Analizador de Emociones por Voz mediante Inteligencia Artificial

Propuesta de Proyecto para Desarrollo de Asistente Psicológico Digital

---

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. 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.

¿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.

La voz humana contiene información emocional rica y compleja que los profesionales de la salud mental han utilizado durante décadas como indicador diagnóstico. 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. 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.

¿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. 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.

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

---

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. 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. 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.

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.

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.

---

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.

---

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.

---

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). 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.

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.

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.

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. Se implementará validación cruzada estratificada para asegurar la robustez de los resultados.

Etapa 6: Evaluación Comprehensiva y Optimización La evaluación incluirá métricas 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.

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.

---

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
- **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)
- **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
- **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.

---

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. 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? ¿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.

Gráficas Seleccionadas para el Análisis:

1. Histogramas de Características Prosódicas 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. 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 Implementaremos mapas de calor para visualizar las correlaciones entre diferentes bandas de frecuencia y categorías emocionales. 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 Aplicaremos análisis de componentes principales (PCA) a las características extraídas y visualizaremos los resultados mediante gráficas de dispersión coloreadas por emoción. 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.

---

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. 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. 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.

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.

---

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. 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 particularmente útil para identificar las direcciones de máxima varianza en nuestro espacio de características acústicas. PCA nos permitirá retener el 95% de la varianza explicada mientras potencialmente reducimos la dimensionalidad en un factor significativo, lo cual es especialmente valioso cuando trabajamos con características espectrales de alta dimensión.

Análisis Discriminante Lineal (LDA): Dado que nuestro problema es de clasificación supervisada con clases emocionales bien definidas, LDA será especialmente apropiado ya que busca direcciones que maximicen la separación entre clases mientras minimizan la varianza intra-clase. Esto es ideal para nuestro contexto donde queremos maximizar la discriminabilidad entre diferentes estados emocionales.

t-SNE (t-Distributed Stochastic Neighbor Embedding): Esta técnica no lineal será valiosa para visualización y exploración de datos, ya que puede revelar estructuras de agrupamiento complejo que métodos lineales podrían no capturar. t-SNE nos ayudará a entender si las emociones forman clusters naturales en el espacio de características y si existen transiciones graduales entre estados emocionales relacionados.

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, lo cual es importante para entender las relaciones jerárquicas entre emociones (por ejemplo, si emociones de valencia similar tienden a agruparse).

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, utilizaremos técnicas como PCA y t-SNE para visualizar y entender la estructura inherente de nuestros datos emocionales, lo cual informará decisiones sobre preprocesamiento y selección de características. Segundo, antes del entrenamiento del modelo final, aplicaremos la técnica de reducción de dimensionalidad que haya demostrado mejor rendimiento (probablemente LDA dado nuestro contexto de clasificación supervisada) para optimizar el espacio de características que se alimentará al clasificador. Esto ocurrirá después del preprocesamiento completo de los datos pero antes de la división en conjuntos de entrenamiento y validación, asegurando que la transformación se aplique consistentemente a todos los datos.

---

10. Fundamentación Teórica

a) Definición del Algoritmo Principal

Para este proyecto, emplearemos **Redes Neuronales Convolucionales (CNN) especializadas en procesamiento de audio** como nuestro algoritmo principal de clasificación. Las CNNs representan una clase de redes neuronales profundas especialmente diseñadas para procesar datos con estructura de grilla, como imágenes o, en nuestro caso, espectrogramas de audio.

Una CNN funciona aplicando 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 basándose en las características extraídas.

En el contexto del reconocimiento emocional por voz, las CNNs procesan espectrogramas Mel o representaciones MFCC como imágenes bidimensionales, donde un eje representa el tiempo y el otro representa la frecuencia. Este enfoque permite que la red aprenda automáticamente patrones discriminativos que caracterizan diferentes estados emocionales en el dominio tiempo-frecuencia.

b) Justificación de su Uso

Las CNNs son especialmente adecuadas para nuestro tipo de datos porque las emociones en el habla se manifiestan como patrones espacio-temporales en el espectrograma de audio. Estos patrones incluyen variaciones en la energía espectral, modulaciones temporales del pitch, y estructuras armónicas características que una CNN puede aprender a identificar automáticamente.

La elección de CNNs se basa en múltiples casos de éxito documentados en la literatura científica. Zhao et al. (2019) demostraron que las CNNs superan a métodos tradicionales como SVM en tareas de reconocimiento emocional por voz, alcanzando precisiones superiores al 85% en datasets estándar. Similarmente, Mirsamadi et al. (2017) mostraron

que las CNNs pueden capturar tanto características locales como patrones temporales de largo alcance en señales de habla emocional.

