ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.872

AI-powered voice analysis to recognize the various emotions of students at a university

Edward Flores[1]; Justo-Pastor Solis-Fonseca[1]; Jose-Hilarion Rosales-Fernandez[1]; Cesar-Raul Cuba-Aguilar[1]
[1]Universidad Nacional Federico Villarreal, Lima-Perú,
eflores@unfv.edu.pe, jsolis@unfv.edu.pe, jrosales@unfv.edu.pe, ccubaa@unfv.edu.pe

Abstract*– This paper addresses the use of artificial intelligence in voice analysis to recognize the emotions experienced by university students. The objective of this project was to create a voice recognition prototype to improve students' emotional well-being and enrich their learning process. The study highlights the importance of AI in education for comprehensively evaluating student participation, intrinsic motivation, and emotional well-being—essential factors for their holistic development. The research was quantitative and experimental, conducted with first-year university students. A total of 4,723 audio recordings were collected and categorized into six emotions: happy, sad, angry, scared, surprised, and neutral. Various techniques were then applied to balance and expand the model's training dataset. The trained machine learning model performed excellently, achieving 98% effectiveness in training and 91% in validation. Neutrality and Sadness were found to be recognized with high consistency; in contrast, Surprise was the most difficult emotion to recognize. The findings highlighted the influence of these emotions on the learning process, providing evidence for the subsequent development of a Socio-emotional-Constructivist Pedagogical Model that integrates pedagogical strategies to develop more comprehensive, supportive, and effective learning environments.

***Keywords*-- emotions, artificial intelligence, MFCC, neural networks.**

Análisis de voz con IA para reconocer las diversas emociones de estudiantes en una universidad.

Edward Flores[1]; Justo-Pastor Solis-Fonseca[1]; Jose-Hilarion Rosales-Fernandez[1]; Cesar-Raul Cuba-Aguilar[1]
[1]Universidad Nacional Federico Villarreal, Lima-Perú,
eflores@unfv.edu.pe, jsolis@unfv.edu.pe, jrosales@unfv.edu.pe, ccubaa@unfv.edu.pe

Resumen*– En el presente trabajo se aborda el análisis de voz mediante inteligencia artificial para reconocer las emociones que atraviesan los estudiantes universitarios. El objetivo de este proyecto fue crear un prototipo de reconocimiento de voz para mejorar el bienestar emocional del estudiante y enriquecer su proceso de aprendizaje. El estudio destaca la importancia de la IA en la educación para evaluar de manera integral la participación, motivación intrínseca y bienestar emocional de los estudiantes, factores esenciales para su desarrollo integral. La investigación fue cuantitativa, experimental, y se realizó con estudiantes del primer año universitario. Se recopilaron 4,723 audios, categorizados en 6 emociones: feliz, triste, enojado, asustado, sorprendido y neutral. Luego, se aplicaron diversas técnicas para equilibrar y ampliar el conjunto de datos de entrenamiento del modelo. El modelo de aprendizaje automático entrenado dió como resultado un desempeño excelente, logrando un 98% de efectividad en el entrenamiento y 91% en la validación. Se observó que las emociones de Neutralidad y Tristeza se identificaron de manera muy consistente; la emoción de Sorpresa fue la más difícil de identificar. Los resultados mostraron la incidencia de estas emociones en el proceso de aprendizaje, dejando evidencia para elaborar posteriormente un Modelo Pedagógico Socioemocional-Constructivista que incorpore estrategias pedagógicas para crear ambientes de aprendizaje más comprensivos, solidarios y efectivos.

Palabras clave—emociones, inteligencia artificial, MFCC, redes neuronales.

I. Introducción

La inteligencia artificial y la inteligencia emocional se han popularizado en los últimos años, y más en el ámbito educativo, ya que el hecho de entender las emociones de los estudiantes afecta sus resultados de aprendizaje. Las nuevas tecnologías de análisis de voz permiten conocer las emociones de los estudiantes. Con la IA procesando grabaciones de voz, los profesores están aprendiendo a medir la participación, el compromiso y el estado de ánimo de los alumnos. Esto les permite crear ambientes de aprendizaje más sensibles y flexibles. La importancia de esta investigación es cómo afecta en el rendimiento académico y en la salud mental. Por ejemplo, estudios han demostrado que poder reconocer y controlar las emociones favorece las relaciones entre estudiantes y profesores y el rendimiento escolar [1]. Pero ahora que las escuelas se enfocan en el aprendizaje socioemocional, el análisis de voz es una manera de fomentar la educación integral, la que desarrolla la mente y las emociones [2]. Una pregunta que surge de la literatura reciente es qué tan bien pueden las técnicas de IA identificar e interpretar señales emocionales en datos de voz. Algunas investigaciones muestran que los algoritmos de aprendizaje automático pueden identificar diferentes estados emocionales en base a características de la voz, como el tono, el timbre y el ritmo [3]. La evidencia empírica demuestra que estos algoritmos son capaces de clasificar emociones como la alegría, la ira o la tristeza en contextos reales como el educativo, donde los cambios emocionales a menudo escapan a la percepción del profesorado [4]. Pero la mayor parte de la literatura se enfoca en adultos, lo que genera una laguna sobre cómo estas tecnologías se pueden adaptar con éxito a las necesidades del estudiantado en diversos contextos educativos [5].

Pero todavía existen carencias de marcos pedagógicos integrales que integren conocimientos tecnológicos y pedagógicos para orientar a los profesores sobre cómo integrarlos en la práctica [6]. Un tema candente de la literatura científica son las implicancias éticas de la IA en la educación. La privacidad y seguridad de los datos son esenciales, sobre todo tratándose de datos sensibles sobre el estado de ánimo del estudiantado [7]. Cómo los profesores pueden combinar la información aprendida por medio de la IA con las preocupaciones éticas es una cuestión compleja que requiere más investigación científica [8]. Además, aunque la literatura ha reconocido el potencial de la IA para poder mejorar la conciencia emocional del estudiantado, la evidencia experimental que respalde estos efectos en el largo plazo aún es escasa. Esto requiere estudios de tipo longitudinal que evalúen cómo el análisis de voz afecta el rendimiento y el desarrollo emocional del estudiante con el tiempo [9].

