Reconocimiento de Estados Emocionales de Personas Mediante la Voz Utilizando Algoritmos de Aprendizaje de Máquina

Nerio Morán
Ingeniería de Sistemas, LaSDAI
Universidad de Los Andes
Mérida, Venezuela
neriojmoran@ula.ve

Jesús Pérez
Ingeniería de Sistemas, LaSDAI
Universidad de Los Andes
Mérida, Venezuela
jesuspangulo@ula.ve

Wladimir Rodriguez
Ingeniería de Sistemas, Dpto. Computación
Universidad de Los Andes
Mérida, Venezuela
wladimir@ula.ve

***Resumen*— El reconocimiento de estados emocionales de las personas se ha popularizado en aras de mejorar las interacciones entre personas y robots. Actualmente, los investigadores han mostrado un creciente interés por desarrollar técnicas que permitan reconocer emociones a través de la voz. Las técnicas más populares para reconocer emociones mediante la voz, utilizan bases de datos con registros de voz de diferentes personas que expresan diferentes emociones, para entrenar algoritmos de aprendizaje de máquina. Particularmente, las emociones humanas pueden ser expresadas de diversas maneras, lo cual afecta la capacidad de reconocimiento de estos algoritmos, y en consecuencia, la capacidad de interacción eficaz de los robots, ya que reconocer todas las formas de expresión de una misma emoción a través de la voz es una tarea compleja. En este sentido, en aras de proporcionar la capacidad a los robots de reconocer emociones de un amplio grupo de personas, en esta investigación se construye una base de datos en condiciones controladas y actuadas de seis emociones (ira, sorpresa, felicidad, miedo, tristeza y asco). Luego, con el propósito de hacer comparaciones, se entrenan tres modelos de aprendizaje automático (Máquinas de Vectores de Soporte, Bosques Aleatorios y Aumento del Gradiente). Posteriormente, se construyen dos bases de datos adicionales (una en condiciones controladas y semi-naturales, y otra en condiciones no controladas y naturales) para probar con mayor rigurosidad los modelos entrenados. Los resultados obtenidos indican que la mejor tasa de reconocimiento se obtiene cuando se hacen predicciones sobre muestras capturadas en la mismas condiciones que las muestras de la base de datos de entrenamiento, y además, para muestras pertenecientes a las otras bases de datos hay resultados prometedores, como por ejemplo, la alta tasa de reconocimiento de la ira en todas las pruebas realizadas.**

Palabras Clave—Emociones, Reconocimiento, Aprendizaje de Máquina, Voz.

I. Introducción

A lo largo de los años, ha sido el ser humano quien se ha adaptado a las diferentes formas de comunicación que ofrecen las computadoras. Investigaciones actuales, están dirigidas por la iniciativa de disminuir la brecha de comunicación entre personas y robots. Para ello, algunos de los aspectos que se consideran son el reconocimiento y adaptación de las computadores según el estado emocional de la persona [1].

El reconocimiento de emociones es realizado mediante diferentes medios, tales como: la voz [2]–[12], imágenes de rostros [13], conductancia de la piel [14], frecuencia cardíaca [15], señales inalámbricas [16], entre otros. Dado que las señales de la voz se consideran fáciles de obtener y es una de las formas de comunicación más usadas, se le considera como una de las fuentes de información más adecuadas para la clasificación de emociones [5].

Dentro de las aplicaciones más resaltantes de los algoritmos de aprendizaje de máquina aplicados al reconocimiento de emociones mediante la voz, están los robots sociales de asistencia personal [17], con la capacidad de detectar emociones y regularlas. El objetivo de estos robots es mantener el bienestar del estado afectivo de las personas ubicadas en un entorno inteligente. Cada robot cuenta con dos componentes principales de reconocimiento: voz y expresiones faciales; los cuales usa de manera conjunta para determinar los estados afectivos y regularlos en caso de ser necesario.

Para entrenar los algoritmos de aprendizaje de máquina, han sido utilizadas múltiples bases de datos, y dentro de las más populares se encuentran: "A Database of German Emotional Speech", también conocida como Emo-DB [18], "Polish Emotional Speech Database" [19], "The eNTERFACE'05 audio-visual emotion database" [20], "Surrey Audio-Visual Expressed Emotion", también conocida como SAVEE [21], entre otras. Estas bases de datos cuentan con múltiples muestras de audio en un idioma específico, etiquetadas con distintos estados emocionales, que son procesadas para extraer diferentes características y servir como entrada a los algoritmos de clasificación.

Las investigaciones actuales, han sugerido la extracción de numerosas combinaciones de características de las señales de audio; dentro de éstas, las más populares para el reconocimiento de emociones son: tono, energía, los Coeficientes Ceptrales de las Frecuencias de Mel (MFFCs, por sus siglas en inglés), los Coeficientes Dinámicos de Energía de Mel (MEDC, por sus siglas en inglés) y los formantes [2], [4]–[12].

Gran parte de las investigaciones relacionadas al reconocimiento de emociones a través de la voz, no son rigurosas en las pruebas que le hacen a los modelos entrenados, gran parte de estas, alcanzando una tasa de reconocimiento considerablemente elevada, no obstante, estos resultados

suelen estar sobre-entrenados y muy pocas veces el modelo es sometido a pruebas utilizando un conjunto más amplio. Dado que el objetivo principal es reconocer emociones en un amplio grupo de personas, muchas consideraciones deben tomarse, principalmente con los criterios utilizados para construir la base de datos de entrenamiento. Esto se debe, a que las bases de datos presentan diversas problemáticas con respecto a la diversidad de las expresiones humanas y las diversas características asociadas a la población utilizada para grabar las emociones. A pesar de que existen bases de datos orientadas al reconocimiento de emociones en español [22]–[24]; en esta investigación, se diseñará y se establecerán diferentes criterios de construcción para 3 bases de datos distintas. Esto, en aras de mantener un mayor control sobre las condiciones de ambiente y de las características asociadas a la población que formará parte de la base de datos. Estas bases de datos permitirán probar con rigurosidad los modelos entrenados, y de esta forma realizar una comparativa entre el desempeño de los algoritmos de aprendizaje de máquina seleccionados. En concordancia con [4] se utilizaran MFCCs y la energía, como parte de las características que se extraen del conjunto de datos. Además, se utilizaran 3 algoritmos de aprendizaje supervisado distintos: Bosques Aleatorios (RF, por sus siglas en inglés), Aumento del Gradiente (GB, por sus siglas en inglés) y Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés).

El documento se organiza de la siguiente manera: la segunda sección es una descripción de los antecedentes que se usaron como parte de la investigación; la tercera sección explica de manera breve los procesos involucrados en la clasificación de las emociones (construcción de las bases de datos, procesamiento de datos y entrenamiento); la cuarta sección muestra los resultados; la quinta sección muestra una discusión; y la sección final muestra las conclusiones y trabajos futuros de esta investigación.

II. Antecedentes

Uno de los contenidos de mayor disponibilidad y de mayor uso en la actualidad, son los archivos de audio. La voz, es el principal medio de comunicación en los seres humanos y como componente para-verbal de la comunicación [25], se considera que contiene mucha información sobre el estado emocional de la persona que la emite. Las investigaciones relacionadas al reconocimiento de emociones mediante la voz, se basan en la extracción de características del audio para obtener una representación matemática. Esta representación, es utilizada para entrenar los algoritmos de aprendizaje de máquina y de esta manera realizar clasificaciones. Gran parte de las investigaciones se basan en la precisión o tasa de reconocimiento de los clasificadores, enfocándose en 3 aspectos: algoritmos de aprendizaje de máquina utilizados, características extraídas del audio y bases de datos empleadas. Los antecedentes de esta investigación se dividen en dos: reconocimiento de emociones, haciendo énfasis en los algoritmos de aprendizaje de máquina, características extraídas, bases de datos y tasas de reconocimiento; y bases de datos orientadas al reconocimiento de emociones, haciendo énfasis en los criterios utilizados para su construcción y aspectos relevantes adicionales.

