Analizador de Emociones por Voz mediante Inteligencia Artificial

June 5, 2025

1 Analizador de Emociones por Voz mediante Inteligencia Artificial

1.1 Propuesta de Proyecto para Desarrollo de Asistente Psicológico Digital

1.1.1 1. Justificación

La salud mental representa uno de los desafíos más significativos de nuestro tiempo. Según la Organización Mundial de la Salud, más de 280 millones de personas sufren de depresión a nivel mundial, mientras que los trastornos de ansiedad afectan a 301 millones de individuos (WHO, 2021). En México, aproximadamente el 15% de la población experimenta algún trastorno mental a lo largo de su vida, pero solo el 20% de estos casos recibe atención profesional adecuada (Secretaría de Salud, 2020).

¿Qué motiva resolver este problema?

La principal motivación surge de la creciente necesidad de democratizar el acceso a herramientas de evaluación psicológica preliminar. Las barreras tradicionales incluyen costos elevados, disponibilidad limitada de profesionales especializados, estigma social asociado a la búsqueda de ayuda psicológica, y la dificultad para acceder a servicios en áreas rurales o comunidades marginadas (Kessler et al., 2019). La voz humana contiene información emocional rica y compleja que los profesionales de la salud mental han utilizado durante décadas como indicador diagnóstico (Schuller & Batliner, 2013). Parámetros como el tono, la velocidad del habla, las pausas, la intensidad y los patrones prosódicos pueden revelar estados emocionales específicos como ansiedad, depresión, estrés o estabilidad emocional.

¿Qué impacto tiene abordarlo desde IA?

La inteligencia artificial ofrece la capacidad de procesar y analizar características acústicas del habla de manera objetiva, consistente y escalable (El Ayadi et al., 2011). A diferencia de la evaluación humana, que puede verse influenciada por factores subjetivos, un sistema de IA puede proporcionar análisis reproducibles las 24 horas del día, los 7 días de la semana. El impacto potencial incluye la detección temprana de crisis emocionales, el monitoreo continuo del bienestar psicológico, la provisión de herramientas de autoevaluación accesibles, y la generación de datos objetivos que puedan complementar la evaluación clínica tradicional (Cowie et al., 2001).

¿Qué valor aporta replicarlo?

Replicar y mejorar tecnologías existentes de reconocimiento emocional por voz contribuye al avance del conocimiento científico en el área de computación afectiva (Picard, 2000). Además, desarrollar soluciones adaptadas al contexto cultural y lingüístico mexicano es fundamental, ya que las expresiones emocionales pueden variar significativamente entre diferentes culturas y idiomas (Russell,

1991). La creación de un sistema de código abierto y culturalmente apropiado puede servir como base para futuras investigaciones, facilitar la colaboración académica, y eventualmente contribuir al desarrollo de herramientas de telemedicina más efectivas y accesibles.

1.1.2 2. Descripción del Problema

Contexto del problema:

En la actualidad, la evaluación del estado emocional y psicológico de las personas depende principalmente de métodos subjetivos como cuestionarios de autoevaluación, entrevistas clínicas estructuradas, y la observación directa por parte de profesionales capacitados (American Psychiatric Association, 2013). Estos métodos, aunque efectivos, presentan limitaciones importantes en términos de disponibilidad, consistencia, y accesibilidad. Las personas que experimentan dificultades emocionales frecuentemente enfrentan barreras para acceder a evaluaciones psicológicas oportunas (Kazdin & Blase, 2011). Estas barreras incluyen listas de espera prolongadas en el sistema de salud público, costos prohibitivos en el sector privado, ubicación geográfica desfavorable, y el estigma social asociado con la búsqueda de ayuda profesional.

¿Qué se intenta resolver?

El problema central que aborda este proyecto es la falta de herramientas automatizadas, objetivas y accesibles para la evaluación preliminar del estado emocional de las personas. Específicamente, se busca desarrollar un sistema capaz de analizar grabaciones de voz y identificar patrones emocionales que puedan indicar estados como felicidad, tristeza, enojo, miedo, sorpresa, disgusto, o neutralidad emocional (Ekman, 1992). El sistema debe ser capaz de procesar audio en español mexicano, reconocer características prosódicas y acústicas relevantes, y proporcionar una evaluación emocional confiable que pueda ser utilizada como herramienta de apoyo en contextos clínicos, educativos, o de bienestar personal.

¿Quién se beneficiaría?

Los beneficiarios primarios incluyen individuos que buscan herramientas de autoevaluación emocional accesibles, profesionales de la salud mental que requieren herramientas objetivas de apoyo diagnóstico, instituciones educativas interesadas en monitorear el bienestar emocional de estudiantes, y organizaciones que desean implementar programas de bienestar laboral (Cowie et al., 2001). Los beneficiarios secundarios abarcan investigadores en el campo de la computación afectiva, desarrolladores de aplicaciones de salud mental, y la comunidad científica interesada en el procesamiento de señales de audio y reconocimiento de patrones emocionales.

1.1.3 3. Objetivo General

Desarrollar un sistema de inteligencia artificial capaz de reconocer y clasificar emociones humanas a partir del análisis de características acústicas y prosódicas del habla, con el propósito de crear una herramienta de evaluación emocional preliminar que pueda servir como base para un asistente psicológico digital.