El reconocimiento de emociones en la voz con IA es un paso para comprender la interacción humana. Esta tecnología, entre la computación afectiva y adaptativa, pretende interpretar alteraciones en el habla como reflejo de estados emocionales, siendo de gran utilidad en áreas como la teleasistencia y la salud electrónica (eHealth) [10]. En concreto, el habla es un buen reflejo de la emoción y se han llegado a obtener mejores clasificadores, como el emotional coupling, que codifican emociones ricas en un espacio

dimensional pequeño [11]. Con el avance de estas técnicas, se podría incorporar la detección de emociones en diversas aplicaciones que revolucionarían la manera en que interactuamos con la tecnología y entre nosotros, mejorando nuestra calidad de vida y comunicación.

El análisis de voz es el estudio de las propiedades acústicas de la persona hablante para inferir información sobre el estado de las emociones y el estado mental de la persona hablante. Este procedimiento hace uso de tecnología capaz de medir la frecuencia y la variabilidad de la voz, las cuales son señales que tienen el nivel de tensión emocional de un individuo. En ese sentido, las tecnologías de agentes empáticos, como las que se plantean en las últimas novedades, se fundamentan en medidas indirectas de la respuesta fisiológica al estrés, siendo esencial el análisis de la voz para determinar las emociones humanas [12]. Además, la integración de prácticas tradicionales de medicina, como es el caso del Qigong, con el análisis de voz por IA está creando instrumentos para predecir patrones emocionales y adaptar las intervenciones médicas [13].

El análisis de voz con IA se convierte en una herramienta clave para reconocer emociones y representa un avance para la interacción humano-máquina. Esta tecnología permite descomponer la voz en sus partes constitutivas (tono, entonación, patrones de habla, etc.) para así reconocer emociones que antes eran difíciles de reconocer. En el mundo académico actualmente, los resultados muestran que los agentes artificiales no solo tienen una expresividad de tipo emocional, sino que además pueden provocar respuestas emocionales en las personas, lo cual está relacionado con la percepción emocional [14]. Además, casos como el de recuperar la información han destacado la necesidad de la IA en las soluciones de análisis de voz, creando conversaciones espaciales entre expertos de diferentes campos [5]. De ahí que el análisis por voz se erija como un nexo entre la tecnología y la emoción humana.

El estudio del análisis de voz con IA para identificar las emociones de los estudiantes es un campo que proporciona información sobre la emoción en el aula. Los investigadores han planteado que las herramientas de IA, con algoritmos de aprendizaje automático, pueden identificar con exactitud los estados emocionales a través de características de voz, como el tono, el ritmo y el timbre [2]. Esta habilidad de reconocimiento emocional sensible trasciende lo académico, impactando en la mejora de la participación, el rendimiento y la creación de ambientes de aprendizaje más receptivos [3], [15]. Con el creciente reconocimiento del aprendizaje socioemocional en las escuelas, la capacidad de la tecnología en el análisis de voz para identificar marcadores emocionales la hace una herramienta potencial para que los maestros ajusten sus métodos de enseñanza y aborden las necesidades emocionales de los estudiantes [4], [5]. Una parte importante de la revisión ha sido identificar las implicaciones éticas de las aplicaciones dadas de la IA en psicología de la educación. Los estudios indican la preocupación permanente de la privacidad y la seguridad de los datos, dada la sensibilidad en la información emocional que se procesa [9], [16]. Y como estas tecnologías van avanzando, es necesario crear bases éticas y hablar sobre cómo balancear los beneficios de las perspectivas que la IA puede aportar con la necesidad de proteger los datos emocionales de los estudiantes [7], [8].

Además, a pesar de los avances que se pueden percibir en la literatura actual, aún existen ciertas restricciones en el enfoque demográfico de los estudios actuales, donde se favorece la población adulta sobre la estudiantil [17]. Esto evidencia la necesidad de mayor investigación hacia la diversidad demográfica estudiantil para que la tecnología de evaluación emocional sea lo suficientemente sensible y contextualizada. El objetivo de la investigación fue determinar la capacidad de la inteligencia artificial (IA) para identificar emociones a través de la voz de estudiantes universitarios peruanos, un campo que ha tomado relevancia en diferentes ámbitos, desde la psicología hasta las tecnologías de comunicación.

II. Método

Para el análisis de voz con IA para el reconocimiento de emociones en estudiantes se utilizó un enfoque cuantitativo y experimental. Para ello, se realizó el recojo de audios mediante grabadoras de voz en una aplicación creada para ello. A continuación, el audio se preprocesará (normalización, eliminación de ruido y segmentación) y se extraerán características acústicas relevantes, tales como coeficientes cepstrales en escala mel (MFCC), tono fundamental (pitch), eliminación de ruido/silencios, energía, realizando una identificación de sub-bandas por cada audio. Estas características serán interpretadas por algoritmos de aprendizaje automático y aprendizaje profundo, previamente entrenados con conjuntos de datos etiquetados emocionalmente. El sistema categorizará las emociones en feliz, triste, enojado, alegre, neutral, sorpresa. Finalmente, los resultados se medirán en términos de precisión, exactitud y F1-score para verificar la efectividad del modelo. Este tipo de aproximación se ha usado en estudios de análisis emocional de voz con buenos resultados en entornos educativos [19].

El universo del estudio estuvo conformado por los estudiantes de una Universidad Nacional, mientras que la población se delimitó específicamente a la Facultad de Ingeniería de dicha universidad, integrada por las escuelas profesionales de Ingeniería Informática, Ingeniería Mecatrónica, Ingeniería Electrónica e Ingeniería de Telecomunicaciones, entendida esta población como el conjunto de casos que cumplen con especificaciones previamente definidas [20]. La muestra se seleccionó mediante un muestreo no probabilístico, considerando que “la muestra es cualquier subconjunto de la población”, y que en

este tipo de muestreo la elección de los participantes no depende del factor de probabilidad sino de las características del propio estudio y de los objetivos del propio investigador; en tal sentido, se tomó como muestra a estudiantes que cursan el primer año de la Facultad de Ingeniería.