A. Reconocimiento de Emociones

En la investigación [26], utilizan 6 tipos de clasificadores para comparar la tasa de reconocimiento en la predicción de las 6 emociones universales (ira, sorpresa, felicidad, miedo, tristeza y asco) [27], utilizando como entrada la voz. La base de datos utilizada en esa investigación se llama eNTERFACE'05 [20], la cual es una base de datos audio-visual, que contiene muestras de las 6 emociones mencionadas anteriormente. Cada video es convertido en formato WAV utilizando la herramienta MATLAB. Las características extraídas del audio fueron las siguientes: los Coeficientes Cepstrales de la Frecuencia de Mel (MFCCs, por sus siglas en inglés), coeficientes de Predicción Lineal Cepstral (LPC, por sus siglas en inglés), método de los momentos, segundo método de los momentos, centroide espectral, punto de caída espectral, flujo espectral, compacidad, variabilidad del centroide espectral, media cuadrática, fracción del marco de baja energía, tasa de cruces por cero, frecuencia máxima mediante la tasa de cruces y la transformada discreta de Fourier. Los algoritmos de aprendizaje supervisado utilizados fueron los siguientes: SVM lineal y polinomial, árboles de decisión, redes neuronales, redes bayesianas, algoritmo de los k-vecinos más cercanos y Bayes ingenuo. Los resultados obtenidos mostraron que el árbol de decisión obtuvo la mejor tasa de reconocimiento la cual fue de 96.21%.

Uno de los clasificadores de mayor popularidad en el área de reconocimiento de emociones a través de la voz es el SVM. La principal motivación se debe a que el SVM ha demostrado ser uno de los algoritmos que mayor tasa de reconocimiento tiene cuando se trata de pruebas dependientes e independientes de la persona. En la investigación [7], se utiliza el clasificador SVM con 4 núcleos distintos. El conjunto de datos utilizado es la base de datos Pocala [19] y las emociones utilizadas fueron: ira, miedo, felicidad, tristeza y aburrimiento. Las características seleccionadas fueron: el tono, los formantes, la tasa de cruces por cero, los MFFCs, y parámetros estadísticos. Se utilizaron los núcleos: lineal, cuadrático, radial y polinomial. Los resultados obtenidos mostraron que el SVM con núcleo radial obtuvo la mejor tasa de reconocimiento del 84%. Como conclusión se obtuvo que los núcleos lineal y cuadrático tienen una mejor tasa de reconocimiento en las emociones: ira, miedo y tristeza. A diferencia, el núcleo polinomial tiene la peor tasa de reconocimiento.

Un factor que influye en el desempeño del algoritmo son los hiper-parámetros. En la investigación [6], utilizan la base de datos Emo-DB [18]. Se utilizan 5 emociones: la ira, la tristeza, la alegría, la neutralidad y el miedo. Las características seleccionadas del audio son: Los MFCCs y los Coeficientes del Espectro Dinámico de la Energía de Mel (MEDC, por sus siglas en inglés). La tasa de reconocimiento obtenida utilizando un clasificador SVM con núcleo radial, fue de 93.75%. En comparación a la investigación [4], a pesar de que las mismas bases de datos fueron usadas, el cambio de las características y una selección óptima de los hiper-parámetros, produjo que

el SVM radial obtuviera una mejor tasa de reconocimiento con respecto a los árboles de decisión.

TABLA I: TASA DE RECONOCIMIENTO PARA DIFERENTES CLASIFICADORES

Ref.	Modelo	N°clases	% Exactitud	Base de Datos
[2]	GSVM	4	67.1%	Susas [28]
[2]	HMM	4	70.1%	Susas [28]
[2]	HMM	2	96.3%	Susas [28]
[2]	GSVM	5	42.3%	Aibo [29]
[4]	Rand-SVM	7	55.89%	Emo-DB [18]
[4]	RF	7	81.05%	Emo-DB [18]
[4]	GB	7	65.23%	Emo-DB [18]
[5]	MCP NN	2	85%	Sin nombre
[6]	RBF-SVM	5	93.75%	Emo-DB [18]
[7]	RBF-SVM	6	84%	Polish-DB [19]
[8]	MLP	7	83.1%	Emo-DB [18]
[8]	RF	7	77.19%	Emo-DB [18]
[8]	PNN	7	94.1%	Emo-DB [18]
[8]	SVM	7	83.1%	Emo-DB [18]
[9]	RBF-SVM	7	86.6%	Emo-DB [18]
[12]	SVM	3	91.30%	Emo-DB [18]
[12]	SVM	3	95.09%	SJTU-DB [12]

Investigaciones han explorado el uso de diferentes clasificadores como: SVM con núcleo de función base radial (RBF-SVM, por sus siglas en inglés), SVM con núcleo gaussiano (GSV, por sus siglas en inglés), modelo oculto de Márkov (HMM, por sus siglas en inglés), Bosques Aleatorios (RF, por sus siglas en inglés), aumento del gradiente (GB, por sus siglas en inglés), modelo de red neuronal de McCulloch y Pits (MCP-NN, por sus siglas en inglés), perceptrón multicapa (MLP, por sus siglas en inglés), red neuronal probabilística (PNN, por sus siglas en inglés), entre otros. El número de clases y la base de datos varía según la investigación. En la Tabla I se muestra de manera sintetizada los resultados de los trabajos relacionados.

En esta investigación se usaran 2 de las características más populares en las investigaciones actuales: la Energía y los Coeficientes Cepstrales de Mel. Bajo el interés de realizar una comparativa del desempeño de los clasificadores sobre muestras en diferentes condiciones, se utilizarán 3 algoritmos de aprendizaje de máquina: Bosques aleatorios (RF), Máquinas de Vectores de Soporte (SVM) y Aumento del Gradiente (GB).

B. Bases de Datos

A pesar de que existen muchas investigaciones que logran una gran precisión reconociendo emociones [6] [8] [12], en la práctica estos clasificadores no suelen tener el mismo desempeño. Esto se debe a muchos factores, tales como las diferencias que existen entre las personas que forman parte del conjunto de entrenamiento y el conjunto de prueba, por ejemplo: las condiciones de ambiente, los dispositivos utilizados, diferencias culturales, el idioma, la edad, entre otros. Existen otros factores que afectan la tasa de reconocimiento en las emociones, por ejemplo el desbalance en las bases de datos, la calidad de las emociones capturadas, la diversidad de sentencias, el número de emociones, entre otros. Muchos son los criterios utilizados para diseñar una base de datos orientada al reconocimiento de emociones, entre las características más relevantes se encuentran: número de personas, origen de las personas, idioma utilizado, declaraciones utilizadas, entre otros. A continuación se presentan las bases de datos más populares orientadas al reconocimiento de emociones.

En la investigación [18], se presenta una base de datos conocida como Emo-DB. Esta base de datos, fue construida utilizando 10 actores (5 mujeres y 5 hombres), simulando o actuado emociones. Las declaraciones seleccionadas conforman un conjunto de oraciones (5 cortas y 5 largas) usadas diariamente e interpretables en todas las emociones aplicadas. Las grabaciones fueron realizadas en una cámara anecoica, con equipo de grabación de alta calidad. La base de datos consistió en 800 sentencias, en las cuales están contenidas 7 emociones: neutralidad, ira, miedo, alegría, tristeza, asco y aburrimiento. La base de datos fue evaluada mediante una prueba de percepción con respecto a su reconocibilidad y su naturalidad. Las sentencias que fueron reconocidas con un porcentaje mayor al 80% y juzgadas con un porcentaje mayor al 60% como natural fueron seleccionadas y etiquetas. Para mejorar la calidad de las muestras, se utilizaron diferentes audios para ayudar a los actores a reproducir cada una de las emociones. Una de las características más notorias de esta base de datos, es que las 10 sentencias son expresadas para las 7 emociones distintas, y aunado a eso, estas sentencias son interpretables en cada de estos casos.

