad y enojo no se alejen tanto del porcentaje obtenido por FELIZNET4 Y ENOJADONET4.

6.5.1.4. MULTICLASS ALEMANA 2

Tanto la matriz de entradas y la división de muestras entrenamiento-validación-prueba permanecieron constantes con respecto a la anterior red neuronal. Para generar mayor profundidad en el entrenamiento, se implementaron treinta capas ocultas en el algoritmo de la red para mejorar la eficiencia del sistema.

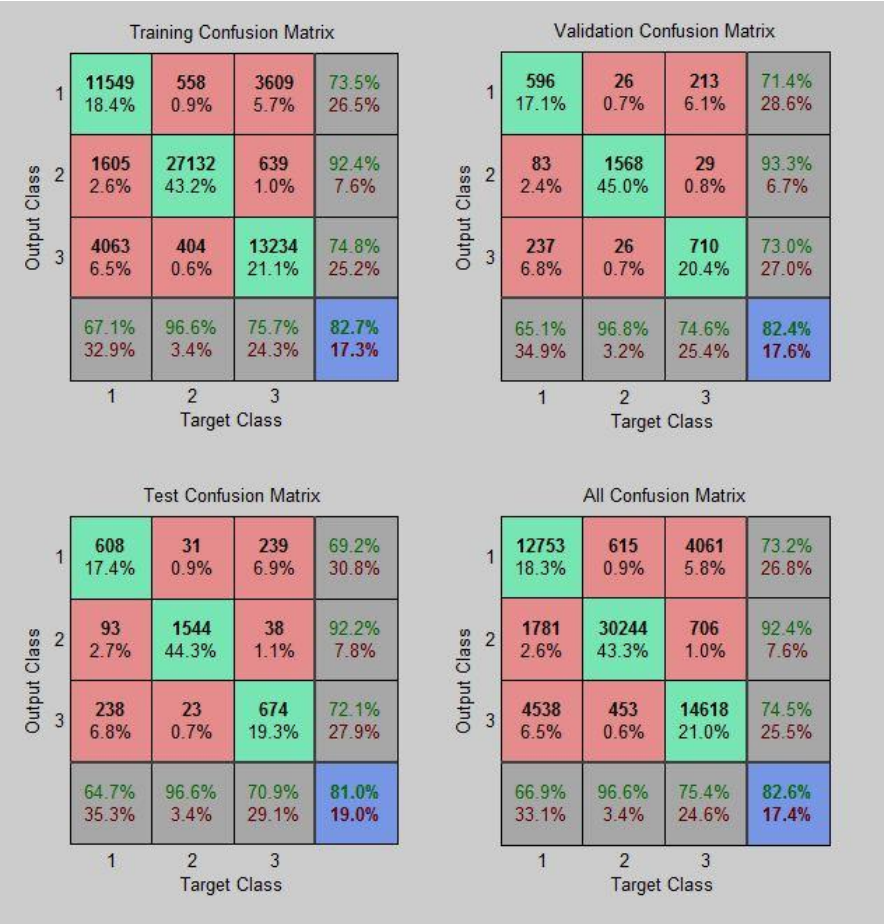


Figura 46. Matrices de confusión para la cuarta red multiclase.

Se puede observar como la eficiencia de la red neuronal implementada supera en un 3.5% a la anterior red con diez capas ocultas.

6.5.1.4. MULTICLASS HÍBRIDA

Como un último análisis, se implementó una red multiclase entrenada con una matriz de parámetros espectrales pertenecientes a archivos de audio de los bases de archivos de audio foráneas y nativas, se seleccionaron los 72 archivos de audio del corpus español con 72 audios aleatorios entre los tres estados emocionales del corpus alemán y se entrenó una red multiclase de treinta capas ocultas y una división de muestras de 90% entrenamiento, 5% validación y 5% pruebas.

Figura 47. Matrices de confusión para la quinta red multiclase.

Aunque la eficiencia global de la red no es el mayor de todas las redes creadas, al estar entrenada con voces españolas y alemanas y tener un porcentaje de reconocimiento de 76.3%, se pueden realizar observaciones importantes en el reconocimiento de patrones emocionales en conjunto de los dos idiomas estudiados.

6.6.1. INTEGRACIÓN DE RESULTADOS A TRAVÉS DE UNA INTERFAZ DE USUARIO

Una vez realizado el análisis de las diferentes redes neuronales y parámetros utilizados para su entrenamiento, se procede a la creación de una interfaz de usuario para poder visualizar resultados de reconocimiento con archivos de audio específicos almacenados y grabados.