La herramienta utilizada para la recogida de datos en el estudio fueron una aplicación de pc que pudieron descargar los participantes de captura de voz desarrollada como herramienta digital para grabar, guardar y etiquetar las muestras de voz junto con sus metadatos (estudiante, duración); micrófonos direccionales o auriculares con micrófono integrado que cumplan con unos mínimos estándares de calidad técnica (frecuencia de muestreo ≥16 kHz y 16 bits de resolución) para garantizar la fidelidad acústica. Un modelo de inteligencia artificial para clasificación emocional, que funcionó como instrumento de análisis indirecto al transformar las características acústicas de la voz en etiquetas emocionales; y una base de datos estructurados, concebido como sistema centralizado de almacenamiento para los audios procesados, los resultados de la clasificación emocional, facilitando así la organización, recuperación y análisis integral de la información recolectada.

La validez del sistema se sustenta, en primer lugar, en la validez de contenido, dado que fue diseñado para captar características acústicas ampliamente reconocidas en la literatura científica como representativas de estados emocionales, tales como los coeficientes cepstrales en las frecuencias de Mel (MFCC), el pitch, la intensidad y otros descriptores prosódicos; en segundo lugar, en la validez de criterio, que puede establecerse mediante la comparación de las salidas del sistema con escalas validadas de autoinforme emocional o con anotaciones realizadas por expertos humanos; y, finalmente, en la validez interna, garantizada a través de la adecuada selección de variables acústicas, el entrenamiento del modelo con datasets previamente etiquetados y la aplicación de estrategias para minimizar sesgos en el modelo de inteligencia artificial. La confiabilidad del sistema se evalúa mediante la consistencia del modelo de IA, utilizando métricas de rendimiento como exactitud (accuracy), sensibilidad (recall), especificidad y F1-score, así como a través de pruebas piloto repetidas bajo condiciones similares para verificar que la detección emocional se mantiene estable en el tiempo. La exactitud mide qué tan bien el sistema acierta en la identificación de emociones en comparación con una etiqueta real o referencia externa, mientras que la precisión indica cuántas de las emociones clasificadas como una categoría específica, por ejemplo “alegría”, corresponden efectivamente a dicha emoción, lo que resulta clave para evitar falsos positivos. Para evaluar estas métricas se emplean instrumentos como la matriz de confusión, que permite visualizar los aciertos y errores por categoría emocional, la curva ROC/AUC para analizar el comportamiento global del clasificador emocional y las pruebas cruzadas (cross-validation), que estiman el rendimiento del modelo en distintos subconjuntos de datos y fortalecen la robustez de los resultados.

El análisis de datos de la investigación se inicia con la recolección de datos realizada, las grabaciones pueden realizarse en tiempo real o de forma diferida, según el diseño del sistema. Posteriormente, los audios son sometidos a un preprocesamiento que incluye la normalización del volumen, la eliminación de ruido de fondo, la segmentación para identificar partes relevantes de la grabación y la conversión a un formato estándar (por ejemplo, .wav, 16 kHz, mono). A continuación, se lleva a cabo la extracción de características acústicas que sirven como insumos para el modelo de inteligencia artificial, tales como los MFCC para capturar el timbre de la voz, el pitch o tono fundamental vinculado a diferentes emociones, los formantes (F1, F2, etc.), la energía de la señal, la tasa de habla y medidas como jitter y shimmer, sensibles a estados emocionales como ansiedad o tristeza. Sobre esta base, se realiza el análisis e interpretación de resultados mediante la visualización de emociones predominantes por audio y su distribución temporal, lo que permite identificar patrones emocionales en el contexto académico. Finalmente, se incorpora una fase de retroalimentación y mejora continua del sistema, evaluando el rendimiento del modelo con métricas como precisión, recall y F1-score, ajustando parámetros y reentrenando con nuevos datos cuando se detectan errores o baja sensibilidad, e integrando la retroalimentación docente para contextualizar adecuadamente los resultados y fortalecer la validez pedagógica del sistema.

II. Resultados.

A continuación, la figura 1 muestra la intensidad de voz promedio de los 4,723 audios recolectados de los estudiantes, en las diferentes emociones descritas para el estudio, como se puede apreciar, la emoción alegría tiene la mayor intensidad de voz reconocida de los estudiantes.

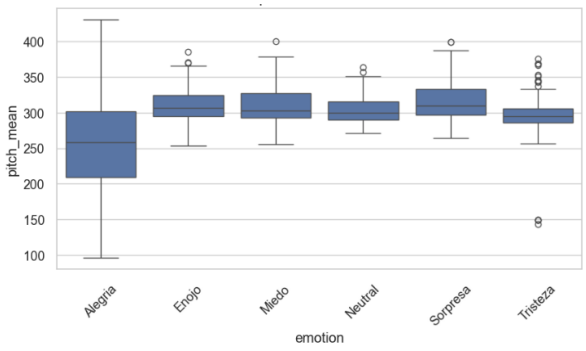

Fig. 1 Comparación de pitch medio por conjunto de datos.

La figura 2 muestra un espectrograma de un audio recolectado donde muestra que al inicio el participante no ha pronunciado nada y ha dejado un espacio inicial en silencio, en este caso, el preprocesamiento se encarga de evaluar los espacios vacíos al inicio y al final del audio para que luego estos puedan se eliminados, dejando solamente el espectrograma con frecuencias identificadas.

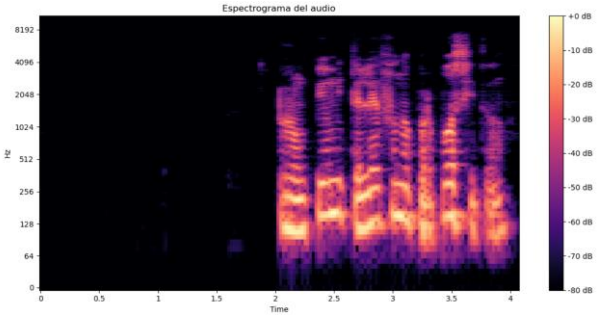

Fig 2. Ejemplo de espectrograma de audios de Enojo.

La figura 3 muestra un audio convertido a su espectrograma, a través de la frecuencia de los coeficientes cepstrales de Mel (MFCC), donde se puede apreciar que existe un ruido o distorsión del audio enviado por el participante, en este caso, adicional a la eliminación de espacios vacíos al inicio y al final, se identifica la frecuencia que origina la distorsión y se procede a eliminar para que el audio hay podido quedar en forma adecuada para el presente estudio. De no ser posible la eliminación de la interferencia, se procederá a la eliminación del audio.