Existen muchas técnicas desarrolladas para reconocer emociones, una de las formas de aumentar la capacidad de reconocer emociones, es mediante el uso de información multimodal. Estos algoritmos utilizan información de diferentes canales de entrada como: la voz y la imagen del rostro; para mejorar la capacidad de reconocimiento de los clasificadores. En la investigación [20], se presenta una base de datos audio-visual conocida como eNTERFACE'05, cuyo propósito es la evaluación de algoritmos de reconocimiento de emociones (unimodal y multimodal). Para reproducir las emociones a cada uno de los participantes se les pidió escuchar 6 historias sucesivas, cada una de ellas evocando una emoción en particular. Cada uno de ellos debía reaccionar en su propio idioma a cada una de las situaciones mientras eran grabados, luego, dos jurados detallaban si el sujeto reaccionaba de

manera auténtica, y según este criterio se añadía la muestra a la base de datos. No obstante, la distribución geográfica de las personas que fueron parte de la base de datos era muy dispersa y debido a esto, las características como las variaciones del tono y la tasa del habla no eran comunes entre los participantes. Por lo tanto, se tomó la decisión de realizar el mismo protocolo pero reaccionando en inglés. En el segundo protocolo se tomó la decisión de predefinir las respuestas ante los distintos escenarios de las historias, debido a que cuando los actores reaccionaron libremente a cada uno de los escenarios no se expresaron de una manera completamente espontánea. El protocolo final se realizó de la siguiente manera: cada sujeto escuchaba una pequeña historia por cada emoción para intentar inmergirse en el escenario, luego el sujeto reaccionaba mediante cada una de 5 sentencias predefinidas. La base de datos consistió en 1166 secuencias de video, de las cuales 264 eran constituidas por mujeres y 902 por hombres. Una de las características más notorias de esta base de datos es la distribución geográfica de las personas que fueron parte de la misma, además, esta base de datos está constituida por 5 sentencias diferentes por emoción.

Una de las dificultades más comunes en la construcción de bases de datos es la captura de emociones auténticas. Aunque gran parte de las bases de datos utilizan emociones actuadas o simuladas, existe un gran esfuerzo por validar la reconocibilidad y naturalidad de cada una de las emociones que conforman la base de datos. En la investigación [19], se presenta una base de datos polaca, la cual esta conformada por muestras extraídas de discusiones naturales en programas de televisión. Esta base de datos está conformada por declaraciones de interacciones espontáneas y además provee un amplio rango de emociones básicas y complejas. Cada una de las muestras extraídas fueron etiquetadas por un amplio grupo de expertos y voluntarios. La base de datos está constituida de 15 estados emocionales, los cuales se dividen en primarias: ira, anticipación, alegría, miedo, sorpresa, tristeza, asco; y secundarias: rabia, molestia, éxtasis, serenidad, terror, detención, dolor, pensamiento. La base de datos consistió en 784 muestras. La característica más notoria de esta base de datos es que está conformada por sentencias espontáneas, además, contiene un rango de emociones mucho más amplio y altamente diferenciado.

Numerosas investigaciones han propuesto diferentes tipos de bases de datos para el reconocimiento de emociones, entre ellas: la base de datos SAVEE [21], la cual cuenta con 480 muestras de audio con las emociones: ira, asco, sorpresa, alegría, miedo, tristeza y neutralidad. Una de las características más notorias de esta base de datos, es que las emociones son inducidas mediante videos. Existen otros tipos de bases de datos cuyo propósito general no fue la evaluación de algoritmos de aprendizaje para el reconocimiento de emociones, no obstante, son utilizadas para ese fin, entre ellas: la base de datos SUSAS [28], cuyo propósito principal fue el análisis y formulación de algoritmos del reconocimiento del habla en condiciones de ruido y estrés. Otras bases de datos populares como AIBO [29], son construidas a partir de escenarios naturales; en este caso, grabaciones de niños mientras interactúan con un robot. La base de datos tiene 110 diálogos y 29200 palabras en 11 categorías emocionales de ira, aburrimiento, enfático, indefenso, ironía, alegría, maternas, represión, descanso, sorpresa y tacto. El etiquetado de los datos se basa en el juicio de los oyentes. Adicionalmente, existen otras bases de datos orientadas para realizar análisis sentimental como [30], la cual presenta una base de datos de videos en español con 105 muestras etiquetadas mediante su polaridad: positiva o negativa.

III. Método

En las investigaciones sobre algoritmos para el reconocimiento de estados emocionales mediante la voz, son imprescindibles tres aspectos: la base de datos, el procesamiento de los datos y el entrenamiento. La base de datos consiste en el conjunto de muestras de audio que serán parte del entrenamiento del algoritmo de clasificación, cuya calidad y diversidad de las muestras se relaciona directamente con la tasa de precisión del algoritmo. El procesamiento de los datos, consiste en la selección de las características del audio apropiadas que permitirán representar las muestras matemáticamente; el procesamiento es imprescindible y la selección correcta de características influye directamente en la tasa de reconocimiento. Por otro lado, el entrenamiento consiste en utilizar la base de datos para entrenar el algoritmo de aprendizaje de máquina seleccionado. De acuerdo a las características de las muestras algunos algoritmos permiten realizar una mejor clasificación. Por esta razón, en esta investigación se realizará una comparativa de los resultados obtenidos de cada uno de los algoritmos de aprendizaje de máquina seleccionados en cada uno de los experimentos realizados.

A. Construcción de las Bases de Datos

Las bases de datos emocionales de audio, pueden ser clasificadas en tres tipos según la forma en que se pide a las personas demostrar las emociones [7]:

- **Lenguaje actuado**: Se pide a los actores expresar directamente una emoción predefinida.
- **Lenguaje de la vida real**: Respuestas naturales de conversaciones, las cuales son auténticas por naturaleza.
- **Lenguaje emocional evocado**: Las emociones son inducidas y son auto-reportadas en lugar de ser etiquetadas, es decir, la persona reconoce su propia emoción y le asigna por si mismo una etiqueta.

Entre las bases de datos que se basan en lenguaje de la vida real se tiene la base de datos: "Polish Emotional Natural Speech Database" [19] y "Automatic Classification of Emotion-Related User States in Spontaneous Children Speech" [29]. Basada en lenguaje actuado: "A Database of German Emotional Speech" [18]; y basada en la evocación de emociones: "The eNTERFACE'05 audio-visual emotion database" [20] y "Surrey Audio-Visual Expressed Emotion (SAVEE) database" [21].

Actualmente, la mayoría de investigaciones relacionadas al reconocimiento de emociones no son rigurosas con el tipo de pruebas que hacen a los modelos. Por esta razón, y en aras de mantener un mejor control sobre las condiciones de ambiente, en esta investigación se construirán 3 bases de datos orientadas al reconocimiento de emociones, una bajo condiciones controladas y actuadas, otra en condiciones controladas y semi-natural, y finalmente, otra en condiciones no controladas y naturales.

1) Bases de Datos en Condiciones Controladas y Actuadas: Para realizar la construcción de esta base de datos se realizó la selección de conjunto de declaraciones por cada una de las emociones como en [20], las cuales fueron sometidas a 3 tipos de validación: validación de las declaraciones de forma textual, validación por parte de los participantes de la base de datos y validación por parte de un jurado de 4 personas.