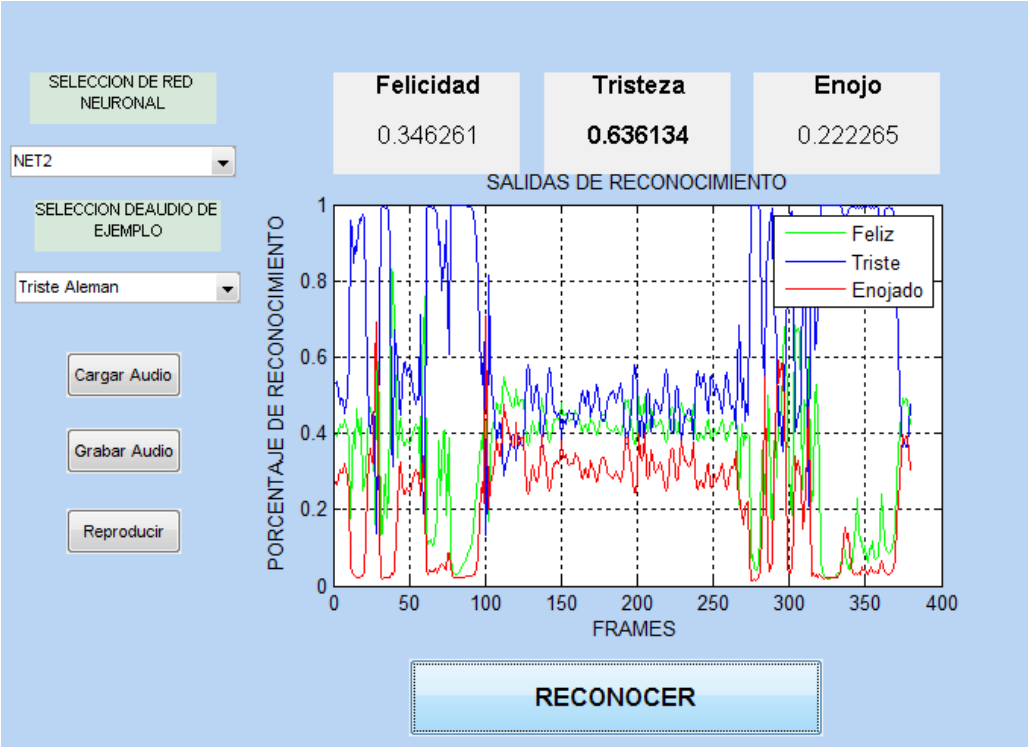


Figura 48. Interfaz de usuario para pruebas adicionales.

Mediante la interfaz, el usuario puede elegir una de los nueve sistemas de redes neuronales artificiales que llevará a cabo el reconocimiento. Dependiendo de la red escogida, la interfaz realizará la extracción de los parámetros de entrada correspondiente a esa red (MFCC, ZCR, Delta, Delta-Delta,Energía) para así generar la matriz de prueba correspondiente.

```
function pushbutton3_Callback(hObject, eventdata, handles)
global fun audio Fs M1 M2 M3
set(handles.text3,'FontWeight','normal')
set(handles.text5,'FontWeight','normal')
set(handles.text7,'FontWeight','normal')
switch fun
    case 1
        test=kannumfcc(26,audio,Fs);
        test=test';
        outputsf=sim(M1,test);
        Feliz=mean(outputsf);
        outputst=sim(M2,test);
        Triste=mean(outputst);
        outputse=sim(M3,test);
        Enojado=mean(outputse);
        set(handles.text3,'String',Feliz)
        set(handles.text5,'String',Triste)
        set(handles.text7,'String',Enojado)
```

Figura 49. Extracción de parámetros para el primer sistema de redes neuronales NET1

Otra entrada del usuario a la interfaz es el archivo de audio a probar. El usuario puede seleccionar entre seis archivos de audio de ejemplo (tres en alemán y tres en español) para verificar el reconocimiento de la red seleccionada. Adicionalmente el usuario puede seleccionar un archivo de audio diferente a los ejemplos mediante el botón de "Cargar Audio".

