TECNIACÚSTICA
CUENCA 2023

RECONOCIMIENTO DE EMOCIONES A TRAVÉS DE LA SEÑAL DE VOZ

Lorena Álvarez Pérez[1], *Álvaro Callejas Ramos*[1]
Fernando Blanco Albendea[2]

[1]Universidad Carlos III de Madrid, Leganés (Madrid), España
[2]Accenture Limited

RESUMEN

El reconocimiento de emociones es una herramienta muy atractiva debido al gran número de aplicaciones y casos de uso que ofrece. La utilización de algoritmos de aprendizaje automático supervisado permite a la máquina "aprender" a partir de las características de la voz humana, para posteriormente predecir la emoción del hablante en señales de audio desconocidas. En este trabajo se presentan dos sistemas de reconocimiento de emociones: uno para voz hablada (8 emociones) y otro para voz cantada (6 emociones). Asimismo, se incluye un estudio previo de las características de la voz humana y de los clasificadores (de aprendizaje máquina) utilizados, que mejor rendimiento presentan en la tarea de reconocimiento de emociones. Se han llevado a cabo experimentos utilizando la base de datos RADVESS (del inglés "*The Ryerson Audio-Visual Database of Emotional Speech and Song*") que contiene 7356 ficheros que corresponden a 24 actores (12 hombres y 12 mujeres) y se han considerado 4 clasificadores: perceptrón multicapa, Random Forest, XGBoost y el algoritmo *k*-NN. Se comprobará que, dependiendo del tipo de problema a resolver, será mejor un clasificador u otro, y la emoción peor clasificada también dependerá de si la señal de audio es voz hablada o voz cantada.

ABSTRACT

Speech emotion recognition refers to the process of predicting human emotions from audio signals using machine learning algorithms. It has become more prominent due to its large number of applications in areas such as, for instance, psychology, medicine, education and entertainment. Extracting relevant features from audio signals is a crucial task in this process aiming at correctly identifying emotions. In this paper, we proposed two automatic emotion recognizer systems: 1) it works for spoken speech (in this case, the speech recordings consist of 8 emotions) and 2) it works for song (in this case, the recording consist of 6 emotions). There are several databases for speech emotions testing. In this paper, we refer to the Ryerson Audio Visual Database of Emotional Speech and Song (RADVESS) because of its diversity and length. It comprises 7356 files and contains 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent. The following 4 classifiers have been the objects for the research: Multilayer Perceptron, Random Forest, XGBoost and the *k*-nearest neighbours algorithm. It will be shown that the classifier that best works will depend on the kind of the problem to be solved. In the same line of reasoning, the worst classified emotion will also depend on whether the audio signal is speech or song.

Palabras Clave— aprendizaje máquina, reconocimiento de emociones a través de la voz, clasificación por tramas y clasificación por pistas.

1. INTRODUCCIÓN

Las emociones son estados afectivos que todo ser humano experimenta de forma constante. Las emociones surgen cuando tiene lugar un suceso relevante, pero la relevancia de cada acontecimiento es subjetiva y no es posible predecir con exactitud lo que siente una persona en concreto simplemente por el estímulo que haya provocado ese acontecimiento. Por ejemplo, si una persona está viendo la televisión y escucha un golpe, su reacción y las emociones desencadenadas dependerán de cómo reciba el estímulo, del grado de interés por lo que estuviera viendo en ese instante en la pantalla, y de muchos otros factores, en su mayoría desconocidos a simple vista.

En los últimos años, las nuevas tecnologías y su interactividad en la vida del ser humano han pasado a formar parte de la normalidad. La aparición de nuevos algoritmos, aplicaciones y herramientas que buscan facilitar el día a día de las personas crece de manera exponencial [1]. La investigación sobre el reconocimiento de emociones continúa avanzando y es necesario mantener este progreso, no sólo para el desarrollo de aplicaciones de uso diario, sino también para expandir su empleo a otras facetas y profesiones, como, por ejemplo, la investigación de accidentes, la lucha contra el terrorismo, la educación, la medicina o la neurociencia.