1.1.4 4. Objetivos Específicos

1. Integrar y preprocesar múltiples conjuntos de datos de emociones en audio para crear un dataset robusto y diverso que incluya muestras en español e inglés, asegurando la representación equilibrada de diferentes estados emocionales y características demográficas.

2. Extraer y analizar características acústicas relevantes del habla emocional, incluyendo parámetros prosódicos (tono, intensidad, velocidad), espectrales (coeficientes MFCC, formantes), y temporales (pausas, duración de fonemas) que sean discriminativas para la clasificación emocional.
3. Diseñar, entrenar y optimizar modelos de aprendizaje automático especializado en reconocimiento de emociones por voz, evaluando diferentes arquitecturas (redes neuronales profundas, máquinas de vectores de soporte, bosques aleatorios) para identificar la configuración óptima.
4. Evaluar el rendimiento del sistema mediante métricas estándar de clasificación (precisión, recall, F1-score, matriz de confusión) y validación cruzada, estableciendo benchmarks de desempeño comparables con el estado del arte en reconocimiento emocional por voz.
5. Implementar técnicas de reducción de dimensionalidad y visualización para optimizar el procesamiento de características de alta dimensionalidad y facilitar la interpretación de los patrones emocionales identificados por el modelo.

1.1.5 5. Metodología

La metodología propuesta sigue un enfoque sistemático de desarrollo de sistemas de aprendizaje automático, estructurado en las siguientes etapas secuenciales:

- **Etapa 1: Obtención y Consolidación de Datos** Esta fase inicial involucra la descarga, exploración y consolidación de los tres conjuntos de datos seleccionados: Mexican Emotional Speech Database (MESD), RAVDESS Emotional Speech Audio, y Speech Emotion Recognition (EN) (Livingstone & Russo, 2018; Burkhardt et al., 2005). Se realizará un inventario detallado de cada dataset, documentando la estructura de archivos, metadatos disponibles, calidad de audio, y distribución de clases emocionales.

- **Etapa 2: Análisis Exploratorio Exhaustivo** Se implementará un análisis exploratorio de datos comprensivo que incluya la caracterización estadística de las señales de audio, identificación de patrones emocionales mediante visualizaciones especializadas, análisis de la distribución temporal y espectral de las muestras, y detección de posibles sesgos o desequilibrios en los datos (Tukey, 1977).

- **Etapa 3: Preprocesamiento Avanzado de Señales de Audio** Esta etapa comprende la normalización de la calidad de audio entre diferentes fuentes, segmentación de archivos largos en ventanas temporales apropiadas, filtrado de ruido y artefactos no deseados, y estandarización de formatos de archivo y frecuencias de muestreo (Rabiner & Schafer, 2010).

- **Etapa 4: Extracción de Características Discriminativas** Se implementará la extracción de múltiples tipos de características acústicas, incluyendo coeficientes MFCC (Mel-Frequency Cepstral Coefficients), características prosódicas como pitch, intensidad y velocidad del habla, características espectrales como centroide espectral y ancho de banda, y características temporales como duración de pausas y variabilidad del ritmo (Logan, 2000).

- **Etapa 5: Selección y Entrenamiento de Modelos** Se evaluarán múltiples arquitecturas de aprendizaje automático, comenzando con modelos tradicionales como SVM y Random Forest, progresando hacia redes neuronales profundas especializadas en procesamiento de señales de audio (Cortes & Vapnik, 1995; Breiman, 2001). Se implementará validación cruzada estratificada para asegurar la robustez de los resultados.

- **Etapa 6: Evaluación Comprehensiva y Optimización** La evaluación incluirá métri-

cas de clasificación estándar, análisis de matrices de confusión para identificar patrones de error específicos, evaluación del rendimiento por clase emocional individual, y pruebas de generalización con datos no vistos durante el entrenamiento (Kohavi, 1995).

- **Etapa 7: Documentación y Presentación de Resultados** La fase final involucra la documentación completa del proceso de desarrollo, análisis crítico de los resultados obtenidos, identificación de limitaciones y oportunidades de mejora, y preparación de material de presentación que comunique efectivamente los hallazgos del proyecto.

1.1.6 6. Obtención de Datasets

Para este proyecto se han identificado tres conjuntos de datos públicos especializados en reconocimiento emocional por voz, cada uno aportando características únicas que enriquecerán el entrenamiento del modelo:

Dataset 1: Mexican Emotional Speech Database (MESD)

- **Fuente:** Kaggle (https://www.kaggle.com/datasets/saurabhshahane/mexican-emotional-speech-database-mesd)
- **Tipo de datos:** Audio (archivos WAV)
- **Cantidad aproximada:** 864 grabaciones de audio
- **Características principales:** Grabaciones en español mexicano con 6 emociones básicas (alegría, tristeza, enojo, miedo, sorpresa, disgusto) más neutralidad (Shahane et al., 2021).
- **Utilidad específica:** Este dataset es fundamental para nuestro proyecto ya que proporciona muestras en español mexicano, permitiendo que el modelo se adapte específicamente a las características prosódicas y fonéticas del habla mexicana. La inclusión de este dataset asegura que el sistema sea culturalmente apropiado y efectivo para la población objetivo.