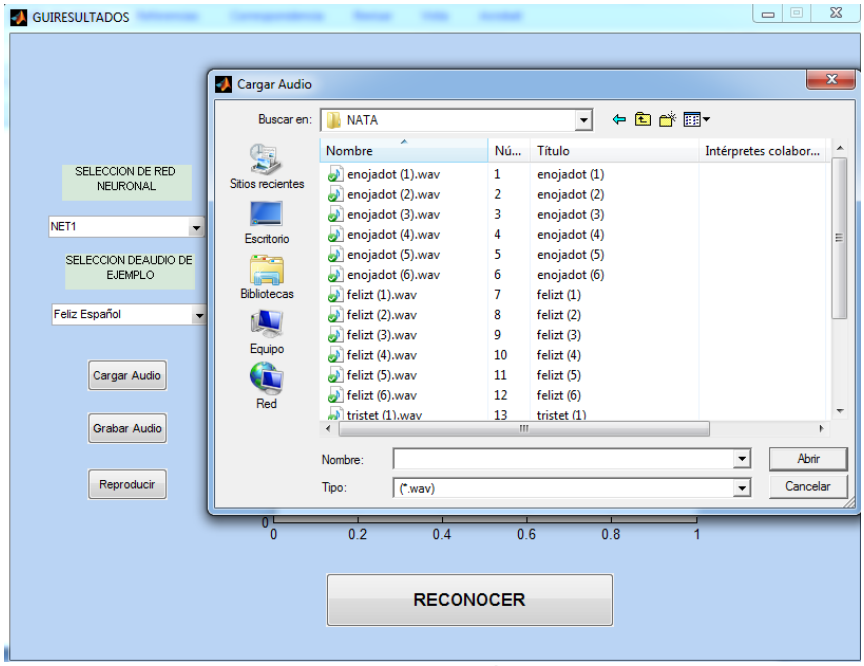


Figura 50. ventana de selección de archivos .wav

Finalmente, el usuario puede elegir realizar la grabación de la voz para su posterior procesamiento. El usuario seleccionará mediante una ventana emergente cuánto tiempo (en segundos) se tomarán para el proceso de grabación. La captura se realiza mediante el comando de wavrecord y la frecuencia de sampleo para este audio se deja fija en 16000 Hz, puesto que todas las redes neuronales fueron entrenados con audios codificados a esta misma frecuencia, y al intentar realizar pruebas con audios de frecuencias de sampleo superiores llevaría a una discordancia entre los bancos de filtros en frecuencia y la frecuencia real que entra a las redes. Cabe anotar que no se realizará ningún preprocesamiento adicional a la captura de los audios, debido que esto forma parte del desarrollo e implementación del algoritmo en tiempo real. Por otra parte, es interesante analizar el nivel de sensibilidad del reconocimiento de las redes neuronales implementadas a parámetros de error como lo son el ruido de fondo y los silencios entre palabras.
En la ventana principal de la interfaz, se presentan dos visualizaciones de evaluación:

- Los valores promedio de los vectores de salida correspondientes a las emociones de feliz, triste, y enojado.

Estos valores presentan un veredicto final al reconocimiento, al realizar una promedio en las salidas emocionales de todos los bloques extraídos de la señal de audio, se puede visualizar un valor general de la emoción más fuerte durante el tiempo total del audio. La interfaz resaltará el valor medio más alto entre las tres emociones, realizando así el reconocimiento. Puesto que los valores promedio se calculan a partir de las salidas de las redes neuronales, que debido a su función de transferencia (logarítmica sigmoide) están acotadas al rango [0,1], estos valores promedio también estarán en dicho rango.

Figura 51. Reconocimiento promedio

La gráfica presentada en la interfaz de usuario muestra los tres vectores de salida de la red neuronal utilizado en el reconocimiento. En ella se puede analizar los cambios emocionales de la voz en el tiempo (cada bloque corresponde a diez milisegundos de la señal de audio), al igual que los cambios y cruces entre las tres emociones para estos intervalos.

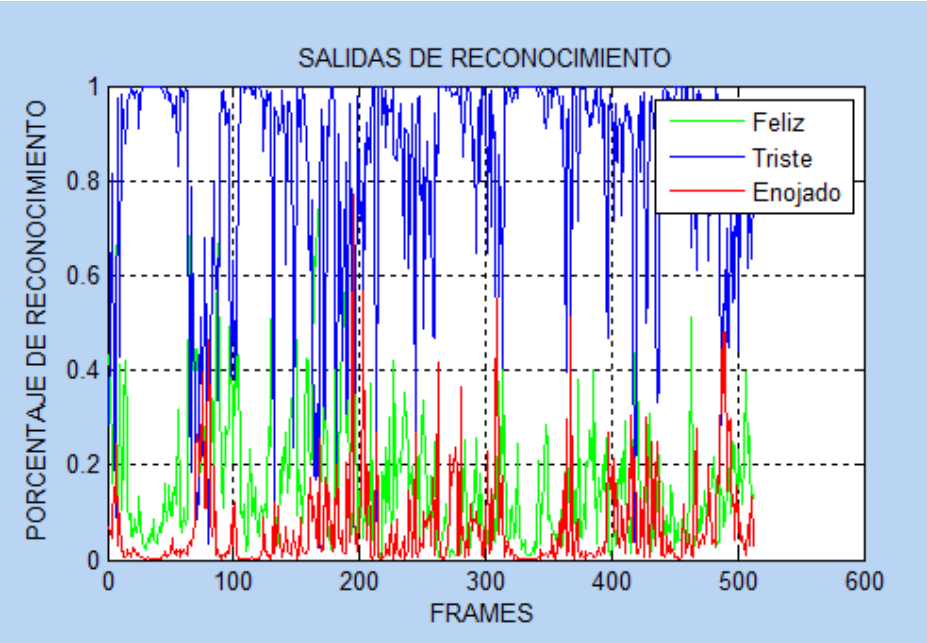


Figura 52. Ventana de reconocimiento emocional en el tiempo.

6.6.2. PRUEBAS ALEATORIAS UTILIZANDO LA INTERFAZ GRÁFICA

Se realizaron pruebas aleatorias para comprobar el funcionamiento de la interfaz, capturando audios de entre 4 y 5 segundos con un micrófono de cámara web

Logitech. Los resultados de reconocimiento de la voz hablando en español fueron altos para las redes neuronales NET1 y multiclasshib.