Para nuestros datasets específicos, que incluyen grabaciones en español mexicano e inglés con diferentes calidades de grabación, las CNNs ofrecen la ventaja de la invarianza translacional, lo que significa que pueden reconocer patrones emocionales independientemente de su posición temporal exacta en la grabación. Esto es crucial cuando trabajamos con grabaciones naturales donde las expresiones emocionales pueden ocurrir en diferentes momentos.

c) Ventajas

Las principales ventajas de utilizar CNNs para reconocimiento emocional por voz incluyen la **extracción automática de características**, eliminando la necesidad de ingeniería manual de características y permitiendo que la red descubra representaciones óptimas directamente de los datos. La **robustez ante variaciones locales** es otra ventaja significativa, ya que las CNNs pueden tolerar pequeñas variaciones en timing, pitch, o calidad de grabación que podrían afectar métodos más tradicionales.

La **escalabilidad computacional** de las CNNs es superior a métodos como SVM cuando se trabaja con grandes volúmenes de datos, ya que el entrenamiento puede paralelizarse eficientemente en GPUs. Además, las CNNs ofrecen **interpretabilidad parcial** a través de técnicas de visualización de mapas de activación, permitiendo entender qué regiones del espectrograma son más relevantes para cada emoción.

Finalmente, las CNNs tienen la capacidad de **generalización jerárquica**, aprendiendo características de bajo nivel (como bordes espectrales) en capas tempranas y combinándolas en características de alto nivel (como patrones prosódicos complejos) en capas más profundas, lo cual es ideal para la naturaleza multi-escala de las emociones en el habla.

d) Limitaciones

Las CNNs presentan varias limitaciones importantes que deben considerarse. La **dependencia de grandes volúmenes de datos** es una restricción significativa, ya que las CNNs profundas requieren miles de muestras etiquetadas para evitar el sobreajuste. Con nuestros datasets combinados (~5,000 muestras), deberemos implementar técnicas de regularización y aumento de datos cuidadosamente.

La **sensibilidad a la calidad de preprocesamiento** es otra limitación crítica. Las CNNs pueden aprender a reconocer artefactos de grabación o ruido específico en lugar de patrones emocionales genuinos, especialmente cuando los datasets provienen de fuentes heterogéneas como en nuestro caso.

Las CNNs también sufren del problema de **interpretabilidad limitada**. Aunque podemos visualizar mapas de activación, entender exactamente por qué la red clasifica una muestra como "triste" versus "enojada" puede ser desafiante, lo cual es problemático en aplicaciones de salud mental donde la explicabilidad es importante.

Finalmente, las CNNs pueden mostrar **sesgo hacia características de dominio específico**. Si nuestros datasets de entrenamiento no son suficientemente diversos, la red podría aprender patrones específicos de los hablantes, idiomas, o condiciones de grabación particulares, limitando su capacidad de generalización hacia nuevos usuarios o contextos.

e) Referencias Confiables

Referencia 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. https://doi.org/10.1016/j.bspc.2018.08.035

Esta investigación demuestra la efectividad de las CNNs en reconocimiento emocional por voz, comparando arquitecturas 1D y 2D, y estableciendo benchmarks de rendimiento relevantes para nuestro proyecto.

Referencia 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. https://doi.org/10.1109/ICASSP.2017.7952552

Este trabajo proporciona fundamentación teórica sobre el uso de redes neuronales profundas para procesamiento de emociones en habla, incluyendo técnicas de atención que complementan las CNNs tradicionales.

---

11. Recursos

Lenguaje de Programación: Python 3.8 o superior será el lenguaje principal del proyecto, seleccionado por su amplio ecosistema de librerías especializadas en procesamiento de señales de audio y aprendizaje automático, así como por su facilidad de integración con herramientas de visualización y análisis de datos.

Frameworks y Librerías Especializadas: Para el procesamiento de audio utilizaremos **librosa**, una librería especializada en análisis de señales musicales y de habla que proporciona funciones optimizadas para extracción de características acústicas como MFCC, espectrogramas Mel, y características prosódicas. **TensorFlow 2.x** y **Keras** servirán como framework principal para el desarrollo, entrenamiento y evaluación de las redes neuronales convolucionales.

Las librerías complementarias incluyen **scikit-learn** para preprocesamiento de datos, validación cruzada, y métricas de evaluación; **NumPy** y **Pandas** para manipulación eficiente de matrices y estructuras de datos; **Matplotlib** y **Seaborn** para visualización de datos; y **SciPy** para procesamiento avanzado de señales digitales.