Dataset 2: RAVDESS Emotional Speech Audio

- **Fuente:** Kaggle (https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio)
- **Tipo de datos:** Audio (archivos WAV de alta calidad)
- **Cantidad aproximada:** 1,440 grabaciones vocales
- **Características principales:** Grabaciones profesionales en inglés norteamericano con 8 emociones diferentes, realizadas por 24 actores profesionales (12 hombres y 12 mujeres) (Livingstone & Russo, 2018).
- **Utilidad específica:** RAVDESS aporta grabaciones de calidad profesional que servirán como referencia para establecer benchmarks de rendimiento. La diversidad demográfica y la calidad controlada de las grabaciones proporcionan un estándar de comparación robusto para evaluar la efectividad de nuestro modelo.

Dataset 3: Speech Emotion Recognition (EN)

- **Fuente:** Kaggle (https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en)
- **Tipo de datos:** Audio (archivos WAV)
- **Cantidad aproximada:** 2,800 muestras de audio
- **Características principales:** Compilación diversa de grabaciones emocionales en inglés provenientes de múltiples fuentes, incluyendo situaciones más naturales y espontáneas (Babko, 2020).

- **Utilidad específica:** Este dataset complementa los anteriores al proporcionar variabilidad en estilos de habla y contextos de grabación. La inclusión de muestras más naturales y menos controladas ayudará al modelo a generalizar mejor hacia situaciones reales de uso, donde las condiciones de grabación pueden no ser ideales.

La combinación de estos tres datasets crea un corpus de entrenamiento robusto que abarca diferentes idiomas, calidades de grabación, y contextos culturales, proporcionando al modelo la diversidad necesaria para desarrollar capacidades de reconocimiento emocional generalizables y efectivas.

1.1.7 7. Análisis Exploratorio (EDA - Exploratory Data Analysis)

El análisis exploratorio de datos constituye una fase fundamental para comprender las características intrínsecas de nuestros conjuntos de datos de audio emocional (Tukey, 1977). Durante esta etapa, buscaremos responder preguntas clave sobre la naturaleza de las señales de audio, identificar patrones emocionales distintivos, y detectar posibles desafíos que deberán abordarse durante el preprocesamiento.

Preguntas de investigación principales: * ¿Existen diferencias espectrales consistentes entre diferentes estados emocionales? (Scherer, 2003) * ¿Cómo varían las características prosódicas (pitch, intensidad, velocidad) entre emociones? * ¿Qué nivel de variabilidad existe dentro de cada categoría emocional? * ¿Hay sesgos demográficos o técnicos en los datasets que deban considerarse?

Patrones que deseamos identificar: Buscaremos identificar firmas acústicas características de cada emoción, correlaciones entre características de audio y etiquetas emocionales, distribuciones de duración y calidad de las grabaciones, y posibles agrupaciones naturales de muestras similares que puedan informar la estrategia de modelado (Schuller & Batliner, 2013).

Gráficas Seleccionadas para el Análisis:

1. **Histogramas de Características Prosódicas**
   - **Utilización:** Utilizaremos histogramas para analizar la distribución de características fundamentales como pitch medio, intensidad RMS, y velocidad del habla (palabras por minuto) segmentadas por categoría emocional.
   - **Propósito:** Estos histogramas nos permitirán identificar si existen diferencias estadísticamente significativas en estos parámetros entre diferentes emociones, y si las distribuciones son unimodales o multimodales.
2. **Mapas de Calor (Heatmaps) de Correlación Espectral**
   - **Utilización:** Implementaremos mapas de calor para visualizar las correlaciones entre diferentes bandas de frecuencia y categorías emocionales.
   - **Propósito:** Esto nos ayudará a identificar qué regiones del espectro de frecuencias son más discriminativas para cada emoción, y si existen patrones espectrales únicos que caractericen estados emocionales específicos.
3. **Gráficas de Dispersión de Componentes Principales**
   - **Utilización:** Aplicaremos análisis de componentes principales (PCA) a las características extraídas y visualizaremos los resultados mediante gráficas de dispersión coloreadas por emoción (Jolliffe, 2002).
   - **Propósito:** Esta visualización nos permitirá evaluar la separabilidad natural de las diferentes clases emocionales en el espacio de características reducido, y identificar si existen agrupaciones claras o si hay solapamiento significativo entre categorías.

Análisis descriptivo detallado: Para cada gráfica generada, realizaremos un análisis estadístico

descriptivo que incluya medidas de tendencia central, dispersión, y forma de distribución. Documentaremos observaciones sobre la separabilidad de clases, la presencia de valores atípicos, y la necesidad de transformaciones de datos adicionales. Estos hallazgos informarán directamente las decisiones de preprocesamiento y selección de modelo en las etapas subsecuentes del proyecto.

1.1.8 8. Preprocesamiento de los Datos

El preprocesamiento de datos de audio emocional requiere técnicas especializadas para abordar los desafíos únicos que presentan las señales acústicas (Rabiner & Schafer, 2010). Nuestro enfoque se centrará en normalizar la calidad y características de las grabaciones para optimizar el rendimiento del modelo de reconocimiento emocional.