La validación de las declaraciones de forma textual, se realizó mediante una encuesta en las cuales participaron 96 personas (66 hombres y 30 mujeres). A cada una de las personas se les presentó un conjunto de declaraciones por cada una de las emociones: ira, sorpresa, felicidad, tristeza, asco y miedo. La encuesta consistía en seleccionar aquellas declaraciones con las cuales expresarían cada una de las emociones. Las declaraciones con las cuales se sintieron identificados gran parte de los participantes de la encuesta fueron: ira, tristeza, felicidad, miedo y sorpresa. Las declaraciones utilizadas para expresar el asco fueron las menos seleccionadas. Adicionalmente, se les pidió a los participantes sugerir qué declaraciones utilizarían ellos para expresar cada emoción. Luego cada una de las declaraciones, junto con las sugerencias de los participantes fueron seleccionadas por 2 jueces y el conjunto de declaraciones resultante fue el siguiente:

El proceso de grabación se realizó en una oficina, con poco ruido. Adicionalmente, todos los participantes fueron ubicados en un mismo sitio para grabar, a una distancia de 40 centímetros del micrófono. El proceso de grabación fue realizado de la siguiente manera:

- A cada uno de los participantes se les pidió sentarse en una silla ubicada a 40 centímetros del micrófono.
- A cada participante se le pidió leer el conjunto de declaraciones de cada emoción, luego, se le pidió reproducir (actuar) cada una de las declaraciones de cada emoción 4 veces de distintas maneras.
- En caso de no expresar correctamente alguna declaración o de que el participante no estuviese satisfecho con el resultado, se le pedía al participante repetir dicha emoción utilizando como ayuda la orientación del operador o muestras de participantes anteriores.
- Para realizar las grabaciones fue utilizado el software Audacity [31]. Se utilizó un solo canal de grabación y la frecuencia de muestreo fue de 48 kHz.
- Luego del proceso de grabación cada una de las declaraciones fue seleccionada por el operador, el cual descartó aquellas declaraciones contaminadas (ruidos de golpes de mesa, movimientos de sillas, entre otros) o de poca calidad (declaraciones incompletas o ambiguas).

TABLA II: CONJUNTO DE DECLARACIONES DE CADA EMOCIÓN

Ira	1) ¿Qué te pasa? 2) !Eso a mi que me importa! 3) ¡O te vas o te boto! 4) ¿Me vas a atender o no? 4) ¿Sabes qué? !Déjalo así! 5) No me molestes!
Sorpresa	1) ¡No puede ser! ¿En serio? 2) ¡Qué! ¡Yo no sabia eso! 3) ¡Jamás lo hubiera creído! 4) ¡No me lo esperaba! 5) ¡No te creo! ¿De verdad? 6) ¿De verdad? ¡No sabia! 7) ¿Es en serio?
Felicidad	1) ¡Gané! 2) ¡Que genial! Pase! 3) ¡No me lo creo! ¡qué suerte! 4) ¡Lo logre! ¡Al fin! 5) ¡No puede ser! ¡qué bien! 6) ¡No lo creo! ¡Funciona!
Miedo	1) No, no me hagas daño 2) Ya no tengo más, no tengo nada. 3) No, no me robes 4) Aléjate, Aléjate 5) Aléjate por favor 6) no, por favor
Asco	1) Esto si está feo! 2) ¿Qué hay en el plato? 3)¡Qué repugnante! 4) ¿Qué asco? ¿Qué es esto? 5) ¡Un bicho! 6) ¿Qué es esto?
Tristeza	1) Todo iba tan bien, no sé qué paso 2) Lo/La extraño pero se fue 3) Ya no será lo mismo 4) Dime que no es verdad 5) Aún sentía algo por ella 6) El/Ella fue parte de mi vida 7) No pude hacerlo

- Luego de seleccionadas las muestras, se recortaron cuidadosamente y se transformaron a 16 kHz. Adicionalmente, cada una de las muestras fue etiquetada.

La validación por parte de los participantes de la base de datos, consistió en reproducir cada una de las declaraciones del participante y hacerle dos preguntas por cada una de ellas: ¿Considera que en esta declaración se expresó la emoción? y ¿Considera que esta declaración pudiera interpretarse de otra manera?. Si algunas de las dos preguntas anteriores eran respondidas de manera negativa, se descartaba la muestra. En el caso particular en donde se descartaban todas las muestras de una declaración, se repetía el proceso de grabación.

Finalmente, la validación por parte de un jurado consistió en cuantificar la validez del contenido mediante la "V" de Aiken. El número de jurados fue 4. A cada uno de los jurados se les pidió calificar cada una de las muestras previamente filtradas por las validaciones anteriores según las preguntas de la siguiente escala:

TABLA III: ESCALA UTILIZADA PARA LA VALIDACIÓN DE LAS MUESTRAS DE AUDIO

Significado	Valor
El audio es entendido e interpretado inequívocamente de una única manera	**3**
Para algunas personas podría tener otro significado	**2**
El audio es susceptible de ser entendido en sentidos diversos	**1**
El audio definitivamente se presta para múltiples interpretaciones.	**0**

Luego de realizar el etiquetado en base a la escala anterior, se obtuvo el coeficiente "V" de Aiken para cada una de las muestras.

$$V = \frac{S}{(n(C-1))} \quad (1)$$

En la ecuación 1, el valor S, representa la suma de los valores de cada jurado por cada muestra. El valor n, representa el número de personas en el jurado, y el valor C, el número de valores en la escala de valoración. Las muestras cuyo cálculo de validación fue mayor a 0.75 fueron aceptadas y formaron parte de la base de datos. El resto de muestras fue descartado.

La base de datos en condiciones controladas y actuadas, fue conformada por un total de 1351 muestras. La frecuencia de las muestras por emoción puede verse en la Figura 1.

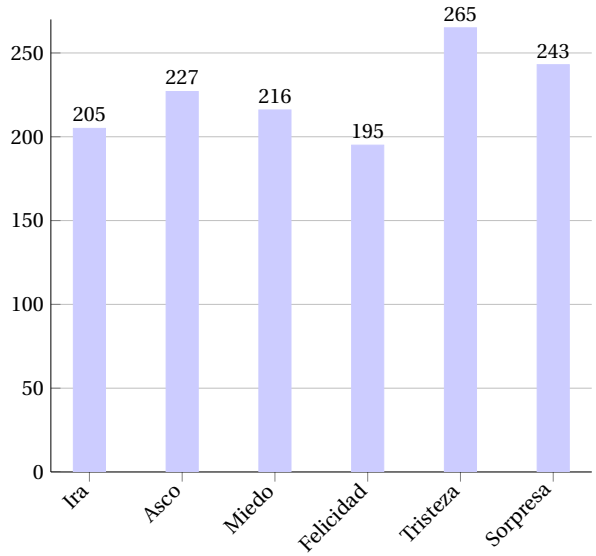

Figura 1: Frecuencia de las Emociones de la Base de Datos en Condiciones Controladas y Actuadas

2) Bases de Datos en Condiciones Controladas y Semi-naturales: Para crear un conjunto que permitiera probar de manera rigurosa los modelos entrenados con la base de datos anterior, se realizó una base de datos en las mismas condiciones pero variando las palabras utilizadas inicialmente. Esta base de datos, de manera similar a la anterior, fue sometida a los 3 tipos de validación descritos en la sección anterior.

En aras de determinar la capacidad de los modelos para reconocer un amplio grupo de expresiones en las personas, esta base de datos consistió en expresar cada una de las declaraciones de cada emoción utilizando sus propias palabras. Es decir expresando el mismo significado y la emoción de la declaración original pero utilizando las palabras que utilizaría el participante en vida cotidiana.

Todo el proceso de grabación fue similar al anterior, el único cambio que se realizó, se basó en que los participantes debían utilizar sus propias palabras para expresar 4 veces cada una de las declaraciones de cada emoción (ver Tabla II).

La base de datos en condiciones controladas y semi-naturales, fue conformada por un total de 1163 muestras. La frecuencia de las muestras por emoción puede verse en la Figura 2.