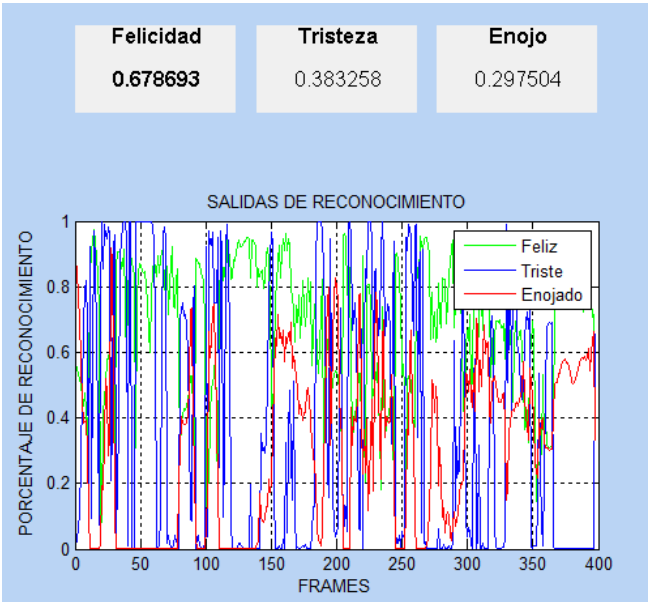


Figura 53. Resultados para prueba de cuatro segundos de captura de audio (reconocimiento correcto)

7.1. ANÁLISIS DE RESULTADOS

7.1.1. COMPARACIÓN DE LAS VOCES EN ESPAÑOL

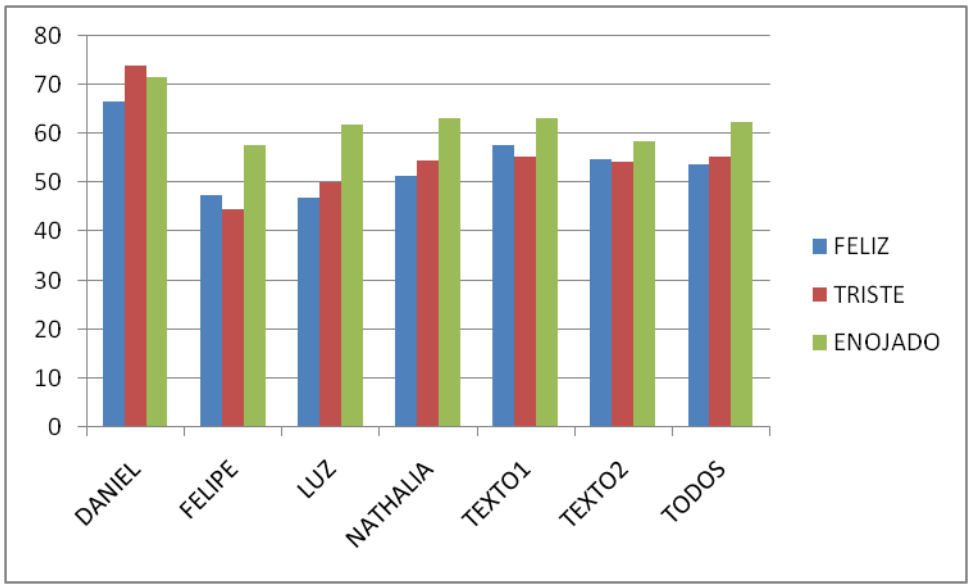


Figura 54. Resultados específicos para las pruebas con voces en español

Aunque se presentaron valores de reconocimiento superiores al 65% para el caso del locutor Daniel Morales, la tendencia de los porcentajes globales están entre el 50% y el 60%. El reconocimiento general de los dos textos analizados están por encima de la media de reconocimiento, por lo tanto fueron apropiados para las intenciones emocionales realizadas en la locución.

7.1.2.RESULTADOS GLOBALES

Los resultados de reconocimiento para los nueve sistemas de redes neuronales implementados en el desarrollo ingenieril se presentan en el siguiente diagrama de barras:

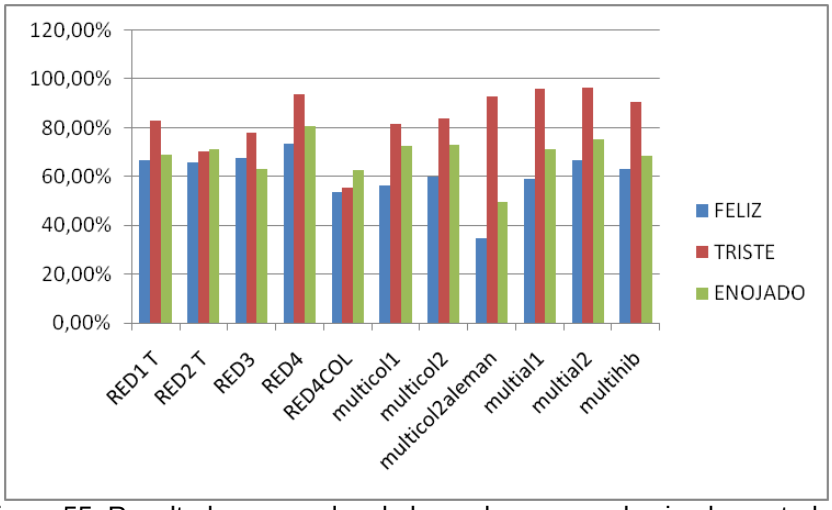


Figura 55. Resultados generales de las redes neuronales implementadas

Adicionalmente a los resultados de las pruebas con voces pertenecientes a la base de datos con la que fueron entrenados los algoritmos, se muestra el reconocimiento de voces en español por la red NET4 entrenada en alemán, y el reconocimiento de voces en alemán con la red multiclasscol2 entrenada en español. Se puede ver que la tristeza tiene un porcentaje de reconocimiento superior con respecto a las otras dos emociones, esto es debido a su distancia en el plano dimensional continuo descrito en **[2.1.3]**con respecto a la felicidad y al enojo.

Aunque el porcentaje de reconocimiento de las voces en español por las redes alemanas es bajo, tasas de reconocimiento por encima del 57% son superiores al reconocimiento emocional de los humanos al intentar percibir las emociones por la voz[20]. Por otra parte, dentro de las pruebas aleatorias llevadas a cabo usando la interfaz gráfica implementada para la integración de los algoritmos, se alcanza una tasa de reconocimiento superior a los resultados presentados, por lo tanto, se puede validar la analogía de los parámetros emocionales extraídos de las voces alemanas con los parámetros provenientes de las voces nativas de Colombia.

[20]OUDEYER, Pierre. *The Production and Recognition of Emotions in Speech: Features and Algorithms".* Paper Released in Science Direct Human Computer Studies, Sony CSL. Paris, Nov 30, 2002.

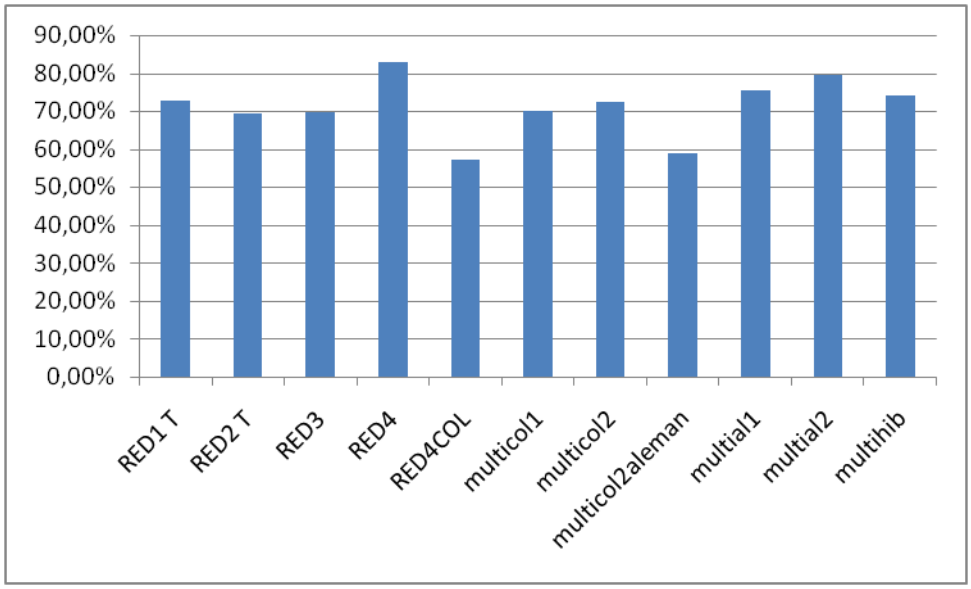


Figura 56. Porcentaje de reconocimiento global para las redes neuronales implementadas

Esta gráfica presenta los porcentajes de reconocimiento global para las nueve redes neuronales implementadas, se alcanza una tasa de eficiencia de 82.77% para la RED4, implementando tres redes neuronales binarias independientes para cada emoción, y de un 79.63% para la red multiclase muclticlassal2. La diferencia entre la eficiencia de las redes entrenadas con la base de datos Emo-DB y el corpus en español capturado para el desarrollo del proyecto se deben a las condiciones de grabación casi ideales del corpus alemán (selección de locutores y condiciones de reverberación) .

7.1.3. COMPARACION CON RESULTADOS INSTITUCIONALES

Los estudios realizados en la Universidad de San Buenaventura que involucran aprendizaje computacional[21], han centrado su investigación en el reconocimiento de voz. Esta es la primera ocasión en la que se realiza un estudio institucional de reconocimiento de emociones en el habla. Sin embargo se pueden realizar comparaciones de los resultados de los algoritmos de reconocimiento de patrones entre los proyectos. En el proyecto realizado en el 2010[22], el porcentaje de acierto en el reconocimiento de voz utilizando redes neuronales artificiales fue de un 14%, mejorándolo a 87% realizando un análisis de formantes en los archivos de audio.