La demanda de sistemas de detección de emociones efectivos continúa creciendo a medida que avanzan la investigación en procesos emocionales y el estudio de la neurociencia afectiva, y según aumenta el número de aplicaciones que facilitan la vida de las personas tanto en el ámbito personal o de ocio, como en el ámbito profesional en algunos casos. Una solución para hacer frente a la creciente demanda y a los

avances científicos es el desarrollo de un sistema de reconocimiento de la voz [2].

El objetivo principal de este trabajo es el diseño de dos sistemas de reconocimiento automático de emociones a través de la señal de voz. Cada sistema recoge la voz del hablante, analiza su estado emocional, y devuelve un parámetro de salida que refleja o indica su emoción. Esta predicción debe ser independiente del significado, el sentido y signos lingüísticos como símbolos, palabras o expresiones que utilice el hablante. El primer sistema utiliza señales de voz cantada, mientras que el segundo utilizará voz hablada. Este diseño requiere un estudio previo de las características de la voz humana y el sonido, y un análisis de los algoritmos y clasificadores de aprendizaje automático que permitan el reconocimiento de emociones a través de las características obtenidas.

El resto de este trabajo se divide tal y como se indica a continuación. La Sección 2 incluye un breve estado del arte en el que se detallan los fundamentos teóricos relativos a las emociones, el modelo de producción de voz y sus características, así como una descripción concisa del término o concepto: aprendizaje máquina. La Sección 3 describe el sistema de reconocimiento de emociones implementado en este trabajo, incluyendo una descripción de la base de datos, el procesado de la señal de voz, así como las características y clasificadores implementados. La Sección 4 resume los resultados obtenidos, y finalmente, en la Sección 5 se exponen las conclusiones extraídas, así como las futuras líneas de trabajo que quedan abiertas.

2. ESTADO DEL ARTE

2.1. Introducción

La señal de voz es el principal medio de comunicación entre los seres humanos, y se considera que contiene una gran cantidad de información sobre el estado emocional del emisor [3]. Sin embargo, la investigación sobre el reconocimiento de emociones ha estado principalmente centrada en el reconocimiento facial debido a las dificultades técnicas que implica un archivo de audio.

Las investigaciones actuales relacionadas con el reconocimiento de emociones se basan en los algoritmos de aprendizaje automático utilizados, las características de audio extraídas y el desarrollo de bases de datos orientadas al reconocimiento de emociones. Por lo tanto, la investigación sobre el reconocimiento automático de emociones en términos generales ya sea facial o a través de la voz, no engloba únicamente el campo del aprendizaje automático, sino también los ámbitos psicológico, lingüístico y biológico. En el terreno psicológico, las emociones son síndromes generados a raíz de diferentes tipos de eventos; en el campo de la lingüística, estas son cambios en la frecuencia fundamental, espectro, duración e intensidad de la voz; y por último, en el campo de la biología, son configuraciones del organismo en las cuales múltiples sistemas internos siguen un patrón que va evolucionando para hacer frente a una situación particular de forma eficiente. De este modo, para poder progresar en el desarrollo de estas aplicaciones, es necesario avanzar en los campos de investigación mencionados [4].

2.2. Fundamentos de las emociones

Se ha considerado que las emociones de los seres humanos han ido cambiando a lo largo de la evolución de la especie. No obstante, la realidad es que las emociones no son características por sí mismas, y surgen como combinaciones de algunas de ellas. Esto hace de la investigación una tarea muy subjetiva, y por este motivo, los investigadores trabajan con emociones características para poder diferenciar unas de otras.

Muchos investigadores están de acuerdo en que las emociones y su interpretación pueden estar condicionados por los fenómenos sociales, expectativas, normas y el entorno en el que vive cada persona. Diferentes sociedades tratan de forma distinta algunas emociones como el amor, la ira o la vergüenza.

En 1972, el psicólogo estadounidense Paul Ekman, se opuso a este planteamiento, y una gran parte de la comunidad científica, hoy en día, sigue cuestionando la validez de sus hallazgos. Ekman descubrió que existen 6 expresiones básicas y universales: alegría, asco, enfado, miedo, sorpresa y tristeza. En la década de los 90, Ekman incluyó una lista de emociones secundarias que surgen a raíz de la combinación de dos o más emociones básicas, y que no siempre se identifican con una expresión facial. Esta lista de emociones se muestra en la Tabla 1 [5] y son las más utilizadas en experimentos de reconocimiento de emociones.