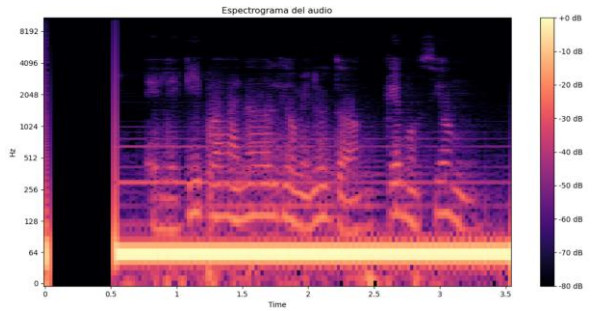

Fig 3. Ejemplo de espectrograma de audios de Alegría con ruido.

La figura 4 muestra un espectrograma de un audio que muestra también un ruido en menor proporción, en este caso, por ser una interferencia de menor grado, se procede a dividir el audio en sub-bandas, con el objetivo de eliminar las bandas más débiles, de tal forma, que el audio quede limpio de ruido. Del mismo modo, se puede apreciar que al inicio hay un tiempo sin datos, el cual se eliminará, maximizando el espacio del audio dentro del espectrograma.

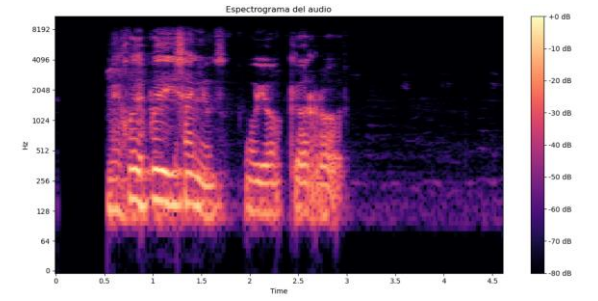

Fig 4. Ejemplo de espectrograma de tristeza.

La figura 5 muestra la compresión de audios, en donde se van a eliminar aquellos picos altos de intensidad del audio, mediante una escala 3 a 1 pasado el umbral permitido, por ejemplo, si el umbral es 10 decibeles, entonces pasados los 10 decibeles de intensidad del audio, en un instante de tiempo, solo se considerará 1 decibel adicional por cada 3. De este modo, si existen picos de audio de 16 decibeles, solo se consideran los diez primeros, más 1 decibel por cada 3 adicionales, por lo tanto, en total el audio quedaría al final con 12 decibeles en ese pico más alto que fue identificado sobre el umbral permitido. De esta forma no se pierde calidad en los audios, sino que se eliminan los picos para que exista una mejor homogeneidad entre los datos.

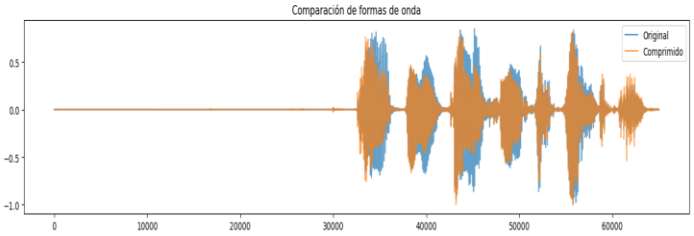

Figura 5. Compresión de audio.

Para la obtención del análisis de disponibilidad de audios por emoción se muestra variaciones significativas en la proporción de archivos válidos según su duración. Para Alegría se contaron 779 audios originales, de los cuales solo 203 fueron válidos (26.1%), requiriéndose 147 archivos adicionales mediante data augmentation; en Enojo se registraron 772 originales y 318 válidos (41.2%), siendo necesarios 32 adicionales; en Miedo hubo 766 originales y 340 válidos (44.4%), por lo que se requieren únicamente 10 archivos más; en Neutral se identificaron 778 audios, con 298 válidos (38.3%) y la necesidad de generar 52 audios adicionales; en Sorpresa se contabilizaron 772 originales y 320 válidos (41.5%), requiriéndose 30 adicionales; finalmente, en Tristeza se registraron 781 audios originales, de los cuales 211 fueron válidos (27.0%), siendo necesario generar 139 archivos mediante data augmentation. Estos resultados evidencian diferencias importantes en la calidad y duración de los audios entre emociones, lo que justifica la aplicación de técnicas de aumento de datos para equilibrar el conjunto de entrenamiento.

El análisis de la base de audios evidencia diferencias importantes entre la cantidad de archivos originales y aquellos que cumplen con la duración y calidad necesarias para su uso en el entrenamiento del modelo. Para cada emoción se estableció un conjunto final balanceado de 350 audios, lo que implicó utilizar el 100% de los archivos válidos disponibles y complementar el resto mediante técnicas de data augmentation de ser el caso. En el caso de Alegría, de 779 audios originales solo 203 fueron válidos, requiriéndose la generación de 147 audios adicionales. Enojo presentó 772 audios originales y 318 válidos, por lo que fue necesario aumentar 32 registros; Miedo

contó con 766 originales y 340 válidos, requiriendo únicamente 10 adicionales; mientras que Neutral, con 778 originales y 298 válidos, necesitó 52 archivos generados. La emoción Sorpresa presentó 772 originales y 320 válidos, por lo que se añadieron 30 audios para completar el total. Finalmente, Tristeza registró 781 audios originales y 211 válidos, lo que demandó 139 archivos aumentados, siendo esta la categoría con mayor déficit de datos válidos. Estos resultados evidencian la necesidad de aplicar estrategias de aumentación para asegurar un conjunto equilibrado y homogéneo por emoción, permitiendo así un entrenamiento más robusto y representativo del modelo de reconocimiento emocional.

A continuación, la tabla 1, indica las características que tiene solo el modelo de la red neuronal, la cual fue aplicada después de los preprocesamientos realizados para la limpieza de audios.