[21]BAEZ, Javier. *Diseño de un Dispositivo para El Reconocimiento de Caracteres Vocálicos, para Ordenar Comandos al Televisor.* Documento de Grado, Universidad San Buenaventura. Bogotá, Septiembre 17, 2009

[22]ALDANA A. PIÑEROS J.*Desarrollo e Implementación de un Algoritmo de Reconocimiento de Voz que Permita Seleccionar una Imagen a Partir de un Banco de Nueve Fotografías Utilizando Redes Neuronales.* Documento de Grado, Universidad San Buenaventura. Bogotá, Oct. 22, 2010

Como se mencionó anteriormente, el análisis de formantes no es competente a la detección emocional en la voz, debido a la variabilidad temporal de los audios de entrenamiento y pruebas. El porcentaje de reconocimiento de patrones utilizando las redes neuronales artificiales implementadas en este proyecto, superan por un gran margen los resultados obtenidos en el antecedente de sus algoritmos de reconocimiento.

7.1.6. COMPARACIÓN CON RESULTADOS NACIONALES

Dentro de una estudio de reconocimiento de emociones en la voz utilizando MFCC realizado en la Universidad Industrial de Santander[23] , la matriz de confusión final presenta valores de reconocimiento de 90 y 100%, sin embargo en el documento no se especifica el algoritmo de reconocimiento (solo se hace referencia al algoritmo K-means), y no establecen los parámetros de entrenamiento y aprendizaje del sistema, por tanto, y teniendo en cuenta los porcentajes de reconocimiento investigados en la documentación, se pone en duda la veracidad de los datos proporcionados en el artículo.

En el proyecto realizado en la Universidad Tecnológica de Pereira en el 2007[24], se utilizaron parámetros no convencionales para encontrar características emocionales en la voz, estas características fueron divididas en parámetros semiperiódicos (Frecuencia fundamental) y Parámetros de perturbación (Jitter, Shimmer). Otros parámetros analizados fueron la transformada Gabor y la transformada Wavelet de la señal. Los resultados de reconocimiento en el proyecto fueron de entre 64,66% y 94,66%. la extracción de características fue realizado en audios en español, y el reconocimiento fue llevado a cabo sobre pruebas en la lengua nativa. Los resultados de reconocimiento de los algoritmos entrenados y probados con las voces alemanas del corpus Emo-DB, alcanzan los valores de reconocimiento propuestos en la tesis nacional.

7.1.7. COMPARACIÓN CON RESULTADOS INTERNACIONALES

- **ESTUDIOS DE APRENDIZAJE COMPUTACIONAL**

[23] RUEDA,E,TORRES,Y. *Identificación de emociones en la voz* .Thesis Degree Paper, Universidad Distrital de Santander.Colombia .2007.

[24]MORALES, M, ECHEVERRY, J, OROZCO, A. *Reconocimiento de emociones empleando procesamiento digital de la señal de voz.* Documento de Grado, Universidad Tecnológica de Pereira. Colombia.

En los estudios realizados en el Instituto Politécnico de Madrid[25], se encontraron valores de reconocimiento de emociones en la voz de un 66.6% para la tristeza, 23.8% para el enojo y 52.4% para la alegría. Los valores de reconocimiento alcanzados en este proyecto superan los resultados de la investigación española, la cual centró su estudio en la extracción de parámetros temporales y energéticos de los archivos de audio (Fonemas y Duración).

Estudios llevados a cabo en la Universidad de Munich, en Alemania[26], realizaron el reconocimiento extrayendo parámetros temporales, energéticos y estadísticos de la voz,(principalmente, aquellos relacionados con la afinación, duración, y energía de la señal), los resultados del algoritmo para el reconocimiento de todo el espectro emocional propuesto en la norma MPEG4 (incluyendo las tres emociones estudiadas en este proyecto), fueron de un 71.62%. Es importante resaltar la similitud en los resultados obtenidos en la investigación alemana y la realizada en el presente documento, teniendo en cuenta que este estudio se enfocó en las envolventes energéticas de la voz.

- **ESTUDIOS DE RECONOCIMIENTO EN INDIVIDUOS**

En la investigación realizada por Pierre-Yves Oudeyer[27], menciona que el resultado de la eficiencia de reconocimiento de emociones a partir de la voz, llevada a cabo por escuchas japoneses, alcanzaba valores máximos de 60% para audios de voces en japonés e inglés (americano). En sus siguientes experimentos, escuchas de diferentes nacionalidades fueron puestos en la tarea de identificar emociones en audios en francés, con resultados homogéneos de un 57% de reconocimiento en promedio. Los resultados globales de reconocimiento de los sistemas neuronales creados en el desarrollo del proyecto (74.03%) representa una tasa mucho mayor de reconocimiento que los resultados obtenidos en pruebas humanas.

[25] MONTERO, GUTIERREZ. *Analysis and modelling of emotional speech in spanish. Documento de Grado*, Universidad Politécnica de Madrid. España.