Tabla 1. Lista de emociones primarias y secundarias según Paul Ekman.

Emociones primarias	Emociones secundarias
Alegría, asco, enfado, miedo, sorpresa y tristeza	Alivio, bochorno, culpa, desprecio, diversión, emoción, orgullo, satisfacción y vergüenza

2.3. Producción de la voz y el habla

La voz es el sonido producido por cualquier especie cuando el aire es expulsado por los pulmones a través de la laringe y hace que vibren las cuerdas vocales. Surgió a raíz de la necesidad de comunicarse del ser humano, y su uso no se limita a la transmisión de mensajes orales, sino que transmite el estado emocional de una persona a través de sus características, independientes del contenido del mensaje.

Sin embargo, el habla no es solamente la emisión del sonido, e incluye su modificación mediante los órganos resonadores y la emisión de las palabras mediante los órganos articuladores. Por un lado, los órganos resonadores son la faringe, la cavidad bucal y las fosas nasales, cavidades en las que el sonido que sale a través de la laringe resuena y se amplifica. Por otro lado, los articuladores son los órganos encargados de convertir el sonido en fonemas. Estos son los labios, los dientes y la lengua [6].

A continuación, se expone el modelo de producción de la voz humana, desde que el aire sale de los pulmones hasta que este sale por los labios, en función del tipo de sonido emitido. Los sonidos emitidos por el ser humano pueden ser tonales, si estos son producidos por la oscilación relajada de las cuerdas vocales, como las vocales y varias consonantes, o no tonales, si durante su emisión las cuerdas vocales permanecen abiertas y el aire expulsado a cierta velocidad produce una turbulencia. En la producción de sonidos tonales, el flujo de aire saliente de los pulmones incrementa la presión y las cuerdas vocales se separan, permitiendo que el aire fluya. Este flujo de aire disminuye la presión, permitiendo que las cuerdas vocales se vuelvan a cerrar.

2.4. Fundamentos del aprendizaje automático

La Inteligencia Artificial (IA) es el término más genérico para hacer referencia al campo de la informática centrado en la creación de programas capaces de mostrar comportamientos considerados "inteligentes". Fue acuñado en el año 1956 por el informático John McCarthy, que la definió como "*la ciencia y el ingenio de hacer máquinas inteligentes*", en la conferencia de Dartmouth [7]. Engloba numerosas técnicas, como algoritmos de búsqueda, estadística, análisis predictivo o el aprendizaje automático (también conocido como *machine learning*, de su nombre en inglés).

El aprendizaje automático, concretamente, es una práctica que consiste en el uso de algoritmos para analizar datos, aprender de ellos y utilizarlos para hacer una predicción o estimación sobre algo externo. A su vez, el aprendizaje automático engloba el aprendizaje profundo, que consiste en utilizar algoritmos para "imitar" el funcionamiento del cerebro humano en el procesado de datos y en su uso para la generación de patrones y toma de decisiones.

Desde el año 2010 hasta la actualidad, los "gigantes informáticos" han desarrollado sus propias herramientas de aprendizaje automático. Algunos ejemplos son Google y la Universidad de Stanford con "GoogleBrain", IBM con "Watson" o Microsoft con "Kinect" [8].

3. SISTEMA DE RECONOCIMIENTO DE EMOCIONES

3.1. Base de datos

La base de datos utilizada para el diseño de los diferentes clasificadores de emociones, y la investigación llevada a cabo en este trabajo, es la denominada "Ryerson Audio-Visual Database of Emotional Speech and Song" (RAVDESS) [9]. Este conjunto de datos se puede descargar de forma gratuita y contiene un total de 7356 registros, cada uno de los cuales, ha sido sometido a 10 evaluaciones en tres aspectos: intensidad, autenticidad y validez emocional.

Los datos utilizados son pistas de audio de duración 3-5 segundos, interpretadas por 24 actores (12 hombres y 12 mujeres) de Toronto (Canadá), que vocalizan dos oraciones diferentes, pero cuyo léxico está emparejado, en un inglés con acento norteamericano neutro para no afectar a los experimentos. El rango de edad de los actores es 21-36 años, siendo la media, 26 años. El formato de las muestras es audiovisual, pero por el alcance del trabajo, únicamente se utiliza el audio.