TABLA I
RESUMEN DE LA RED CONVOLUCIONAL

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 13, 137, 32)	896
batch_normalization (BatchNormalization)	(None, 13, 137, 32)	128
max_pooling2d (MaxPooling2D)	(None, 7, 69, 32)	0
dropout (Dropout)	(None, 7, 69, 32)	0
conv2d_1 (Conv2D)	(None, 7, 69, 64)	18,496
batch_normalization_1 (BatchNormalization)	(None, 7, 69, 64)	256
max_pooling2d_1 (MaxPooling2D)	(None, 4, 35, 64)	0
dropout_1 (Dropout)	(None, 4, 35, 64)	0
conv2d_2 (Conv2D)	(None, 4, 35, 128)	73,856
batch_normalization_2 (BatchNormalization)	(None, 4, 35, 128)	512
max_pooling2d_2 (MaxPooling2D)	(None, 2, 18, 128)	0
dropout_2 (Dropout)	(None, 2, 18, 128)	0
flatten (Flatten)	(None, 4608)	0
dense (Dense)	(None, 256)	1,179,904
dropout_3 (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 6)	1,542

La evaluación del modelo se llevó a cabo utilizando un conjunto de pruebas de 350 muestras equilibradas, representadas por seis emociones distintas. La alegría, el enojo, el miedo, la neutralidad, la sorpresa y la tristeza. La eficacia del modelo fue evaluada a través de las métricas estándar que corresponden a la clasificación: precisión, recall y puntaje F1-score, adicionalmente de la exactitud global. Estas métricas señalan un equilibrio razonable entre las categorías, sin una predominancia evidente por alguna categoría, dado que el conjunto de datos se encontraba equilibrado por diseño.

TABLA II
RESULTADOS DEL ENTRENAMIENTO

Reporte de clasificación:

	precision	recall	f1-score	support
Alegria	0.96	0.92	0.94	105
Enojo	0.96	0.86	0.90	105
Miedo	0.79	0.92	0.85	105
Neutral	0.96	0.91	0.94	105
Sorpresa	0.95	0.84	0.89	105
Tristeza	0.88	1.00	0.94	105
accuracy			0.91	630
macro avg	0.92	0.91	0.91	630
weighted avg	0.92	0.91	0.91	630

La optimización del rendimiento se dio en las emociones Neutral, Enojo y Alegría, con un promedio de F1-score de 0.96, lo que indica que el modelo es capaz de reconocer estas emociones de manera precisa y consistente.

Prestación intermedia: La emoción Tristeza fue la que mejor rindió (0.8 F1), lo que sugiere que esta emoción tiene huellas acústicas muy distintas.

Miedo obtuvo el menor F1-score (0.79), lo que indica dificultades para identificar esta emoción, posiblemente porque se parece mucho en su sonoridad a otras emociones como Sorpresa o Tristeza, ya que comparten características acústicas (ej., tono alto, energía repentina).

La matriz de confusión (figura 6), permite visualizar el grado de identificación de los audios obtenidos por los participantes con la emoción asociada.

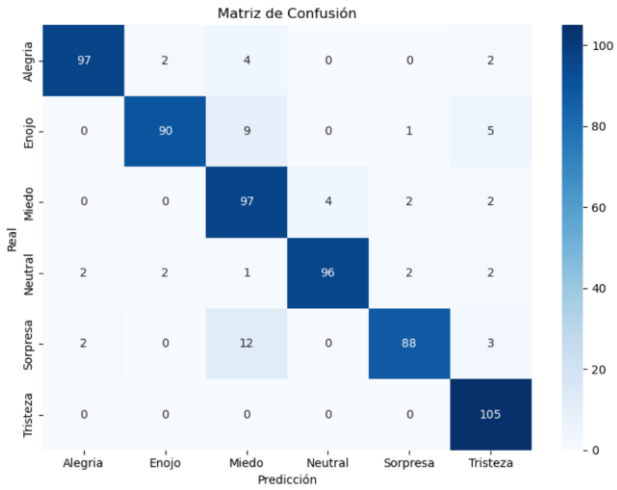

Fig 6. Matriz de confusión.

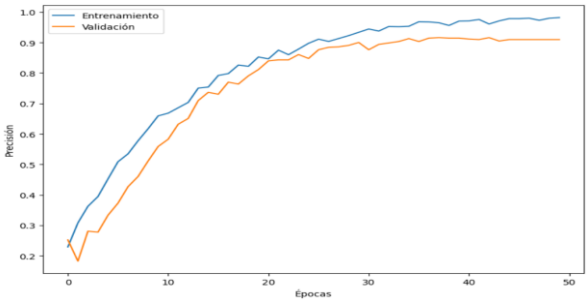

Fig 7. Precisión durante el entrenamiento

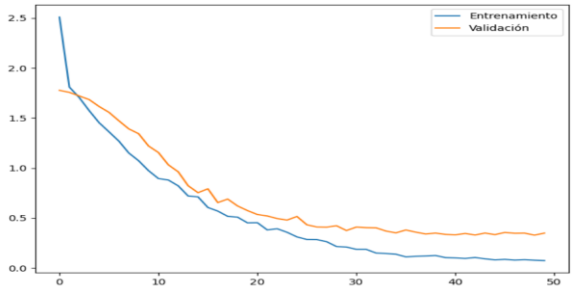

Fig 8. Pérdida durante el entrenamiento.

Análisis técnico de la figura 7.

Fase de inicio (épocas 0–10):

La precisión de entrenamiento inicia alrededor del 25% y muestra un incremento rápido, alcanzando aproximadamente 65% hacia la época 10.

La curva de validación inicia ligeramente por debajo (≈23%) y presenta fluctuaciones leves durante las primeras épocas, algo típico mientras el modelo ajusta pesos iniciales.

A partir de la época 5–6, ambas curvas comienzan a mostrar una tendencia ascendente sostenida, señal de que el modelo empieza a capturar patrones relevantes en los datos.

En esta fase el modelo aprende rasgos básicos, disminuye el error y se estabilizan los gradientes. La sincronía progresiva entre ambas curvas sugiere un inicio de entrenamiento saludable.

2. Fase media (épocas 10–25):

Entre las épocas 12 y 18 ocurre una convergencia parcial: ambas curvas se mantienen cercanas, con precisiones entre 70% y 85%.

No se observa una separación brusca entre entrenamiento y validación, lo cual indica que el modelo aún no presenta sobreajuste significativo.