[26] SCHULLER, Bjorn, RIGOLL, Gerhard. *Hidden Markov model based speech emotion recognition. Thesis Paper*, Universidad Técnica de Munich. Alemania. 2003

[27] OUDEYER, Pierre. *The Production and Recognition of Emotions in Speech: Features and Algorithms".* Paper Released in Science Direct Human Computer Studies, Sony CSL. Paris, Nov 30, 2002.

7.2. CONCLUSIONES

- El análisis energético de las señales de audio para la extracción de parámetros es un método eficaz de obtener datos que pueden ser utilizados en el entrenamiento de algoritmos de reconocimiento emocional. El análisis de los coeficientes cepstrales, al igual que de su velocidad y aceleración, validan las similitudes y marcadas diferencias de las emociones estudiadas con respecto a sus componentes de potencia, valencia y activación.

- Las redes neuronales artificiales son algoritmos de reconocimiento que, por medio de aprendizaje computacional, permiten realizar predicciones específicas de estados emocionales al ser entrenadas con parámetros energéticos.

- El uso de las redes neuronales artificiales fue útil para encontrar una generalización en los patrones de reconocimiento, de esta forma se validan los resultados obtenidos en las redes binarias creadas en las primeras iteraciones del proyecto.

- La diferencia en el porcentaje de reconocimiento general de diferentes voces en español es muy baja, por tanto los parámetros emocionales extraídos de la voz son independientes del género, edad y grado de experticia de los locutores, este resultado puede corroborarse en investigaciones futuras que involucren la creación de un corpus de voces con emociones no actuadas.

- Los resultados de las pruebas realizadas entre idiomas se deben a las fuertes diferencias entre las lenguas analizadas (Anglosajona y Romance), al igual que las diferencias entre las condiciones de captura de los dos corpus de audio utilizados (cámara anecóica, preselección de locutores, entre otros.).

- El porcentaje de reconocimiento promedio general de todas las redes neuronales implementadas fue de un 74.03%, un valor mucho más elevado que el porcentaje de reconocimiento emocional alcanzado por sujetos de prueba humanos.

- Los resultados obtenidos en el desarrollo del proyecto son comparables, y en algunos casos superan, los resultados de investigaciones similares a nivel nacional e internacional.

- Al poder realizar un reconocimiento satisfactorio de las emociones en la voz utilizando algoritmos de reconocimiento entrenados con voces foráneas, se

comprueba la teoría de parámetros universales propuesta por Charles Darwin [28], para voces Colombianas.

[28]DARWIN, Charles. *The expression of the Emotions in Man and Animals.*Escocia. 1872

8. BIBLIOGRAFÍA

- FRIBERG, Anders*. Digital Audio Emotions - An Overview of Computer Analysis and Synthesis of Emotional Expression in Music.* Paper. Espoo, Finland.September 1-4, 2008.
- MOURJO*POU*LOS*,* J*.*, TSOUK*A*LA*S,* D. *Neural Network Mapping to Subjective Spectra of Music Sounds.* Presented at the 90th AES Convention. Paris. February 19-22, 1991.
- KOSTEK, Bozena. *Application of Learning Algorithms to Musical Sound Analyses.* Presented at the 97th AES Convention. San Francisco. November 10-13, 1994.
- FABBR*I* Richard J*. Neuralphoneme recognition during a cocktail party.* Presented at the 94th AES Convention. Berlin. March 16-19, 1993.
- CHRISTENSEN, Niels Sander; CHRISTENSEN, Karl Ejner; WORM, Henning*. Classification of Music Using Neural Net.* Presented at the 92th AES Convention. Vienna. March 24-27, 1992.
- KOSTEK, Bozena*. Parametric Representation of Musical Phrases.* Presented at the 101th AES Convention. Los Angeles. November 8-11, 1996.
- KOSTEK, Bozena*. Feature Extraction Methods for the intelligent processing of musical signals.* Presented at the 99th AES Convention. New York. October 6-9, 1995.
- PALOMAKI, Kalle, *Neural Network Approach to Analyze Spatial Sound.* Paper 16th AES International Conference. Espoo, Finlandia.
- SCHMIDMER, Keyh *.A combined measurement tool for the objective, perceptual based evaluation of compressed speech and audio signals.* Presented at the 106th AES Convention. Munich, May 8-11, 1999.
- SZCZERBA, Marek. *Recognition and Prediction of Music -A Machine Learning Approach.* Presented at the 106th AES Convention. Munich, May 8-11, 1999.
- OUDEYER, Pierre. *The Production and Recognition of Emotions in Speech: Features and Algorithms".* Paper Released in Science Direct Human Computer Studies, Sony CSL. Paris, Nov 30, 2002.
- PRADIER, Melanie. *Emotion Recognition from Speech Signals and Perception of Music.*Thesis Degree Paper, Stuttgart University.Germany.2011
- SIDOROVA,Julia. *Speech Emotion Recognition.*PhD Paper, Universitat Pompeu Fabra. España. Jul 4, 2007
- TREJOS, H. URIBE C. *Motor Computacional de Reconocimiento de Voz: Principios básicos para su Construcción.* Documento de Grado. Universidad Tecnológica de Pereira.Nov 2007
- ALDANA A. PIÑEROS J.*Desarrollo e Implementación de un Algoritmo de Reconocimiento de Voz que Permita Seleccionar una Imagen a Partir de un Banco de Nueve FotografíasUtilizando Redes Neuronales.* Documento de Grado, Universidad Sanbuenaventura. Bogotá, Oct. 22, 2010
- PANG, Yixiong. *Speech Emotion Recognition Using Support Vector Machine.* Paper, JiaoTong University. China. 2012
- MIYARA, Federico. *Pasos del algoritmo k-means.* [En línea]. Curso virtual de Procesamiento Digital de Señales en Voz. Facultad de Ciencias Exactas, Ingeniería y Agrimensura. Argentina. Disponible en la Web: "http://www.fceia.unr.edu.ar/prodivoz/Clustering_bw.pdf"
- DO, Min *.An Automatic Speaker Recognition System.* [En línea]. DSP Mini Project. Universidad de Illinois. USA. Disponible en la Web: "http://www.ifp.illinois.edu/~minhdo/teaching/speaker_recognition/"
- LI, Eldon. *Artificial neural networks and their business applications.* Paper, National Chung ChengUniversity.China.1994
- DELGADO, Alberto. *Aplicación de las Redes Neuronales en Medicina.*Paper, Universidad Nacional de Colombia.Colombia.1994