Las oraciones son: "*Kids are talking by the door*" / "*Dogs are sitting by the door*". Ambas oraciones tienen 7 sílabas y están emparejadas en cuanto a frecuencia y familiaridad. Cada actor tiene un total de 44 pistas de audio cantado reflejando 6 emociones distintas y 60 pistas de audio hablado, siendo 8 el número de emociones reflejado en este caso. Las emociones analizadas son: neutro, calma, alegría, tristeza, enfado, miedo, disgusto y sorpresa. Estas dos últimas emociones sólo se grabaron en formato hablado, por lo que la clasificación de emociones en las dos modalidades se realiza por separado.

Cada uno de los actores tiene pistas expresando cada emoción con dos intensidades distintas: normal y fuerte. Por cada emoción, actor y modalidad (cantada o hablada), son 4 pistas con intensidad normal y otras 4 pistas con intensidad fuerte, exceptuando la emoción neutra, para la cual no hay nivel de intensidad fuerte y, por tanto, sólo incluye 4 pistas por actor. Cabe que no se distinguirá intensidad de emociones durante las clasificaciones.

Con esto en mente, a continuación, se resume el número de pistas para cada modalidad explorada en este trabajo:

- Voz hablada: 1440 pistas
- Voz cantada: 1012 pistas

La frecuencia de muestreo es 16 kHz y se utilizan 16 bits para la codificación. Como se ha comentado, las pistas de audio tienen una duración de 3-5 segundos, de los cuales, aproximadamente los primeros y últimos 0,5 segundos son silencio. Estos valores son ruido y se eliminan antes de proceder a la extracción de características, ya que no aportan información al clasificador y reducen la precisión de este.

3.2. Extracción de características

A continuación, se describe cómo se transforma la señal de audio (ya sea voz hablada o voz cantada) en una serie de variables o características que representa la información del

sonido y que se utilizan como entrada al clasificador (en lugar de la señal de audio).

3.2.1. MFCCs

Los MFCCs (de su nombre en inglés, *Mel Frequency Cepstral Coefficients*) [10] son un tipo de coeficientes cepstrales utilizados por músicos y productores para modelar audio y música. Durante un largo tiempo, ha sido la característica más utilizada para el reconocimiento de voz debido a su habilidad para representar el espectro de amplitud de la voz de forma compacta. Estos coeficientes derivan de la aplicación del Cepstrum sobre una ventana de tiempo de la señal de voz. El Cepstrum es un operado que, transformando una convolución temporal en una suma en el dominio espectral, consigue extraer la excitación y el tracto vocal de la señal.

Para el cálculo de los coeficientes de Mel, se siguen los siguientes pasos:

- Pre-énfasis: la señal de audio se pasa por un filtro de pre-énfasis para compensar la atenuación de -20 dB/década que produce el mecanismo humano para la producción del habla.
- Enventanado: la señal de voz es un proceso aleatorio y no estacionario, lo que supone un inconveniente para su análisis. Si se tiene en cuenta que es estacionaria a muy corto plazo (~ms), es posible obtener segmentos muy cortos en el tiempo y solapados entre sí, que sí son estacionarios. En este trabajo, se utiliza una ventana de tipo Hanning con una duración de 20 ms, y un solapamiento entre ventanas de 10 ms. De este modo, se obtienen coeficientes cada 10 ms.
- Cálculo de la Transformada Discreta de Fourier (DFT). A partir de este momento, se trabaja sólo con la amplitud del espectro y se descarta la fase.
- Banco de filtros de Mel. La señal se multiplica por un banco de filtros triangulares de área unidad, espaciados según las frecuencias de la escala Mel. En este trabajo se han utilizado 40 filtros. Tras multiplicar el módulo de la DFT por el banco de filtros Mel, se calcula la energía correspondiente en cada uno de los filtros.
- Logaritmo de la señal transformada. Se calcula el logaritmo de la energía.
- Cálculo de la Transformada Discreta del Coseno.