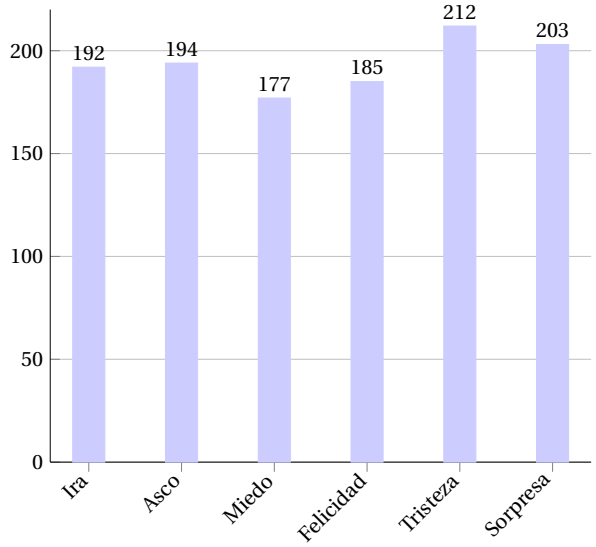

Figura 2: Frecuencia de las Emociones de la Base de Datos en Condiciones Controladas y Semi-naturales

3) Bases de Datos en Condiciones no Controladas y Naturales: La base de datos en condiciones no controladas, cuenta con 105 muestras, 70 muestras provenientes de videos Internet, y 30 muestras provenientes de 3 personas (1 hombre y 2 mujeres).

El proceso de grabación fue realizado de la siguiente manera:

- A cada participante se le muestran segmentos de audio que usaran de guía para reproducir la emoción.
- Todos los segmentos de audio contienen información sobre una situación en particular.
- No se tomó en consideración la ubicación donde se realizaron las grabaciones.

- Cada participante es grabado mediante un micrófono convencional utilizando la biblioteca PyAudio [32].
- De la grabación se extraen las declaraciones que se consideran espontaneas y naturales.
- Cada segmento de audio tiene un tamaño entre 2 y 6 segundos.
- Cada segmento de audio, se graba con una frecuencia de muestreo de 16 kHz y se almacena en formato wav.

Para realizar la validación de cada una de estas muestras se le pide al participante escuchar las declaraciones seleccionadas y luego se valida la emoción tomando en consideración la opinión de dos jueces y la del participante. Si dos de tres opiniones coinciden, entonces la muestra es etiquetada y luego añadida a la base de datos.

Las muestras de Internet fueron obtenidas y procesadas mediante la biblioteca Youtube-dl [33]. Todas las muestras de audio son en español, principalmente países de América Latina. El proceso de obtención de muestras de Internet fue realizado de la siguiente manera:

- Fueron seleccionados videos en español, en los cuales habla una sola persona sin sonidos musicales de fondo.
- Para cada video fueron registrados los segmentos que se corresponden con una emoción particular.
- Se utilizó el tiempo de inicio y fin, el enlace del video y la etiqueta para obtener el segmento de audio correspondiente mediante la biblioteca youtube-dl [33].
- Todos los segmentos fueron transformados a formato wav con una frecuencia de muestreo de 16 kHz.

Para realizar la validación de cada una de las muestras de Internet, se utilizó la opinión de 3 jueces, los cuales de manera similar al proceso anterior, se basó en la opinión de cada uno de ellos.

La base de datos en condiciones no controladas, contiene audios correspondientes a las 6 emociones universales descritas por [27]: ira, miedo, felicidad, asco, tristeza y sorpresa. Adicionalmente, contiene la neutralidad. Si dos de tres opiniones coinciden, entonces la muestra es etiquetada y luego añadida a la base de datos. La base de datos fue conformada por un total de 105 muestras. La frecuencia de cada una de las emociones puede verse en la Figura 3.

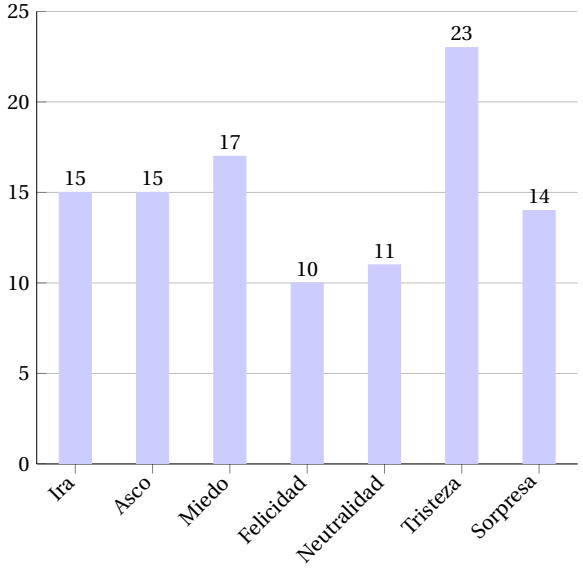

Figura 3: Frecuencia de las Emociones de la Base de Datos en Condiciones no Controladas

B. Procesamiento de los Datos

Los segmentos de audio se procesaron para obtener las características que serán representadas como un vector. Para realizar esto, un proceso de extracción a largo plazo fue llevado a cabo con la ayuda de la biblioteca de análisis de audio PyAudioAnalysis [34].

El proceso de extracción de características a largo plazo, consiste en obtener el promedio de características de mediano plazo que a su vez depende del procesamiento a corto plazo de la señal de audio. Esta forma de procesar el audio también se le conoce como segmental (corto y mediano plazo) y suprasegmental (largo-plazo) [35]. Para cada audio se utilizaron marcos de 20 mili-segundos en el procesamiento a corto plazo sin solapamiento, y segmentos de 1 segundo para el procesamiento a mediano plazo con solapamiento, para finalmente obtener un vector con 28 estadísticas.

En la Figura 4, se puede observar un esquema que muestra como se lleva a cabo el proceso de extracción de características. En la primera fase del procesamiento (análisis a corto plazo), se obtienen las características $c1, c2, ...c_N$ de cada marco $(m_1, m_2, ..., m_b)$; en la segunda fase del proceso (análisis a mediano plazo), se extraen estadísticas de las características particulares de cada uno de los marcos del bloque, en este caso el promedio μ_N y la desviación estándar σ_N; Por último, se realiza un procesamiento suprasegmental (análisis a largo plazo), el cual consiste en obtener el promedio μ'_{2N} de las estadísticas de la fase anterior.

Las señales de audio, y en particular aquellas que contienen contenido emocional, se caracterizan por tener un gran número de información. Una de las cosas más importantes en las investigaciones de reconocimiento de emociones a través de la voz, es seleccionar un conjunto de características adecuadas de tal manera que se pueda representar lo mejor posible cada una de las muestras de audio.

Para esta investigación se utilizaron 2 tipos de características: del dominio del tiempo y del dominio cepstral. Todas las características fueron obtenidas mediante la biblioteca PyAudioAnalysis [34], una descripción formal de las características junto con sus algoritmos puede encontrarse en [36]. A continuación se muestra una descripción de las características que se seleccionaron para esta investigación.

1) Energía o Potencia de la señal: La energía se define como la suma de los cuadrados de las muestras, que usualmente se normaliza dividendo entre la longitud de la muestra. La energía es la característica más básica en el procesamiento de señales de la voz. Ésta juega un papel importante en el reconocimiento de emociones.