- Grupo de Investigación de Redes Neuronales Artificiales. *Perceptrón.* [En línea]. Universidad Carlos III. España. Disponible en la Web: "http://www.lab.inf.uc3m.es/~a0080630/redes-de-neuronas/perceptron-simple.html"
- SERGIU, Ciumac. *Feedforward Artificial Neural Network.* [En línea]. Universidad de Moldavia. Disponible en la Web: http://www.codeproject.com/Articles/175777/Financial-predictor-via-neural-network#conclusion "
- HUDSON, Martin. *Neural Network Toolbox User's Guide.* Reference.2011
- BURKHARDT, PAESCHKE. *A Database of German Emotional Speech.* Paper, Technical University of Berlin. Alemania. 2000
- RUEDA,E,TORRES,Y. *Identificación de emociones en la voz .*Thesis Degree Paper, Universidad Distrital de Santander. Colombia .2007.
- MORALES, M, ECHEVERRY, J, OROZCO, A. *Reconocimiento de emociones empleando procesamiento digital de la señal de voz.* Documento de Grado, Universidad Tecnológica de Pereira. Colombia.
- MONTERO, GUTIERREZ. *Analysis and modelling of emotional speech in spanish. Documento de Grado*, Universidad Politécnica de Madrid. España.
- SCHULLER, Bjorn, RIGOLL, Gerhard. *Hidden Markov model based speech emotion recognition. Thesis Paper*, Universidad Técnica de Munich. Alemania. 2003
- DARWIN, Charles. *The expression of the Emotions in Man and Animals.* Escocia. 1872
- Moving Pictures Experts Group *. Overview of the MPEG-4 Standard*. Norma. 2002.

ANEXO A

Anexos digitales.

A continuación de nombran los archivos que se encuentran en el medio magnético anexo a este documento.

A1. Tablas en Excel con los datos de las comparaciones realizadas.

- resultados.xlsx
- resultadoscolombia..xlsx

A2. Funciones implementadas para la creación, entrenamiento y prueba de los algoritmos de reconocimiento

- Libreria VOICEBOX
- Kannumfcc.m
- ZCRMoral.m
- PRUEBASINDIVIDUALES.m
- TESTCREATE.m
- TESTCREATE2.m
- TESTCREATEMELCEPST.m
- TESTCREATECOLOMBIA.m
- TRAININGCREATE.m
- TRAININGCREATE2.m
- TRAININGCREATE3.m
- TRAININGCREATEMELCEPST.m
- TRAININGCREATEMULTICLASS.m
- TRAININGCREATEMULTICLASSCOLOMBIA.m
- TRAININGCREATEMULTICLASSALEMAN.m
- MATRIZCREATEDELTA.m

A3. Redes neuronales entrenadas y utilizadas en el desarrollo del proyecto

- ENOJADONET1.mat
- ENOJADONET2.mat
- ENOJADONET3.mat
- ENOJADONET4.mat
- FELIZNET1.mat
- FELIZNET2.mat
- FELIZNET3.mat
- FELIZNET4.mat
- TRISTENET1.mat
- TRISTENET2.mat
- TRISTENET3.mat

- TRISTENET4.mat
- multiclassnetcol.mat
- multiclassnetcol2.mat
- multiclassal.mat
- multiclassal2.mat
- multiclasshib.mat

A4. Aplicación desarrollada en el software MATLAB.

- GUIRESULTADOS.FIG
- GUIRESULTADOS.m

A5. Archivos de audio, en español y en alemán, para los entrenamientos y pruebas realizadas.