3.2.2. ΔMFCCs

Asimismo, se han calculado los coeficientes Delta-MFCC (o ΔMFCCs), conocidos como coeficientes de velocidad. Estos coeficientes contienen información sobre las trayectorias de los MFCCs con respecto al tiempo.

Es conveniente utilizar estas características junto con los MFCCs porque las señales de voz varían en el tiempo y son un flujo constante. Aunque en la lingüística, se describa la voz como una "concatenación de secuencia de fonemas", una descripción más precisa es una "secuencia de transiciones entre fonemas". Estos nuevos coeficientes no son muy precisos, ya que las derivadas tienden a amplificar el ruido blanco y la salida es más ruidosa, pero son muy fáciles de calcular y ofrecen un claro beneficio al concatenarse con los MFCCs instantáneos [11].

3.2.2. Frecuencia fundamental

La frecuencia fundamental (o también conocida como "*pitch*") es una característica interesante por ser una de las percepciones básicas del sonido y una variable con mucha información emocional. Dicha frecuencia es la más baja del espectro de frecuencias, y las frecuencias dominantes pueden expresarse como múltiples de esta.

Esta característica se obtiene a partir de las tramas obtenidas tras el enventanado del proceso de extracción de los MFCCs. Posteriormente se calcula la DFT, seguida de una interpolación cuadrática de la magnitud espectral. Esta es una herramienta muy potente para estimar la frecuencia instantánea cerca de un pico en el espectro. Consiste en aproximar la forma de un espectro cerca de un pico mediante una parábola. Si bien es cierto que en este caso se utiliza una ventana de Hanning, esta aproximación está justificada para cualquier tipo de ventana [12].

3.3. Clasificadores

A continuación, se expone cada uno de los clasificadores utilizados para los experimentos, así como sus parámetros más importantes.

3.2.1. Perceptrón Multicapa (MLP)

El MLP [13] es un tipo de red neuronal artificial utilizado, mayormente, para clasificación, que combina neuronas artificiales en una capa de entrada, una o más capas ocultas, y una capa de salida. La dimensión de la capa de entrada es igual al número de características del conjunto de datos, la dimensión de la capa de salida es igual al número de etiquetas o clases, y las neuronas de capas contiguas se conectan entre sí para conformar la red neuronal.

En los experimentos llevados a cabo en este trabajo, se han utilizados MLP con una única capa oculta, cuyo número de neuronas se ha obtenido por validación cruzada.

3.2.2. Random Forest (RF)

RF es un clasificador muy útil para problemas multiclase por su alta capacidad de generalización. Se trata de una combinación de árboles predictores, cada uno de los cuales depende de un vector aleatorio con la misma distribución [14].

En el clasificador utilizado en este trabajo, los árboles predictores utilizan la impureza de Gini para evaluar la calidad de la división. Asimismo, se han ajustado el valor de 2 hiperparámetros, utilizando validación cruzada, con objeto

de reducir el coste computacional y evitar el sobreajuste del modelo. El primero es la profundidad máxima de cada árbol y el segundo es el número de árboles o estimadores, ya que utilizar más árboles de los necesario no implica una mejor eficiencia del clasificador.

3.2.3. XGBoost

XGBoost (de su nombre en inglés, "*eXtreme Gradient Boosting*") [15] es uno de los algoritmos dominantes en el campo del aprendizaje automático. Se sitúa entre los clasificadores de "aprendizaje conjunto", al igual que RF, ya que utiliza múltiples modelos generados y combinados estratégicamente con el objetivo de resolver un problemaconcreto, obteniendo un mejor rendimiento predictivo que el que se podría obtener utilizando cualquiera de los modelos por separado.

En este tipo de clasificadores, se van añadiendo secuencialmente nuevos árboles predictores para corregir los errores cometidos por los árboles ya existentes hasta que no haya más mejoras que hacer. De forma muy sencilla, se puede decir que, en la secuencia de creación de árboles, en el instante t, se asigna un peso menor a las salidas de las muestras clasificadas correctamente en la iteración $t-1$, y se asignan pesos mayores a las salidas erróneas.