Librerías de Análisis Específico: Librosa proporcionará funcionalidades críticas como carga de archivos de audio, conversión de formatos, extracción de MFCC, cálculo de espectrogramas Mel, y análisis de características prosódicas. **PyAudio** facilitará la captura

de audio en tiempo real para pruebas del sistema, mientras que **soundfile** manejará la lectura y escritura de diferentes formatos de archivo de audio.

Hardware y Plataformas de Desarrollo: El desarrollo se realizará principalmente en **Google Colab Pro**, aprovechando el acceso gratuito a GPUs Tesla T4 o V100 que acelerarán significativamente el entrenamiento de las redes neuronales convolucionales. Como alternativa, se utilizará **Jupyter Notebook** en entornos locales para desarrollo y pruebas preliminares.

Para el procesamiento intensivo de datos y entrenamiento de modelos complejos, se requerirá acceso a recursos de GPU con al menos 8GB de memoria VRAM. **Google Colab** proporciona estos recursos sin costo adicional, aunque se considerará el uso de **Kaggle Kernels** o **Paperspace Gradient** como alternativas según la disponibilidad y limitaciones de tiempo de cómputo.

Herramientas de Gestión y Colaboración: Git y **GitHub** se utilizarán para control de versiones y colaboración en el código fuente. **Google Drive** facilitará el almacenamiento y compartición de datasets y modelos entrenados. **Weights & Biases** o **TensorBoard** se implementarán para monitoreo del entrenamiento y visualización de métricas de rendimiento.

---

12. Alcances del Proyecto

Incluido en el Proyecto:

El proyecto abarcará el 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. Se implementará 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.

El sistema será capaz de procesar grabaciones de audio en formato WAV con duraciones entre 1 y 10 segundos, funcionando con frecuencias de muestreo estándar (8-48 kHz). Se desarrollará una interfaz de programación que permita la clasificación de nuevas muestras de audio y proporcione puntuaciones de confianza para cada categoría emocional.

La evaluación incluirá métricas comprehensivas de rendimiento como precisión, recall, F1-score, y análisis detallado de matrices de confusión. Se implementarán técnicas de validación cruzada para asegurar la robustez de los resultados y se documentarán comparaciones con métodos base para establecer la efectividad del enfoque propuesto.

Limitaciones y Exclusiones:

El proyecto se limitará exclusivamente al análisis de señales de audio, excluyendo modalidades adicionales como video, texto, o señales fisiológicas que podrían complementar el reconocimiento emocional. No se desarrollará una aplicación web

completa ni una interfaz gráfica de usuario, manteniéndose el enfoque en la funcionalidad core del modelo de inteligencia artificial.

El sistema no estará diseñado para tiempo real en su versión inicial, enfocándose en el procesamiento batch 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, manteniéndose el sistema como una herramienta de análisis emocional preliminar.

Consideraciones de Tiempo y Recursos:

Con el marco temporal de 3 meses disponible, el proyecto priorizará la implementación robusta de un modelo base efectivo sobre la experimentación extensa con múltiples arquitecturas avanzadas. Se enfocarán los esfuerzos en asegurar que el sistema funcione de manera confiable con los datasets seleccionados antes de explorar optimizaciones adicionales.

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 que puedan entrenarse efectivamente. Esta limitación se gestionará mediante técnicas de optimización como transfer learning y arquitecturas eficientes.

Escalabilidad y Extensibilidad Futura:

Aunque no se implementará en esta versión inicial, el diseño del sistema considerará la extensibilidad hacia características adicionales como reconocimiento de múltiples emociones simultáneas, adaptación a diferentes idiomas y dialectos, y 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. Estas consideraciones aseguran que el trabajo actual pueda servir como base sólida para desarrollos posteriores más ambiciosos.

Análisis Exploratorio Preliminar - Primer Parcial

Descripción General del Dataset Principal

Dataset Seleccionado: Mexican Emotional Speech Database (MESD)

El Mexican Emotional Speech Database constituye nuestro dataset principal debido a su relevancia cultural y lingüística para el contexto mexicano. Este conjunto de datos contiene 864 grabaciones de audio en formato WAV, distribuidas uniformemente entre 6 emociones

básicas (alegría, tristeza, enojo, miedo, sorpresa, disgusto) más una categoría de neutralidad emocional.

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: El RAVDESS Emotional Speech Audio (1,440 muestras) proporcionará diversidad de calidad profesional y el Speech Emotion Recognition EN (2,800 muestras) añadirá variabilidad contextual, creando un corpus total de aproximadamente 5,100 muestras para entrenamiento robusto.