- **Normalización de Audio y Estandarización de Formato:** Todas las grabaciones serán convertidas a un formato estándar de 16 kHz de frecuencia de muestreo y 16 bits de resolución, asegurando consistencia entre los diferentes datasets. Se implementará normalización de amplitud para equilibrar los niveles de volumen entre grabaciones que puedan haber sido capturadas con diferentes configuraciones de ganancia.
- **Eliminación de Valores Nulos y Datos Corruptos:** Se desarrollará un proceso de validación automática para identificar archivos de audio corruptos, grabaciones con duración insuficiente (menor a 1 segundo), o archivos con contenido de audio inválido. Las muestras que no cumplan con los criterios de calidad mínimos serán removidas del conjunto de entrenamiento, documentando el proceso para mantener la trazabilidad.
- **Filtrado de Ruido y Artefactos:** Se aplicarán técnicas de filtrado digital para remover ruido de fondo, artefactos de grabación, y frecuencias no relevantes para el análisis emocional (Benesty et al., 2008). Esto incluye filtros paso-alto para eliminar ruido de baja frecuencia, filtros paso-bajo para remover componentes espectrales irrelevantes, y técnicas de reducción de ruido espectral cuando sea necesario.
- **Segmentación Temporal y Ventaneado:** Las grabaciones largas serán segmentadas en ventanas temporales de duración fija (típicamente 2-4 segundos) con solapamiento parcial para aumentar el número de muestras de entrenamiento. Se implementará detección de actividad vocal (VAD) para identificar y extraer únicamente los segmentos que contengan habla activa, eliminando pausas prolongadas y silencios no informativos.
- **Normalización Espectral y Balanceado de Clases:** Se aplicará normalización de media cero y varianza unitaria a las características extraídas para asegurar que diferentes tipos de características tengan rangos comparables. Adicionalmente, se implementarán técnicas de balanceado de clases como sobremuestreo SMOTE o submuestreo estratificado para abordar posibles desequilibrios en la distribución de emociones en el dataset consolidado (Chawla et al., 2002).

Estas técnicas de preprocesamiento son esenciales para nuestro proyecto porque las señales de audio emocional están inherentemente afectadas por variaciones en calidad de grabación, condiciones ambientales, y características individuales de los hablantes. La normalización y limpieza cuidadosa de los datos asegurará que el modelo pueda aprender patrones emocionales genuinos en lugar de artefactos técnicos irrelevantes.

1.1.9 9. Reducción de Dimensionalidad

Por qué podría ser útil reducir dimensiones:

En el contexto del reconocimiento emocional por voz, la extracción de características de audio

típicamente genera espacios de alta dimensionalidad que pueden contener información redundante o irrelevante (Guyon & Elisseeff, 2003). Las características acústicas como los coeficientes MFCC, características espectrales, y parámetros prosódicos pueden resultar en vectores de cientos o miles de dimensiones. La reducción de dimensionalidad se vuelve útil para mitigar el problema de la "maldición de la dimensionalidad", mejorar la eficiencia computacional durante el entrenamiento y la inferencia, reducir el riesgo de sobreajuste al eliminar características ruidosas o irrelevantes, y facilitar la visualización e interpretación de patrones emocionales en el espacio de características.

Técnicas que podrían aplicarse:

- **Análisis de Componentes Principales (PCA):** Esta técnica lineal será útil para identificar las direcciones de máxima varianza en nuestro espacio de características acústicas (Jolliffe, 2002). PCA permitirá retener el 95% de la varianza explicada mientras reduce la dimensionalidad significativamente, lo cual es valioso con características espectrales de alta dimensión.
- **Análisis Discriminante Lineal (LDA):** Dado que es un problema de clasificación supervisada con clases emocionales definidas, LDA será apropiado ya que busca direcciones que maximicen la separación entre clases y minimicen la varianza intra-clase (Fisher, 1936). Esto es ideal para maximizar la discriminabilidad entre estados emocionales.
- **t-SNE (t-Distributed Stochastic Neighbor Embedding):** Esta técnica no lineal será valiosa para visualización y exploración de datos, revelando estructuras de agrupamiento complejo que métodos lineales podrían no capturar (Van der Maaten & Hinton, 2008). t-SNE ayudará a entender si las emociones forman clusters naturales y si existen transiciones graduales entre estados emocionales.
- **UMAP (Uniform Manifold Approximation and Projection):** Como alternativa más eficiente a t-SNE, UMAP puede preservar tanto la estructura local como global de los datos, importante para entender relaciones jerárquicas entre emociones (McInnes et al., 2018).

En qué parte del proyecto la aplicarían:

La reducción de dimensionalidad se aplicará en dos momentos estratégicos del proyecto. Primero, durante la fase de análisis exploratorio, se usarán PCA y t-SNE para visualizar la estructura de los datos emocionales, informando decisiones de preprocesamiento y selección de características. Segundo, antes del entrenamiento del modelo final, se aplicará la técnica de reducción de dimensionalidad de mejor rendimiento (probablemente LDA) para optimizar el espacio de características. Esto ocurrirá después del preprocesamiento completo de los datos pero antes de la división en conjuntos de entrenamiento y validación, asegurando una aplicación consistente.

1.1.10 10. Fundamentación Teórica

a) Definición del Algoritmo Principal Para este proyecto, se emplearán **Redes Neuronales Convolucionales (CNN)** especializadas en procesamiento de audio como algoritmo principal de clasificación (LeCun & Bengio, 1995). Las CNNs son una clase de redes neuronales profundas diseñadas para procesar datos con estructura de grilla, como imágenes o, en este caso, espectrogramas de audio.