Los parámetros principales que ajustar son la profundidad máxima de los árboles y el número de árboles que forman el conjunto. En ambos casos, los valores óptimos se han obtenido a través de validación cruzada. Otro parámetro que afecta al rendimiento del modelo es la tasa de aprendizaje η, que indica la medida utilizada para reducir los pesos después de cada iteración en el proceso de *boosting*. En los experimentos llevados a cabo, se ha comprobado empíricamente que $\eta = 0{,}3$ es el valor que mejor resultados proporciona.

3.2.4. k-NN

El clasificador *k*-NN (del inglés, *k-Nearest Neighbours*) [16] es uno de los clasificadores más sencillos e intuitivos y se aplica en una gran variedad de ámbitos. En este caso, cada muestra del conjunto de test se clasifica según la clase a la que pertenezcan las *k* muestras más cercanas a esta y, por lo tanto, las muestras del conjunto de entrenamiento se utilizan directamente en la fase de test.

En este caso, se ha ajustado el valor del hiperparámetro *k* que representa el número de vecinos e influye en el rendimiento del modelo.

3.4. Sistema de clasificación de emociones

La Figura 2 muestra un diagrama del sistema de reconocimiento de emociones desarrollado en este trabajo, desde la pista de audio original hasta la clasificación de la emoción correspondiente a dicha pista o a cada una de las tramas que la forman, según el tipo de clasificación.

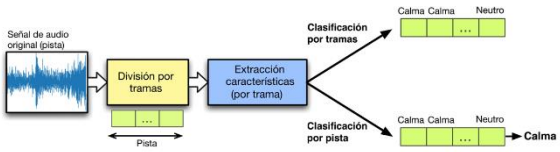

Figura 2. Diagrama del sistema de reconocimiento de emociones.

4. RESULTADOS

Para las pistas de voz cantada, se han realizado dos tipos de clasificación: por tramas y por pistas. Sin embargo, para el conjunto de pistas de voz hablada, únicamente se ha llevado a cabo una clasificación por pistas debido al alto coste computacional, teniendo en cuenta que son un total de 1440 pistas (442513 tramas).

4.1. Experimentos con voz cantada

4.1.1. Clasificación por tramas

La Tabla 2 resume los resultados obtenidos utilizando únicamente los MFCCs y ΔMFCCs. Como se puede observar, el clasificador con mayor precisión es XGBoost.

Tabla 2. Clasificación por tramas para voz cantada.

Clasificador	Tasa de error (%)
MLP	$33{,}4 \pm 2{,}0$
RF	$33{,}9 \pm 1{,}0$
XGBoost	$\mathbf{31{,}7 \pm 1{,}0}$
k-NN	40

4.1.2. Clasificación por pistas

La Tabla 3 resume los resultados obtenidos utilizando únicamente los MFCCs y ΔMFCCs. En este caso, el clasificador con mayor precisión es el MLP. Es común que la clasificación por pistas aporte mejores resultados que la clasificación por tramas, ya que en este caso un error en la clasificación de una trama perteneciente a una pista correctamente clasificada no se tiene en cuenta, y no aumenta por tanto la probabilidad de error.

Tabla 3. Clasificación por pistas para voz cantada.

Clasificador	Tasa de error (%)
MLP	$19{,}8 \pm 3{,}0$
RF	$34{,}95 \pm 2{,}58$
XGBoost	$\mathbf{26{,}43 \pm 2{,}96}$
k-NN	31,53

4.2. Experimentos con voz hablada

La Tabla 4 recoge los resultados obtenidos para voz cantada. En este caso, el clasificador con mayor precisión es el MLP, pero la tasa de error aumenta considerablemente respecto a los experimentos con voz cantada. Esto se debe principalmente al aumento de emociones en las clases.

Tabla 4. Clasificación por pistas para voz hablada

Clasificador	Tasa de error (%)
MLP	$30{,}2 \pm 4{,}08$
RF	$44{,}86 \pm 3{,}12$
XGBoost	$35{,}17 \pm 1{,}20$
k-NN	36,77

5. CONCLUSIONES Y FUTURAS LÍNEAS

En este trabajo se han diseñado dos sistemas de reconocimiento automático de emociones a través de la voz. Ambos sistemas utilizan la base de datos RADVESS.