La progresión es suave y ascendente, sin picos ni caídas abruptas, lo que evidencia que el modelo ha alcanzado una fase de aprendizaje eficiente.

En este intervalo el modelo comprende las características más representativas del conjunto, mejora la generalización y se afianza la estabilidad del entrenamiento.

3. Fase final (épocas 25–50):

La precisión de entrenamiento continúa incrementándose de forma casi monótona hasta alcanzar valores cercanos al 98%, lo que demuestra que el modelo puede ajustar muy bien los datos de entrenamiento.

La curva de validación se estabiliza alrededor de 90–91%, mostrando ligeras oscilaciones, pero sin descensos drásticos.

Se mantiene una brecha moderada entre ambas curvas (~7–8%), lo que representa un leve sobreajuste, considerado normal en modelos bien ajustados.

El modelo entra en una fase de refinamiento donde la capacidad de aprendizaje se acerca a su límite. La estabilidad de la validación indica buena capacidad de generalización y ausencia de sobreajustes severos.

Análisis técnico de la figura 8.

1. Fase inicial (épocas 0–10): La pérdida de entrenamiento inicia alrededor de 2.5, lo cual es esperable cuando los pesos del modelo están sin optimizar. Se observa un descenso rápido y pronunciado en las primeras épocas, señal de que el modelo propuesto está aprendiendo patrones fundamentales del conjunto de datos. La pérdida de validación comienza alrededor de 1.75, siguiendo una caída similar, pero menos pronunciada, a la del entrenamiento. Ambas curvas siguen patrones paralelos, lo que sugiere estabilidad en la optimización temprana. La sincronía de ambas curvas indica que no existen problemas de inicialización ni de mala configuración del learning rate.

2. Fase intermedia (10-25): La pérdida de entrenamiento sigue disminuyendo, aunque cada vez a menor ritmo, hasta estabilizarse en torno a 0.6-0.4 hacia la época 20. La pérdida de validación también se identifica que reduce, encontrándose alrededor de 0.8 a 0.5, lo que indica aprendizaje generalizable. No se aprecia separación brusca entre las dos curvas, lo que es señal de que no hay sobreajuste importante en esta fase. Las fluctuaciones en la curva de validación son normales y son un indicativo de la sensibilidad a la variabilidad del conjunto de validación. El modelo logra un buen equilibrio entre ajuste a los datos de entrenamiento y generalización a datos no vistos.

3. Fase final (épocas 25–50): La pérdida de entrenamiento desciende hasta valores tan bajos como 0.05, lo cual indica que el modelo logra un ajuste muy fino al conjunto de entrenamiento. La pérdida de validación se estabiliza entre 0.35 y 0.40, mostrando una tendencia relativamente plana a partir de la época 30. Se mantiene una brecha entre validación y entrenamiento, pero esta diferencia es estable y moderada, típica de modelos bien entrenados.

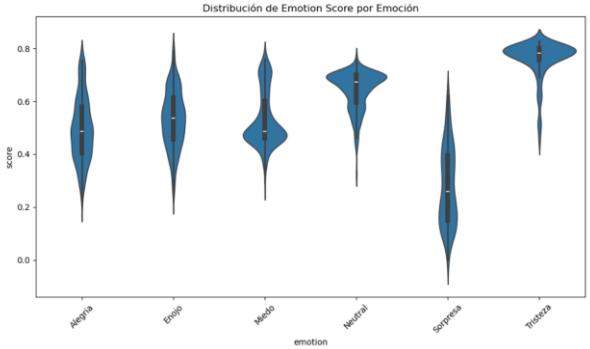

Fig 9. Datos agrupados por emoción.

Análisis técnico de la figura 9

La gráfica representa, mediante violin plots, la distribución del puntaje (score) asignado por el modelo a cada categoría emocional. Este tipo de visualización permite evaluar simultáneamente:

la densidad de los valores,

la dispersión (variabilidad interna),

la asimetría,

y la concentración central (mediana y cuartiles).

A continuación, el análisis por fases y por emoción:

1. Alegría: La distribución se concentra entre 0.40 y 0.65, con una mediana cercana a 0.50.

Presenta colas largas hacia valores bajos (<0.30), lo cual indica cierta incertidumbre del modelo en algunos casos. La forma simétrica refleja estabilidad, sin grandes desviaciones. El modelo identifica Alegría con fiabilidad media-alta, aunque algunos ejemplos generan scores bajos que podrían corresponder a muestras ambivalentes o ruidosas.

2. Enojo: Distribución muy similar a Alegría, con densidad entre 0.45 y 0.65, y mediana alrededor del 0.52. Tiene una base inferior más amplia, lo que indica más casos con scores moderados. El modelo distingue el enojo de forma adecuada, pero mantiene variabilidad, posiblemente por similitud acústica con otras emociones intensas (miedo, sorpresa).

3. Miedo: La distribución es más estrecha y se agrupa fuertemente entre 0.45 y 0.60, con menos dispersión que otras emociones. Se observa una ligera asimetría hacia valores altos. El modelo tiende a asignar scores más consistentes a Miedo, mostrando buena estabilidad interna. La menor dispersión sugiere que las muestras de esta categoría son más homogéneas.

4. Neutral: Presenta una concentración marcada en valores altos (0.65–0.75), con una mediana aproximada de 0.70. Es una de las distribuciones más compactas y con menor presencia de valores bajos. El modelo identifica la emoción Neutral con alta precisión y baja incertidumbre, probablemente por características acústicas más estables (tono plano, baja intensidad, baja variabilidad prosódica).

5. Sorpresa: La distribución es más amplia y dispersa: valores desde 0.15 hasta 0.50, con una mediana cerca de 0.28. Es la emoción con menor puntaje promedio, mostrando mayor dificultad de clasificación para el modelo. Sorpresa exhibe alta variabilidad acústica (puede ser positiva o negativa), lo cual hace que el modelo tenga más incertidumbre. Es la categoría que posiblemente requiera más datos o mejores características acústicas.

6. Tristeza: Distribución muy concentrada en la parte superior (0.75–0.82), con una mediana elevada (~0.80). Es la emoción mejor definida y con menor dispersión de todas. El modelo distingue Tristeza con alta robustez. La señal acústica de esta emoción suele tener patrones distintivos (bajo pitch, ritmo lento, energía reducida), facilitando su identificación.