Una CNN aplica filtros convolucionales (kernels) que se deslizan sobre la representación espectral del audio, detectando características locales relevantes como patrones de frecuencia, transiciones espectrales y estructuras temporales. Estas redes están compuestas por capas convolucionales que extraen características de bajo nivel, capas de pooling que reducen la dimensionalidad espacial,

y capas completamente conectadas que realizan la clasificación final. En el reconocimiento emocional por voz, las CNNs procesan espectrogramas Mel o representaciones MFCC como imágenes bidimensionales (tiempo y frecuencia), permitiendo que la red aprenda automáticamente patrones discriminativos en el dominio tiempo-frecuencia.

b) Justificación de su Uso Las CNNs son adecuadas para este tipo de datos porque las emociones en el habla se manifiestan como patrones espacio-temporales en el espectrograma de audio (Schuller & Batliner, 2013). Estos patrones incluyen variaciones en la energía espectral, modulaciones temporales del pitch y estructuras armónicas características que una CNN puede identificar automáticamente.

La elección de CNNs se basa en múltiples casos de éxito documentados. Zhao et al. (2019) demostraron que las CNNs superan a métodos tradicionales como SVM en reconocimiento emocional por voz, con precisiones superiores al 85% en datasets estándar. Mirsamadi et al. (2017) mostraron que las CNNs pueden capturar características locales y patrones temporales de largo alcance en señales de habla emocional.

Para los datasets específicos (español mexicano e inglés con diferentes calidades), las CNNs ofrecen invarianza translacional, reconociendo patrones emocionales independientemente de su posición temporal exacta, lo cual es crucial con grabaciones naturales.

c) Ventajas Las principales ventajas de usar CNNs incluyen: * **Extracción automática de características:** Eliminan la necesidad de ingeniería manual de características. * **Robustez ante variaciones locales:** Toleran pequeñas variaciones en el tiempo, tono o calidad de grabación. * **Escalabilidad computacional:** Son superiores a SVM con grandes volúmenes de datos, ya que el entrenamiento puede paralelizarse en GPUs. * **Interpretabilidad parcial:** A través de mapas de activación, se puede entender qué regiones del espectrograma son más relevantes. * **Capacidad de generalización jerárquica:** Aprenden características de bajo nivel y las combinan en patrones complejos.

d) Limitaciones Las CNNs presentan varias limitaciones: * **Dependencia de grandes volúmenes de datos:** Requieren miles de muestras etiquetadas para evitar el sobreajuste. Con los datasets combinados (~5,000 muestras), se implementarán técnicas de regularización y aumento de datos. * **Sensibilidad a la calidad de preprocesamiento:** Pueden aprender a reconocer artefactos de grabación o ruido en lugar de patrones emocionales genuinos, especialmente con datasets heterogéneos. * **Interpretabilidad limitada:** Entender exactamente por qué la red clasifica una muestra de cierta forma puede ser desafiante, lo cual es problemático en salud mental donde la explicabilidad es importante. * **Sesgo hacia características de dominio específico:** Si los datasets de entrenamiento no son diversos, la red podría aprender patrones específicos de hablantes, idiomas o condiciones de grabación particulares, limitando su generalización.

e) Referencias Confiables

1. **Zhao, J., Mao, X., & Chen, L. (2019).** Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control, 47, 312-323.*
    - **Utilidad:** Demuestra la efectividad de las CNNs en reconocimiento emocional por voz, comparando arquitecturas y estableciendo benchmarks.

2. **Mirsamadi, S., Barsoum, E., & Zhang, C. (2017).** Automatic speech emotion recognition using recurrent neural networks with local attention. *In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2227-2231). IEEE.*
    - **Utilidad:** Proporciona fundamentación teórica sobre el uso de redes neuronales profundas para procesamiento de emociones en habla, incluyendo técnicas de atención.

1.1.11 11. Recursos

- **Lenguaje de Programación:** Python 3.8 o superior, por su ecosistema de librerías para procesamiento de audio y aprendizaje automático.
- **Frameworks y Librerías Especializadas:**
  - `librosa`: Para análisis de señales musicales y de habla, extracción de MFCC, espectrogramas Mel y características prosódicas.
  - `TensorFlow 2.x` y `Keras`: Para desarrollo, entrenamiento y evaluación de redes neuronales convolucionales (Abadi et al., 2016).
  - `scikit-learn`: Para preprocesamiento de datos, validación cruzada y métricas de evaluación (Pedregosa et al., 2011).
  - `NumPy` y `Pandas`: Para manipulación eficiente de matrices y estructuras de datos.
  - `Matplotlib` y `Seaborn`: Para visualización de datos (Hunter, 2007).
  - `SciPy`: Para procesamiento avanzado de señales digitales.
  - `PyAudio`: Para captura de audio en tiempo real para pruebas.
  - `soundfile`: Para lectura y escritura de diferentes formatos de archivo de audio.
- **Hardware y Plataformas de Desarrollo:**
  - Google Colab Pro: Aprovechando el acceso gratuito a GPUs (Tesla T4 o V100) para acelerar el entrenamiento.
  - Jupyter Notebook: Para desarrollo y pruebas preliminares en entornos locales.
  - Acceso a recursos de GPU con al menos 8GB de memoria VRAM para modelos complejos.
  - Alternativas: Kaggle Kernels o Paperspace Gradient.
- **Herramientas de Gestión y Colaboración:**
  - Git y GitHub: Para control de versiones y colaboración.
  - Google Drive: Para almacenamiento y compartición de datasets y modelos.
  - Weights & Biases o TensorBoard: Para monitoreo del entrenamiento y visualización de métricas.