Las características extraídas han sido los 13 primeros MFCCs y sus correspondientes coeficientes de velocidad o ΔMFCCs, además de la frecuencia fundamental. Se ha explorado el uso de cuatro clasificadores: MLP, RF, XGBoost y el algoritmo *k*-NN. Para todos ellos, se han ajustado sus hiperparámetros más importantes a través de validación cruzada.

El primer sistema es un clasificador para voz cantada en el que se distinguen 6 emociones. Se ha llevado a cabo dos tipos de clasificación: por tramas y por pistas. Los resultados obtenidos han demostrado que es mejor una clasificación por pistas, siendo el clasificador XGBoost el que mejor resultados proporciona.

El segundo sistema es un clasificador para voz hablada en el que se distinguen 8 emociones. En este caso, el clasificador que mejor resultados proporciona es el MLP.

Analizando las matrices de confusión de los resultados en ambos sistemas, se ha observado que la emoción peor clasificada es la emoción neutra, por su similitud con calma y tristeza.

Finalmente, se indican las posibles futuras líneas de investigación:

- Técnicas de aprendizaje profundo
- Sistemas de clasificación para voz cantada y hablada
- Sistema de clasificación jerárquico

6. REFERENCIAS

[1] M. Jalife. “Exponencial crecimiento de patentes de inteligencia artificial”. En *El Financiero* (febrero de 2019). Último acceso: septiembre de 2023. URL: https://www.elfinanciero.com.mx/opinion/mauricio-jalife/exponencial-crecimiento-de-patentes-de-inteligencia-artificial/

[2] J. Lope, and M. Graña, “An ongoing review of speech emotion recognition”, *Neurocomputing*, vol. 528, no. 1, pp. 1-11, 2023.

[3] N. Morán, J. Pérez, and W. Rodríguez, “Reconocimiento de Estados Emocionales de Personas Mediante la Voz Utilizando Algoritmo de Aprendizaje de Máquina”, *Revista Venezolana de Computación*, vol. 5, pp. 41-51, 2018.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis S. Kollias, W. Fellenez, and J. G. Taylor, “Emotion-recognition in human-computer interaction”, *IEEE Signal Processing Magazine*, vol. 18, no.1, pp. 32-80, 2001.

[5] P. Ekman, and H. Oster, “Expresiones faciales de la emoción”, *Annual Review of Psychology*, vol. 30, pp. 527-554, 1979.

[6] R. Dosal González. “Producción de la voz y el habla. La fonación”. Trabajo Fin de Grado (pp. 3-4). Universidad de Cantabria, 2013. URL: https://repositorio.unican.es/xmlui/bitstream/handle/10902/5583/DosalGonzalezR.pdf

[7] J. Moor, “The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years”, *AI Magazine*, vol. 27, pp. 87-91, 2006.

[8] M. Velázquez, “Historia y evolución del Machine Learning”. En *RecluiT, atracción de* talento. Último acceso: septiembre de 2023. URL: https://recluit.com/historia-y-evolucion-del-machine-learning/

[9] R. S. Livingstone, and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RADVESS) [Data set]. En *PLoS ONE (1.0.0)*, vol. 13, no. 5, p. e0196391. Zenodo. 2018.

[10] B. Logan, “Mel Frequency Cepstral Coefficients for Music Modeling”, in Proc. 1st Int. Symposium Music Information Retrieval (Plymouth, Massachusetts), 2000.

[11] T. Bäckström, “Introduction to Speech Processing. Deltas and Delta-deltas”. En *Aalto University Wiki* (2019). Último acceso: agosto de 2023. URL: https://wiki.aalto.fi/display/ITSP/Deltas+and+Delta-deltas

[12] P. Magron, R. Badeau, and D. Bertrand, “Phase recovery by unwrapping: applications to music signal processing”, 2016.

[13] F. Murtagh, “Multilayer perceptrons for classification and regression”, *Neurocomputing*, vol. 2, no. 5-6, pp. 183-197, 1991.

[14] L. Breiman, “Random Forests”, *Machine Learning*, vol. 45, pp. 111-132, 2001.

[15] T. Chen, and C. Guestrin, “XGBoost: A Scalable Tree Boosting System”. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, USA), pp. 785-794, 2016.

[16] P. Cunningham, and S. Delany, “k-Nearest neighbour classifiers”, *Mult Classif System*, vol. 54, 2007.