Por ejemplo, las emociones como la felicidad o la ira contienen una mayor energía en comparación a la tristeza. La mayoría de las investigaciones utilizan esta característica [2], [4], [8], [9], [11], [12].
Sea $X_i(n), n = 1, ..., W_L$ la secuencia de muestras de audio en el i-ésimo marco, donde W_L es el tamaño del marco. La energía a corto plazo es calculada como sigue:

$$E(i) = \sum_{n=1}^{W_L} |X_i(n)|^2 \quad (2)$$

Usualmente la energía es normalizada dividiéndola por el tamaño del marco W_L para remover la dependencia de la longitud del marco, quedando el cálculo de la siguiente manera:

$$E(i) = \frac{1}{W_L} \sum_{n=1}^{W_L} |X_i(n)|^2 \quad (3)$$

2) MFFCs: los coeficientes cepstrales de las frecuencias de Mel han sido muy populares en el campo del análisis de voz. En la práctica, los MFCCs son los coeficientes discretos de la transformación coseno del espectro de potencia logarítmica en la escala de Mel. Los MFCCs han sido ampliamente utilizados en el reconocimiento de voz, agrupamiento de altavoces, reconocimiento de emociones y muchos otros tipos de aplicaciones de análisis de audio y aprendizaje de máquina. Caracterizan la magnitud del espectro y por lo general son usados los 12 primeros coeficientes. En la gran mayoría de investigaciones los MFFCs han mostrado ser la característica que mejores cualidades tiene para el reconocimiento de emociones [2], [4], [6]–[9], [11], [12].
Para extraer los coeficientes cepstrales de las frecuencias de Mel de un marco, son necesarios los siguientes pasos:

a) La transformada discreta de Fourier (DFT, por sus siglas en inglés) es calculada. Esta es usada para derivar la representación de la señal en el dominio de la frecuencia (espectral), la cual sirve como entrada para la obtención de muchas características importantes.
Dada una señal discreta en el dominio del tiempo $x(n), n = 0, ..., N-1$, con N muestras de longitud, su DFT es calculada como sigue:

$$X(k) = \sum_{n=0}^{N-1} x(n) exp(-j\frac{2\pi}{N}kn),\ k = 0, ..., N-1, \quad (4)$$

b) El espectro resultante es utilizado como entrada a un banco de filtros de la escala de Mel que consiste en L filtros. Los filtros usualmente tienen una frecuencia triangular superpuesta. La escala de Mel introduce una función de distorsión de frecuencia que intenta ajustarse a ciertas observaciones psicoacústicas. A través de los años varias funciones de distorsión de frecuencias han sido propuestas por ejemplo:

$$f_w = 2595 * log(1 + f/700) \quad (5)$$

Si $\tilde{O}_k, k = 1, ..., L$, es la potencia en la salida del k-ésimo filtro, entonces los MFCCs están dados por la siguiente ecuación

$$C_m = \sum_{k=1}^{L} (log\tilde{O}_k) cos[m(k - \frac{1}{2})\frac{\pi}{L}],\ m = 1, ..., L. \quad (6)$$

En total se genera un vector de 14 características $(c_1, c_2, c_3, ..., c_{14})$ por cada marco (1 valor correspondiente a la energía y 13 coeficientes de Mel), que será usado para generar un vector de una dimensión igual a 28 (μ'_{2N}), cuyos elementos corresponden al promedio μ y desviación estándar σ de las 14 características obtenidas mediante el procesamiento a largo plazo.

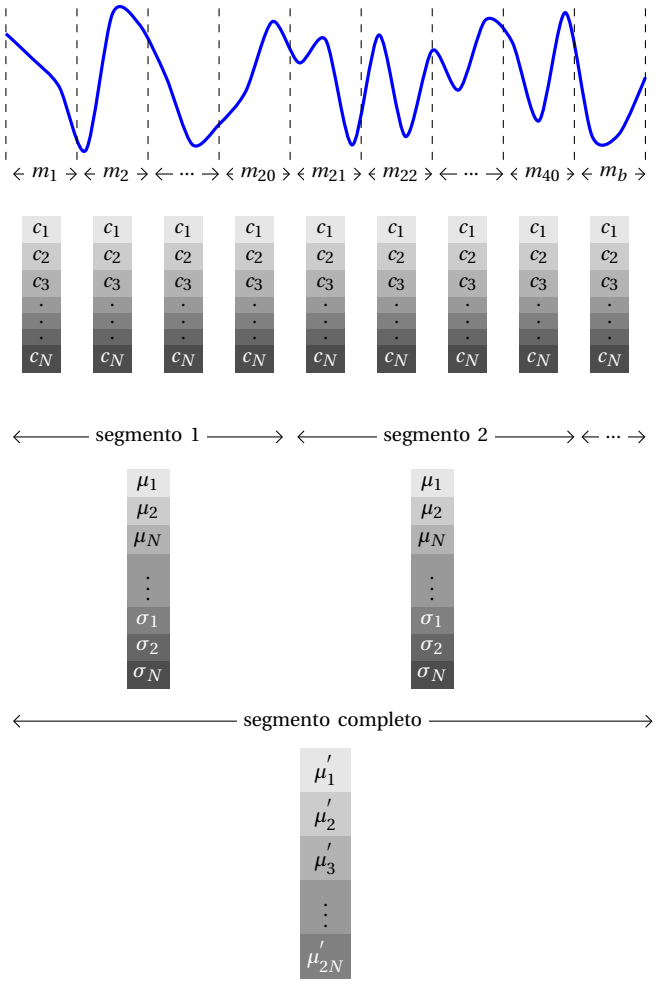

Figura 4: Procesamiento a Largo Plazo o Suprasegmental del Audio

C. Entrenamiento

Muchos algoritmos de aprendizaje de máquina han sido utilizados en diferentes investigaciones sobre el reconocimiento

de emociones a través del audio. Una lista de diferentes algoritmos y su desempeño se puede ver en la Tabla I. Todos los clasificadores necesitan datos de entrenamiento y datos de prueba, este último es usado para calcular la tasa de reconocimiento del clasificador.

En esta investigación se realizará un modelo por cada uno de los algoritmos de aprendizaje de máquina utilizados (SVM, GB y RF). Cada uno de estos modelos será entrenado con el 70% de las muestras pertenecientes a la base de datos en condiciones controladas y actuadas. Posteriormente, se realizarán 3 tipos de pruebas a cada uno de los modelos:

1) P1: Pruebas utilizando el 30% restante de la base de datos en condiciones controladas y actuadas.
2) P2: Pruebas utilizando toda la base de datos en condiciones controladas y semi-naturales.
3) P3: Pruebas utilizando toda la base de datos en condiciones no controladas, a excepción de la neutralidad.

El proceso de clasificación utilizará los vectores provenientes del módulo de extracción de características para su entrenamiento y prueba. En la Figura 5 se puede observar el diagrama del proceso de clasificación. Se utilizarán 3 tipos de clasificadores: bosques aleatorios, Aumento del Gradiente y Máquinas de Vectores de Soporte. La biblioteca Scikit-learn [37], es utilizada para la implementación.

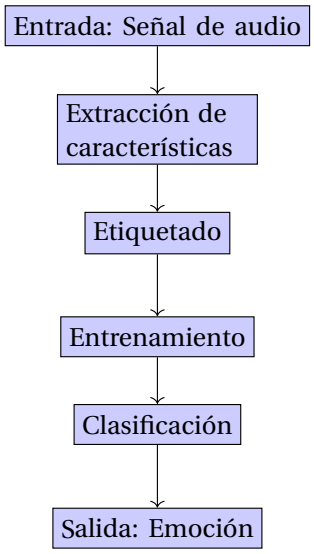

Figura 5: Diagrama del Proceso de Clasificación

IV. Resultados

A continuación se presentan los resultados obtenidos mediante cada una de las pruebas mencionadas anteriormente.

En la Tabla IV, se pueden apreciar las tasas de reconocimiento de cada uno de los modelos implementados en cada una de las pruebas. En este caso se puede apreciar que el modelo SVM, obtuvo la mejor tasa de reconocimiento en las pruebas 1 y 2, mientras que el modelo GB obtuvo mejores resultados en la prueba 3.

TABLA IV: Resultados de las Tasas de Reconocimiento para cada una de las Pruebas

	P1	P2	P3
SVM	83%	68%	17%
GB	79%	66%	27%
RF	79%	67%	23%

A continuación se presentan los resultados individuales de las tasas de reconocimiento de cada emoción de manera individual para cada uno de los modelos.