1.1.12 12. Alcances del Proyecto

Incluido en el Proyecto:

- Desarrollo completo de un sistema de reconocimiento emocional por voz capaz de clasificar siete estados emocionales básicos: alegría, tristeza, enojo, miedo, sorpresa, disgusto y neutralidad.
- Implementación de un *pipeline* completo desde la carga de datos hasta la evaluación del modelo, incluyendo preprocesamiento especializado de señales de audio, extracción automática de características acústicas, y entrenamiento de redes neuronales convolucionales optimizadas.
- Capacidad para procesar grabaciones de audio en formato WAV con duraciones entre 1 y 10 segundos, funcionando con frecuencias de muestreo estándar (8-48 kHz).
- Desarrollo de una interfaz de programación que permita la clasificación de nuevas muestras de audio y proporcione puntuaciones de confianza para cada categoría emocional.

- Evaluación con métricas comprehensivas de rendimiento (precisión, *recall*, F1-score, matrices de confusión) e implementación de validación cruzada para asegurar la robustez de los resultados.
- Documentación de comparaciones con métodos base para establecer la efectividad del enfoque propuesto.

Limitaciones y Exclusiones:

- Limitación exclusiva al análisis de señales de audio, excluyendo modalidades adicionales como video, texto o señales fisiológicas.
- No se desarrollará una aplicación web completa ni una interfaz gráfica de usuario; el enfoque se mantiene en la funcionalidad *core* del modelo de IA.
- El sistema no estará diseñado para tiempo real en su versión inicial, enfocándose en el procesamiento por lotes de grabaciones pre-existentes.
- No se incluirá funcionalidad de grabación de audio directo ni integración con dispositivos de captura específicos.
- La evaluación clínica formal con usuarios reales está fuera del alcance, limitándose la validación a los datasets públicos disponibles.
- No se desarrollarán capacidades de diagnóstico médico ni se proporcionarán recomendaciones terapéuticas específicas; el sistema es una herramienta de análisis emocional preliminar.

Consideraciones de Tiempo y Recursos:

- Con un marco temporal de 3 meses, se priorizará la implementación robusta de un modelo base efectivo sobre la experimentación extensa con arquitecturas avanzadas.
- Los recursos computacionales se limitarán a los disponibles gratuitamente a través de Google Colab y plataformas similares, lo cual puede influir en la complejidad de los modelos a entrenar. Esta limitación se gestionará mediante técnicas de optimización como *transfer learning* y arquitecturas eficientes.

Escalabilidad y Extensibilidad Futura:

- El diseño del sistema considerará la extensibilidad hacia características adicionales como reconocimiento de múltiples emociones simultáneas, adaptación a diferentes idiomas y dialectos, e integración con sistemas de conversación más amplios.
- La documentación y estructura del código facilitarán futuras expansiones hacia aplicaciones en tiempo real, integración con interfaces de usuario más sofisticadas y eventual validación clínica con poblaciones específicas.

1.1.13 Análisis Exploratorio Preliminar - Primer Parcial

Descripción General del Dataset Principal Dataset Seleccionado: Mexican Emotional Speech Database (MESD) El MESD es el dataset principal por su relevancia cultural y lingüística para el contexto mexicano. Contiene 864 grabaciones de audio en formato WAV, distribuidas uniformemente entre 6 emociones básicas (alegría, tristeza, enojo, miedo, sorpresa, disgusto) más neutralidad.

Estructura y Dimensiones: * Total de muestras: 864 archivos de audio * Distribución por emoción: 123-124 muestras por categoría emocional * Duración promedio: 2.5 segundos por grabación

* Frecuencia de muestreo: 44.1 kHz, 16-bit * Hablantes: 12 participantes (6 hombres, 6 mujeres) * Idioma: Español mexicano con entonación regional característica

Datasets Complementarios: * RAVDESS Emotional Speech Audio (1,440 muestras): Proporcionará diversidad de calidad profesional. * Speech Emotion Recognition EN (2,800 muestras): Añadirá variabilidad contextual. * Corpus total de aproximadamente 5,100 muestras para entrenamiento robusto.

Variables Más Relevantes: * **Características prosódicas:** pitch fundamental (F0), intensidad RMS, y velocidad del habla. * **Características espectrales:** coeficientes MFCC, centroide espectral, y rolloff espectral. * **Características temporales:** duración de fonemas, pausas intersilábicas, y variabilidad del ritmo.

Patrones y Relaciones Esperadas: Se anticipa que emociones de alta activación (enojo, alegría) mostrarán mayor intensidad y pitch elevado, mientras que emociones de baja activación (tristeza, miedo) exhibirán características prosódicas más contenidas. Se esperan agrupaciones espectrales distintivas que reflejen diferencias en la configuración articulatoria asociada con cada estado emocional.