Por lo cual, al final de la interpretación técnica se puede inferir que Los resultados obtenidos permiten afirmar que el modelo presenta un desempeño sólido y consistente con una alta precisión de entrenamiento (98%) y validación (91%), con una brecha controlada, reducción estable y progresiva de la pérdida sin signos de divergencia o inestabilidad; y distribuciones de puntajes coherentes, con tres emociones de alto desempeño (Tristeza, Neutral, Miedo), dos emociones de desempeño medio (Alegría, Enojo) y una emoción desafiante (Sorpresa).

A continuación, se muestran resultados con otros modelos que son utilizados para análisis de emociones y con el dataset obtenido para el presente estudio.

La Figura 10, nos indica los resultados obtenidos con Ramdom Forest, donde el Accuracy obtenido fue de 45%.

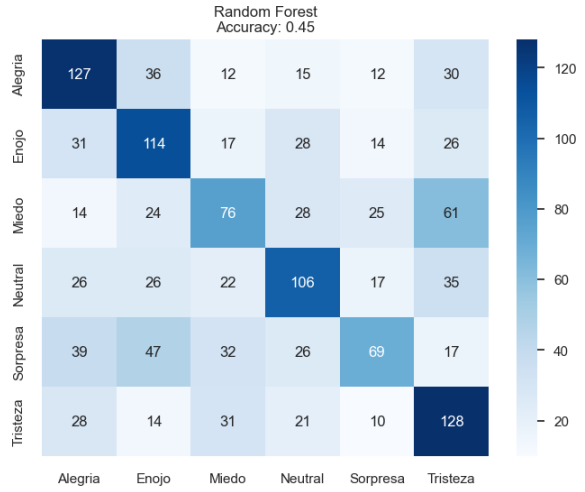

Fig 10. Matriz de confusión con modelo Random Forest.

La Figura 11, nos muestra los resultados con el modelo pre entrenado YAMNet, el cual es un modelo que identifica diversos estados de audios y emociones, con el cual se obtuvo solo un 36% de exactitud.

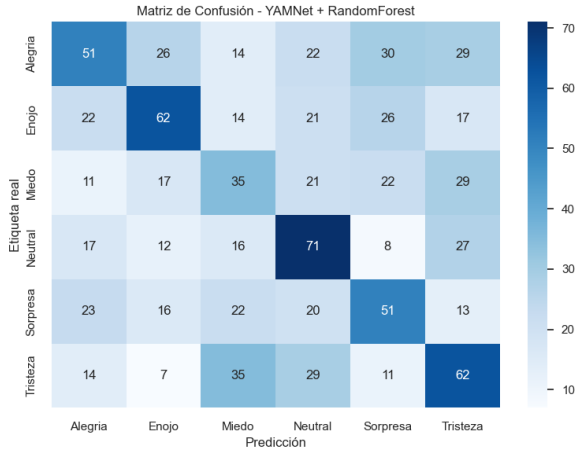

Fig 11. Matriz de confusión con modelo YAMNet + Random Forest.

La figura 12 muestra la matriz de confusión de los resultados obtenidos con SBC (Sub-Band Coding) para el reconocimiento de emociones, donde se obtuvieron resultados de 40% de precisión.

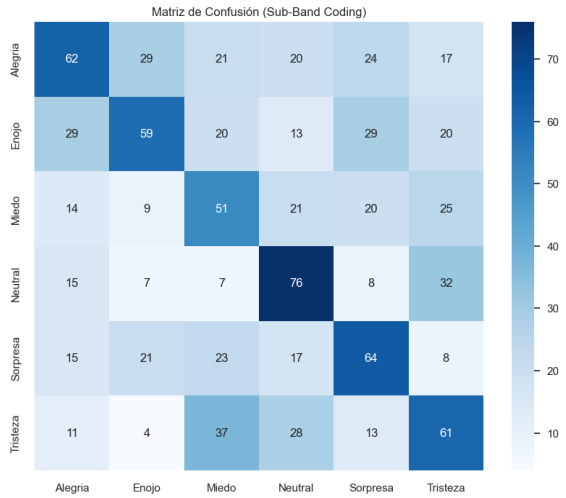

Fig 11. Matriz de confusión con modelo SBC.

IV. Discusión

Los resultados obtenidos en el presente estudio muestran un comportamiento consistente con las tendencias actuales en el reconocimiento de emociones en voz mediante modelos de aprendizaje profundo. La progresión estable de la precisión y la pérdida observada durante el entrenamiento sugiere que el modelo logra aprender representaciones discriminativas de manera eficaz y generalizable. Este comportamiento concuerda con [23], en que los modelos basados en deep representations tienden a tener curvas de entrenamiento estables y disminuciones sucesivas de pérdida al ser entrenados con buenas características espectro-temporales. Además, la pequeña diferencia entre la exactitud del entrenamiento (98%) y la del conjunto de validación (91%) indica un ligero sobreajuste, común en modelos profundos entrenados en voz. Según [24], este tipo de brechas no necesariamente afecta la capacidad de generalización si la curva de validación continúa en ascenso o se estabiliza, como en este caso. La falta de fluctuaciones bruscas o aumentos en la pérdida valida indica que la optimización fue estable y que el modelo encontró un buen equilibrio entre aprendizaje y generalización.

Por otro lado, el estudio de la distribución del emotion score reveló diferencias entre las categorías emocionales. Emociones como Tristeza y Neutral tuvieron una dispersión más pequeña y valores medianos altos, lo que sugiere que el modelo puede reconocer patrones acústicos consistentes relacionados con estas emociones. Estudios recientes también encontraron resultados similares como [25] que demostraron que las emociones con menor variabilidad prosódica y patrones más constantes son mejor reconocidas por los modelos CNN, mientras que Sorpresa tuvo mayor dispersión y mediana más baja, lo que sugiere dificultad para ser clasificada. Este comportamiento se alinea con la literatura, la cual indica que emociones con alta variabilidad en pitch y energía tienden a crear mayor incertidumbre en los clasificadores acústicos [24]. Estas diferencias indican que la emoción está menos representada en el conjunto de datos o que sus características acústicas son más variables, lo que dificulta su modelado.