TABLA V: Resultados de las Tasas de Reconocimiento para cada una de las Emociones utilizando el Modelo SVM

	Ira	Tristeza	Asco	Felicidad	Sorpresa	Miedo
P1	82%	89%	81%	84%	78%	**94%**
P2	52%	**87%**	69%	60%	56%	75%
P3	**57%**	0%	0%	16%	0%	20%

TABLA VI: Resultados de las Tasas de Reconocimiento para cada una de las Emociones utilizando el Modelo GB

	Ira	Tristeza	Asco	Felicidad	Sorpresa	Miedo
P1	75%	84%	70%	83%	73%	**90%**
P2	64%	**87%**	70%	55%	49%	69%
P3	**80%**	0%	50%	14%	0%	12%

TABLA VII: Resultados de las Tasas de Reconocimiento para cada una de las Emociones utilizando el Modelo RF

	Ira	Tristeza	Asco	Felicidad	Sorpresa	Miedo
P1	89%	80%	74%	79%	77%	**92%**
P2	65%	**90%**	70%	54%	53%	64%
P3	**53%**	0%	60%	0%	0%	18%

En las Tablas V, VI y VII, se pueden apreciar los porcentajes de reconocimiento obtenidos para cada modelo en cada una de las pruebas realizadas. El miedo y la tristeza, fueron las emociones que mejor se reconocieron en cada uno de los modelos para las pruebas 1 y 2, la sorpresa fue la emoción que menor tasa de reconocimiento obtuvo en las pruebas 1 y 2. Adicionalmente, la tristeza y la sorpresa no fueron reconocidas en ningunos de los modelos para la prueba 3.

V. Discusión

Las emociones humanas pueden ser expresadas de diversas maneras, por lo que registrar un conjunto representativo de este

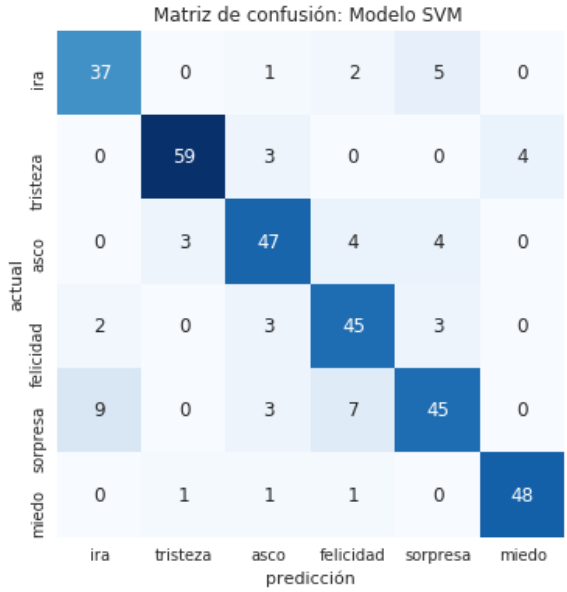

Figura 6: Matriz de Confusión: Resultados del Modelo SVM en la Prueba 1. Utilizando el Subconjunto de Muestras Actuadas en Condiciones Controladas

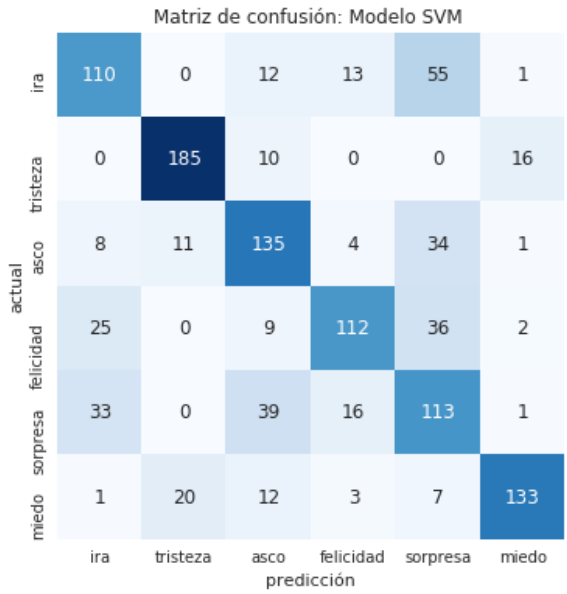

Figura 7: Matriz de Confusión: Resultados del Modelo SVM en la Prueba 2. Utilizando el Subconjunto de Muestras Semi-naturales en Condiciones Controladas

rango es una tarea realmente compleja. Esto, limita la capacidad de los modelos para reconocer emociones y por consiguiente, limita la capacidad de las aplicaciones robóticas que hacen uso de estos modelos. Es por esta razón, que en esta investigación se realizaron 3 bases de datos distintas, para probar rigurosamente la capacidad de los modelos para reconocer emociones en un amplio rango de expresiones y personas.

Los resultados obtenidos mostraron que el algoritmo Máquinas de Vectores de Soporte con núcleo radial obtuvo la mejor tasa de reconocimiento 83% y 68% para las pruebas 1 y 2 respectivamente, en las Figuras 6 y 7 se pueden apreciar las respectivas matrices de confusión. Los resultados obtenidos en las pruebas 1 y 2 para cada una de las emociones fue mayor al 52%, lo que se considera un buen resultado tomando en consideración la cantidad de muestras en cada una de las pruebas a las cuales fue sometido. Adicionalmente, la emoción que mejor se reconoció en la prueba 3 fue la ira.

En el caso del modelo Aumento de Gradiente, los resultados obtenidos para estas dos pruebas, fueron mayores al 49% para las pruebas 1 y 2. Para la prueba 3, este modelo fue el que mejor reconoció la ira, alcanzando un porcentaje de reconocimiento del 80%.

En caso del modelo Bosques Aleatorios, los resultados obtenidos para las pruebas 1 y 2 fueron mayores al 53%, lo cual supera a los modelos SVM y GB descritos anteriormente. Para la prueba 3, este modelo fue el que mejor reconoció el asco, con un porcentaje de reconocimiento del 60%.

Las emociones que mejor se reconocieron en todas las pruebas, fueron el miedo y la tristeza. Para el miedo, el mejor resultado lo obtuvo el modelo SVM con un porcentaje de reconocimiento del 94%, mientras que para la tristeza, el mejor resultado lo obtuvo el modelo RF con un porcentaje de reconocimiento del 90%. En el caso de la tristeza, se puede atribuir este resultado a que esta emoción es la que tiene mayor número de muestras en las bases de datos con condiciones controladas, ver Figuras 1 y 2. No obstante, el miedo es la emoción que menor número de muestras tiene en la base de datos en condiciones controladas y semi-naturales, por lo tanto, este resultado se puede atribuir a que el miedo fue expresado de manera muy consistente tanto en la base de datos en condiciones controladas y actuadas (prueba 1) como en la base de datos en condiciones controladas y semi-naturales.

En el caso de la base de datos en condiciones no controladas, se pudo observar que se obtuvieron los peores resultados, incluso algunas emociones como la tristeza y la sorpresa no fueron reconocidas por ningún modelo. A pesar de que estas muestras están correctamente validadas, este resultado puede ser atribuido a que una emoción puede ser expresada de distintas maneras y en diferentes intensidades como en [19]. Por lo tanto, en aras de reconocer un amplio grupo de emociones en las personas, es necesario entrenar los modelos con un grupo altamente representativo con todas las variaciones en las cuales una emoción puede ser expresada, concretamente mediante la voz.

A pesar de que la prueba 3 fue la que peores porcentajes de reconocimiento obtuvo (condiciones no controladas). Estas pruebas permitieron descubrir que en el caso particular de la ira, se pudo observar que en todos los modelos fue la emoción que mejor se reconoció en esta prueba. Esto puede atribuirse a que a diferencia de otras emociones, la ira es expresada de una manera muy consistente entre todas las personas.