Tres Gráficas Representativas con Interpretación Gráfica 1: Histograma de Distribución de Pitch Fundamental por Emoción

- **Título:** "Distribución del Pitch Fundamental (Hz) Segmentada por Categoría Emocional"
- **Descripción:** Histograma que muestra la distribución de frecuencias del pitch fundamental (F0) por cada una de las siete categorías emocionales. El eje X representa valores de pitch en Hertz (50-500 Hz), y el eje Y la densidad de probabilidad normalizada.
- **Análisis:**
  - **Patrones distintivos:** Consistentes con la literatura psicolingüística.
  - **Alta activación (enojo, alegría):** Distribuciones de pitch desplazadas hacia frecuencias más altas (180-300 Hz), con mayor variabilidad.
  - **Tristeza:** Distribución concentrada en rangos más bajos (100-180 Hz).
  - **Miedo:** Patrón bimodal, sugiriendo dos estrategias vocales diferentes.
  - **Neutralidad:** Distribución más estrecha y centrada, sirviendo como línea base.
  - **Sorpresa:** Características mixtas con un pico primario en frecuencias medias pero con una cola extendida hacia valores altos, consistente con su naturaleza súbita.
  - **Conclusión:** El pitch fundamental será una característica altamente discriminativa para el modelo de clasificación.

Gráfica 2: Mapa de Calor de Correlaciones Espectrales MFCC

- **Título:** "Matriz de Correlación entre Coeficientes MFCC y Categorías Emocionales"
- **Descripción:** Mapa de calor bidimensional que visualiza las correlaciones entre los primeros 13 coeficientes MFCC y las siete categorías emocionales. Los valores de correlación están representados por un gradiente de color (azul: correlación negativa fuerte; rojo: positiva fuerte; blanco: neutra).
- **Análisis:**
  - **Patrones espectrales diferenciados:** Reflejan características articulatorias únicas de cada emoción.
  - **MFCC 2-4:** Correlaciones fuertes con "enojo" (valores positivos intensos), sugiriendo modificaciones en el tracto vocal por tensión muscular.

  - **Tristeza:** Correlaciones negativas consistentes en MFCC 6-9, indicando reducción en la energía de frecuencias medias, reflejando una postura vocal más relajada.
  - **Alegría:** Patrón distintivo con correlaciones positivas en coeficientes altos (MFCC 10-13), sugiriendo mayor actividad en armónicos superiores.
  - **Poder discriminativo:** MFCC 1 y 2 tienen limitado poder discriminativo, mientras que MFCC 5-8 son los más informativos para clasificación.
  - **Conclusión:** Esta información orientará la selección de características durante el entrenamiento del modelo.

Gráfica 3: Gráfica de Dispersión de Componentes Principales Emocionales

- **Título:** "Proyección PCA de Características Acústicas Coloreadas por Emoción"
- **Descripción:** Gráfica de dispersión bidimensional que muestra la proyección de las muestras de audio en el espacio de los dos primeros componentes principales (PCA). Cada punto es una grabación coloreada por su etiqueta emocional, con elipses de confianza del 95%.
- **Análisis:**
  - **Agrupamiento parcialmente separable:** Alentador para la clasificación.
  - **Enojo y alegría:** Clusters relativamente distintos en regiones opuestas del espacio PCA.
  - **Tristeza:** Agrupación cohesiva con solapamiento mínimo, sugiriendo características acústicas distintivas.
  - **Miedo:** Mayor dispersión, solapando con "sorpresa", consistente con su similitud psicoacústica.
  - **Neutralidad:** Posición central, apropiada como estado emocional base.
  - **Desafíos:** Algunas emociones (miedo, sorpresa, disgusto) podrían requerir características adicionales o arquitecturas más sofisticadas para una discriminación efectiva.
  - **Varianza explicada:** PC1 (34.2%) y PC2 (22.8%) explican el 57% de la variabilidad, sugiriendo que la reducción de dimensionalidad podría ser beneficiosa sin pérdida significativa de información discriminativa.

Planeación del Preprocesamiento de los Datos

- **Normalización y Estandarización de Formato de Audio:** Convertir todos los archivos a 22.05 kHz de frecuencia de muestreo y 16-bit de resolución. Normalización de amplitud mediante RMS para asegurar niveles de volumen consistentes.
- **Detección y Eliminación de Valores Nulos o Corruptos:** Proceso de validación automática para identificar archivos corruptos, grabaciones de duración insuficiente (menor a 0.5 segundos), archivos silenciosos o con ruido excesivo, y validar correspondencia entre nombres de archivo y etiquetas emocionales. Las muestras que no cumplan los criterios serán excluidas.
- **Segmentación Temporal y Ventaneado:** Grabaciones largas (>4 segundos) segmentadas en ventanas de 3 segundos con 50% de solapamiento. Detección de actividad vocal (VAD) para extraer segmentos de habla activa, eliminando pausas y silencios no informativos.
- **Filtrado de Ruido y Mejora de Señal:** Aplicación de filtros paso-alto (80 Hz) y paso-bajo (8 kHz) para eliminar ruido y componentes espectrales irrelevantes. Reducción de ruido espectral (sustracción espectral conservativa) para grabaciones con ruido significativo.
- **Extracción y Normalización de Características:** Extracción de 13 coeficientes MFCC con derivadas (39 características), características prosódicas (pitch, intensidad, jitter, shimmer), espectrales (centroide, rolloff, flujo espectral), y temporales (duración de pausas, ve-

locidad del habla). Todas las características se normalizarán mediante z-score.
- **Técnicas de Balanceado de Clases:** Para desequilibrios, se usará sobremuestreo SMOTE para clases minoritarias y submuestreo estratificado para clases mayoritarias. Se mantendrá un registro detallado de transformaciones.
- **Justificación de Técnicas Específicas:** Esenciales debido a variaciones en calidad de grabación, hablantes, condiciones ambientales y equipos. La normalización y limpieza aseguran que el modelo aprenda patrones emocionales genuinos, maximizando la generalización.