Los hallazgos del estudio respaldan la evidencia científica actual y demuestran que los modelos profundos, al ser entrenados apropiadamente, pueden alcanzar alta precisión en tareas de reconocimiento emocional. Pero las diferencias entre categorías emocionales señalan la necesidad de explorar estrategias específicas de mejora, como data augmentation enfocada en las emociones más variables o la inclusión de características prosódicas adicionales.

En cuanto a la distribución emocional, Tristeza y Neutral presentan alta consistencia, lo que indica estados relativamente estables durante las sesiones; sin embargo, la presencia moderada de Alegría, Enojo y Miedo, junto con la variabilidad de Sorpresa, sugiere efectos potenciales sobre el aprendizaje: las emociones positivas se asocian con mayor participación, interés y retención, mientras que emociones negativas o altamente variables pueden introducir barreras cognitivas y afectar la atención o la comprensión del discurso docente [24]. Además, la predominancia de lo neutral puede reflejar un entorno estable pero no necesariamente estimulante; la Alegría, aunque moderada, apunta a momentos de motivación e interacción, mientras que Miedo y Enojo podrían indicar tensión o dificultades. La Sorpresa, pese a su variabilidad acústica, puede vincularse a activación cognitiva, descubrimiento y pensamiento crítico, funcionando como marcador de momentos clave cuando se introducen conceptos nuevos o actividades inesperadas [25].

La evidencia emocional recopilada indica que el análisis automático de emociones es una herramienta valiosa para comprender el estado emocional del grupo y su influencia en el proceso de enseñanza–aprendizaje. Al integrar estos hallazgos en la práctica pedagógica, el docente puede ajustar oportunamente sus estrategias didácticas, anticipar dificultades de comprensión o participación y fortalecer la creación de entornos de aprendizaje más inclusivos, efectivos y emocionalmente saludables.

V. Conclusiones

Según lo que se ha desarrollado en la presente investigación, se llega a la conclusión de que:

La creación de una base de datos de voz etiquetada por emociones contextualizadas en el ambiente universitario de una universidad pública en la ciudad de Lima, Perú, dio como

resultado 4,723 audios de emociones de alegría, tristeza, neutral, enojo, sorpresa y miedo clasificados por cada uno de ellos adecuadamente.

Se pudieron desarrollar algoritmos de preprocesamiento de voz en Python, utilizando diversas técnicas como eliminación de ruidos, silencios, compresión de audios, clasificación de tonos, etc., para extraer características, que permitan aplicar los MFCC y de esta forma, elaborar los espectrogramas para el trabajo con una red convolucional, las cuales se explican en el presente trabajo.

Se pudo entrenar un modelo a partir de audios recopilados con la finalidad de reconocer emociones en estudiantes que ingresan a una universidad pública en Lima, Perú. Este modelo tiene características adecuadas para esa ciudad, como su forma y voz.

Se pudo crear un prototipo de reconocimiento de voz con IA para mejorar el estado emocional de los estudiantes en una universidad, con un porcentaje de efectividad de reconocimiento de emociones en entrenamiento de 98% y 91% en validación.

REFERENCIAS

[1] Y. K. Dwivedi et al., "Defining the future of digital and social media marketing research: Perspectives and research proposals," International Journal of Information Management, vol. 59, p. 102168, 2020.

[2] B. S. et al., "Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review," 2024.

[3] M. Y. A., "The use of artificial intelligence in education for measuring student engagement, motivation, and emotional well-being," 2024.

[4] P. II et al., "Emotional fluctuations in academic settings," 2023.

[5] L. Azzopardi et al., "Report on the Information Retrieval Festival (IRFest2017)," 2017.

[6] J. Rudolph et al., "ChatGPT: ¿Un despilfarro o el fin de las evaluaciones tradicionales en la educación superior?" Revista de Aprendizaje y Enseñanza Aplicados, vol. 6, no. 1, 2023.

[7] S. C. Sivek, "Análisis ubicuo de emociones y cómo nos sentimos hoy," 2018.

[8] B. Grawemeyer et al., "Affective learning: improving participation and learning with conscious affect feedback," 2017.

[9] B. Gilleran, "Identification of fake news through emotion analysis," 2019.

[10] L. D. Ball et al., "Linking recorded data with emotive and adaptive computing in an eHealth environment," 2011.

[11] H. Atassi, "Emotion Recognition from Acted and Spontaneous Speech," 2014.

[12] E. L. van den Broek, Empathic Agent Technology (EAT), 2005. [Online]. Available: https://core.ac.uk/download/pdf/11482531.pdf

[13] B. A. Dethlefs, K. L. Lee, S. C. Li, and R. Loudon, Breathing Signature as Vitality Score Index Created by Exercises of Qigong: Implications of Artificial Intelligence Tools Used in Traditional Chinese Medicine, 2019. [Online]. Available: https://core.ac.uk/download/323075335.pdf

[14] E. S. Cross, F. Hekele, and R. Hortensius, The Perception of Emotion in Artificial Agents, 2018. [Online]. Available: https://core.ac.uk/download/157582956.pdf

[15] S. Abdul et al., "Implications of voice analysis technology for educators," 2024.

[16] A. Au et al., "Ethical considerations in AI applications in educational psychology," 2018.

[17] Y. E. Soelistio et al., "Towards automatic speech and emotion recognition of Indonesian (I-SpEAR)," 2017.

[18] I. L. C. Connon et al., "Critically Envisioning Biometric Artificial Intelligence in Law Enforcement," 2023.

[19] M. E. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[20] R. Hernández, C. Fernández, and M. del P. Baptista, Metodología de la investigación, 6th ed. México: McGraw-Hill Education, 2014.

[21] A. Barrasa et al., "Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review," 2024.

[22] F. Asaolu et al., "Enhancing mental health with Artificial Intelligence: Current trends and future prospects," 2024.

[23] S. Latif et al., "Survey of deep representation learning for speech emotion recognition," IEEE Transactions on Affective Computing, vol. 14, no. 2, pp. 1634–1654, 2023.

[24] H. Lian et al., "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," Entropy, vol. 25, no. 10, p. 1440, 2023.

[25] W. Zheng et al., "Identifying stable features for speech emotion recognition using convolutional neural networks," Sensors, vol. 21, no. 5, p. 1693, 2021.