VI. Conclusiones

En esta investigación se construyeron 3 bases de datos orientadas al reconocimiento de emociones en español. Una de ellas en condiciones controladas y actuadas, otra en condiciones controladas y semi-naturales; y finalmente una en condiciones no controladas.

La base de datos de emociones en condiciones controladas y actuadas fue utilizada para entrenar los algoritmos de aprendizaje de máquina seleccionados para esta investigación. Adicionalmente, se realizaron 3 pruebas distintas para probar la capacidad de reconocimiento de los modelos, utilizando muestras actuadas, semi-naturales y naturales.

Como se puede apreciar en los resultados de la Tabla IV, los mejores resultados se obtienen cuando se prueban los modelos con muestras de la misma base de datos con la cual ha sido entrenada. A medida que las pruebas realizadas se salen del marco del conjunto de entrenamiento del modelo, este disminuye su capacidad de reconocimiento.

Considerando los resultados obtenidos, se puede concluir que generalizar un rango de emociones es una tarea realmente compleja. Incluso cuando se varían condiciones pequeñas, por ejemplo, en el caso de la base de datos con muestras semi-naturales, los resultados disminuyen considerablemente.

Finalmente, trabajos futuros se orientan en la construcción de una base de datos con un rango más amplio de emociones y diferentes niveles de intensidad. Así como la aplicación de nuevos algoritmos y métodos de procesamiento de audio para el reconocimiento de emociones.

Referencias

[1] R. W. Picard, "Toward Computers that Recognize and Respond to User Emotion," *IBM Systems Journal*, vol. 39, no. 3.4, pp. 705–719, 2000.

[2] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion Recognition by Speech Signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[3] M. P. D. Das and M. S. Sengupta, "An Emotion Based Speech Analysis," 2015.

[4] M. Ghai, S. Lal, S. Duggal, and S. Manik, "Emotion Recognition on Speech Signals Using Machine Learning," in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, March 2017, pp. 34–39.

[5] V. Kirandzhiska and N. Ackovska, "Sound Features Used in Emotion Classification," 2012.

[6] Y. D. Chavhan, B. S. Yelure, and K. N. Tayade, "Speech Emotion Recognition using RBF Kernel of LIBSVM," in *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, Feb 2015, pp. 1132–1135.

[7] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker Dependent Speech Emotion Recognition Using MFCC and Support Vector Machine," in *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, Sept 2016, pp. 1080–1084.

[8] T. Iliou and C.-N. Anagnostopoulos, "Classification on Speech Emotion Recognition a Comparative Study," *animation*, vol. 4, pp. 5, 2010.

[9] P. Chandrasekar, S. Chapaneri, and D. Jayaswal, "Emotion Recognition From Speech using Discriminative Features," vol. 101, pp. 31–36, 09 2014.

[10] H. Palo, P. Kumar, and N. Mohanty, "Emotional Speech Recognition Using Optimized Features," Vol. 5, pp. 4–9, 12 2017.

[11] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema, and S. Rajan, "Emotion Recognition From Audio Signals using Support Vector Machine," in *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, Dec 2015, pp. 139–144.

[12] Y. Pan, P. Shen, and L. Shen, "Feature Extraction and Selection in Speech Emotion Recognition," *Proceeding of the onlinepresent. org*, vol. 2, pp. 64–69, 2012.

[13] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion Recognition in Context," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1960–1968.

[14] M. Ménard, P. Richard, H. Hamdi, B. Daucé, and T. Yamaguchi, "Emotion Recognition Based on Heart Rate and Skin Conductance." in *PhyCS*, 2015, pp. 26–32.

[15] H. W. Guo, Y. S. Huang, C. H. Lin, J. C. Chien, K. Haraikawa, and J. S. Shieh, "Heart Rate Variability Signal Features for Emotion Recognition by using Principal Component Analysis and Support Vectors Machine," in *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, Oct 2016, pp. 274–277.

[16] A. Pradhan, A. Singh, and S. Saraswat, "Emotion Recognition through Wireless Signal," in *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN)*, Feb 2017, pp. 91–95.

[17] J. C. Castillo, Á. Castro-González, F. Alonso-Martín, A. Fernández-Caballero, and M. Á. Salichs, "Emotion Detection and Regulation From Personal Assistant Robot in Smart Environment," in *Personal Assistants: Emerging Computational Technologies*. Springer, 2018, pp. 179–195.

[18] F. Burkhardt, A. Paeschke, M. A. Rolfes, W. F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *INTERSPEECH*, 2005.

[19] S. Grocholewski, "Corpora-speech Database for Polish Diphones," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[20] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The Enterface'05 Audio-Visual Emotion Database," in *Proceedings of the 22 nd International Conference on Data Engineering Workshops*, ser. ICDEW '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 8–. [Online]. Available: http://dx.doi.org/10.1109/ICDEW.2006.145

[21] P. Jackson and S. Haq, "Surrey Audio-Visual Expressed Emotion (SAVEE) Database," *University of Surrey: Guildford, UK*, 2014.

[22] R. Barra Chicote, J. M. Montero Martínez, J. Macías Guarasa, S. L. Lutfi, J. M. Lucas Cuesta, F. Fernández Martínez, L. F. D'haro Enríquez, R. San Segundo Hernández, J. Ferreiros López, R. D. Córdoba Herralde *et al.*, "Spanish Expressive Voices: Corpus For Emotion Research in Spanish," 2008.

[23] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J. M. Blanco, D. Bernadas, J. M. Oliver, D. Tena, and L. Longhi, "Validation of an Acoustical Modelling of Emotional Expression in Spanish using Speech Synthesis Techniques," 2000.

[24] F. Burkhardt and W. F. Sendlmeier, "Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis," in *ISCA Tutorial and Research Workshop (ITRW) on speech and emotion*, 2000.

[25] C. J. Van der Hofstadt Román, *El Libro de las Habilidades de Comunicación*. Ediciones Díaz de Santos, 2005.

[26] J. G. R´azuri, D. Sundgren, R. Rahmani, A. Larsson, A. M. Cardenas, and I. Bonet, "Speech Emotion Recognition in Emotional Feedback for Human-Robot Interaction," *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 2, 2015. [Online]. Available: http://dx.doi.org/10.14569/IJARAI.2015.040204

[27] R. Bates Graber, "Ekman: The Face of Man: Expressions of Universal Emotions in a New Guinea Village," 05 2017.

[28] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting Started with SUSAS: a Speech Under Simulated and Actual Stress Database," in *EUROSPEECH*, 1997.

[29] S. Steidl, "Automatic Classification of Emotion Related User States in Spontaneous Children's Speech," 2009.

[30] V. P. Rosas, R. Mihalcea, and L.-P. Morency, "Multimodal Sentiment Analysis of Spanish Online Videos," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38–45, 2013.

[31] A. Team, "Audacity (version 2.0. 0)," *Audio editor and recorder*, 2012.

[32] H. Pham, "Pyaudio: Portaudio v19 Python Bindings," *URL: https://people. csail. mit. edu/hubert/pyaudio*, 2006.

[33] R. G. Gonzalez, "Youtube-dl: Download Videos From Youtube.com," 2006.

[34] T. Giannakopoulos, "Pyaudioanalysis: An Open-source Python Library for Audio Signal Analysis," *PloS one*, vol. 10, no. 12, 2015.

[35] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and Classifiers for Emotion Recognition From Speech: a Survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, Feb 2015. [Online]. Available: https://doi.org/10.1007/s10462-012-9368-5

[36] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: a MATLAB Approach*. Academic Press, 2014.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[38] D. Ververidis and C. Kotropoulos, "A Review of Emotional Speech Databases," in *Proc. Panhellenic Conference on Informatics (PCI)*, 2003, pp. 560–574.

[39] H. Nguyen, K. Kotani, F. Chen, and B. Le, "A Thermal Facial Emotion Database and its Analysis," in *Image and Video Technology*, R. Klette, M. Rivera, and S. Satoh, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 397–408.