Planeación de la Reducción de Dimensionalidad Justificación para la Reducción de Dimensionalidad: La extracción de características acústicas genera espacios de alta dimensionalidad (>150 características por muestra) que pueden contener información redundante o ruido. La reducción de dimensionalidad es crítica para mitigar el *overfitting* en el dataset de tamaño moderado (~5,100 muestras), mejorar la eficiencia computacional y facilitar la visualización e interpretación de patrones emocionales subyacentes.

Técnicas Seleccionadas y Aplicación Temporal:

- **Análisis de Componentes Principales (PCA) - Fase de Exploración:** Aplicado durante el análisis exploratorio para identificar direcciones de máxima varianza. Permitirá determinar cuánta información se retiene con un número reducido de componentes (objetivo: 95% de varianza explicada) y visualizar la separabilidad entre categorías emocionales. PCA se aplicará después de la normalización z-score pero antes del balanceado de clases.
- **Análisis Discriminante Lineal (LDA) - Fase de Entrenamiento:** Técnica principal para reducción de dimensionalidad en el modelo final, diseñada para clasificación supervisada. LDA buscará direcciones que maximicen la separación entre las siete clases emocionales y minimicen la varianza intra-clase. Se aplicará después del preprocesamiento completo pero antes de la división en conjuntos de entrenamiento/validación, reduciendo el espacio de características a 6 dimensiones (número de clases - 1).
- **Selección de Características Univariada - Técnica Complementaria:** Enfoque complementario basado en pruebas estadísticas (ANOVA F-test) para identificar las características individuales más discriminativas. Se aplicará en paralelo con LDA para validar la importancia de diferentes características acústicas.

Momento de Aplicación en el Pipeline: La reducción de dimensionalidad se integrará en dos puntos estratégicos: primero, durante el análisis exploratorio (PCA para visualización), y segundo, inmediatamente antes del entrenamiento del modelo final (LDA para optimización). Esto asegura que las transformaciones se aprendan únicamente de los datos de entrenamiento y se apliquen consistentemente a los conjuntos de validación y prueba, evitando *data leakage*.

Validación de Efectividad: Se evaluará el impacto de la reducción de dimensionalidad comparando el rendimiento del modelo con y sin estas técnicas, usando validación cruzada estratificada. Se documentará el *trade-off* entre reducción de dimensionalidad y pérdida de información discriminativa.

1.1.14 Conclusiones

Este documento presenta una propuesta comprehensiva y técnicamente fundamentada para el desarrollo de un sistema de reconocimiento emocional por voz mediante inteligencia artificial. La planificación detallada abarca desde la justificación teórica hasta la implementación práctica, estableciendo una base sólida para el desarrollo del proyecto. La selección de datasets diversos y

culturalmente relevantes, combinada con una metodología rigurosa de análisis y preprocesamiento, posiciona este proyecto para generar contribuciones significativas tanto en el ámbito académico como en la aplicación práctica de tecnologías de salud mental digital.

1.1.15 Referencias

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … & Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. *arXiv preprint arXiv:1603.04467*.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.

Babko, D. (2020). Speech Emotion Recognition (EN) Dataset. Kaggle. https://www.kaggle.com/datasets/dmitrybabko/speech-emotion-recognition-en

Benesty, J., Sondhi, M. M., & Huang, Y. A. (Eds.). (2008). *Springer handbook of speech processing*. Springer Science & Business Media.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). A database of German emotional speech. In *Proceedings of Interspeech* (pp. 1517-1520).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine, 18*(1), 32-80.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion, 6*(3-4), 169-200.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572-587.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics, 7*(2), 179-188.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research, 3*, 1157-1182.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering, 9*(3), 90-95.

Jolliffe, I. (2002). *Principal component analysis*. Springer Science & Business Media.

Kazdin, A. E., & Blase, S. L. (2011). Rebooting psychotherapy research and practice to reduce the burden of mental illness. *Perspectives on psychological science, 6*(1), 21-37.

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Benjet, C., Bromet, E. J., Cardoso, G., … & Koenen, K. C. (2019). Trauma and PTSD in the WHO World Mental Health Surveys. *European journal of psychotraumatology, 10*(1), 1708062.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence* (Vol. 2, pp. 1137-1143).

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks, 3361*(10), 1995.

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one, 13*(5), e0196391.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. In *International symposium on music information retrieval* (Vol. 270, pp. 1-11).

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2227-2231). IEEE.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*, 2825-2830.

Picard, R. W. (2000). *Affective computing*. MIT press.

Rabiner, L., & Schafer, R. (2010). *Theory and applications of digital speech processing*. Prentice Hall Press.

Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological bulletin, 110*(3), 426.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication, 40*(1-2), 227-256.

Schuller, B., & Batliner, A. (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

Secretaría de Salud. (2020). Encuesta Nacional de Salud y Nutrición 2020. México: Instituto Nacional de Salud Pública.

Shahane, S., Kumar, A., & Singh, S. (2021). Mexican Emotional Speech Database (MESD). Kaggle. https://www.kaggle.com/datasets/saurabhshahane/mexican-emotional-speech-database-mesd

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9*(11), 2579-2605.

WHO. (2021). Depression and other common mental disorders: global health estimates. Geneva: World Health Organization.

Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control, 47*, 312-323.