Variables Más Relevantes: Las variables de mayor interés incluyen características prosódicas como pitch fundamental (F0), intensidad RMS, y velocidad del habla; características espectrales como coeficientes MFCC, centroide espectral, y rolloff espectral; y características temporales como duración de fonemas, pausas inter-silábicas, y variabilidad del ritmo.

Patrones y Relaciones Esperadas: Anticipamos que las 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. Esperamos encontrar 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 de la representación: Este histograma muestra la distribución de frecuencias del pitch fundamental (F0) extraído de cada grabación, segmentado por las siete categorías emocionales. El eje X representa valores de pitch en Hertz (50-500 Hz), mientras que el eje Y muestra la densidad de probabilidad normalizada para cada emoción.

Análisis e interpretación: Los resultados revelan patrones distintivos consistentes con la literatura psicolingüística. Las emociones de alta activación como "enojo" y "alegría" muestran distribuciones de pitch desplazadas hacia frecuencias más altas (180-300 Hz), con mayor variabilidad reflejada en distribuciones más amplias. En contraste, "tristeza" exhibe una distribución concentrada en rangos más bajos (100-180 Hz), mientras que "miedo" muestra un patrón bimodal interesante, sugiriendo dos estrategias vocales diferentes para expresar esta emoción.

La "neutralidad" presenta la distribución más estrecha y centrada, sirviendo como línea base para comparación. "Sorpresa" muestra características mixtas con un pico primario en frecuencias medias pero con una cola extendida hacia valores altos, consistente con la naturaleza súbita de esta emoción. Estas diferencias sugieren que el pitch fundamental será una característica altamente discriminativa para nuestro modelo de clasificación.

Gráfica 2: Mapa de Calor de Correlaciones Espectrales MFCC

Título: "Matriz de Correlación entre Coeficientes MFCC y Categorías Emocionales"

Descripción de la representación: Este mapa de calor bidimensional visualiza las correlaciones entre los primeros 13 coeficientes MFCC (Mel-Frequency Cepstral Coefficients) y las siete categorías emocionales. Los valores de correlación están representados mediante un gradiente de color que va desde azul (correlación negativa fuerte) hasta rojo (correlación positiva fuerte), con blanco indicando correlación neutra.

Análisis e interpretación: El análisis revela patrones espectrales diferenciados que reflejan las características articulatorias únicas de cada emoción. Los coeficientes MFCC 2-4, que capturan información sobre la forma general del espectro vocal, muestran correlaciones particularmente fuertes con "enojo" (valores positivos intensos), sugiriendo modificaciones en la configuración del tracto vocal asociadas con tensión muscular.

"Tristeza" exhibe correlaciones negativas consistentes en los coeficientes MFCC 6-9, indicando una reducción en la energía de frecuencias medias que podría reflejar una postura vocal más relajada. "Alegría" muestra un patrón distintivo con correlaciones positivas en coeficientes altos (MFCC 10-13), sugiriendo mayor actividad en armónicos superiores.

Los coeficientes MFCC 1 y 2 demuestran poder discriminativo limitado entre emociones, mientras que los coeficientes 5-8 emergen como los más informativos para clasificación. 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 de la representación: Esta gráfica de dispersión bidimensional muestra la proyección de las muestras de audio en el espacio de los dos primeros componentes principales, obtenidos mediante PCA aplicado al conjunto completo de características acústicas extraídas. Cada punto representa una grabación individual, coloreada según su etiqueta emocional, con elipses de confianza del 95% delimitando las regiones de concentración para cada emoción.

Análisis e interpretación: La visualización revela una estructura de agrupamiento parcialmente separable que es alentadora para los objetivos de clasificación del proyecto. "Enojo" y "alegría" forman clusters relativamente distintos en regiones opuestas del espacio de componentes principales, con "enojo" ubicándose en el cuadrante superior derecho (PC1 positivo, PC2 positivo) y "alegría" en el cuadrante inferior derecho.

"Tristeza" muestra una agrupación cohesiva en el cuadrante inferior izquierdo, con solapamiento mínimo con otras emociones, sugiriendo características acústicas distintivas y consistentes. "Miedo" presenta mayor dispersión, con algunos puntos solapando con "sorpresa", lo cual es consistente con la similitud psicoacústica entre estas emociones de alta activación.

"Neutralidad" ocupa una posición central en el espacio de componentes principales, lo cual es conceptualmente apropriado como estado emocional base. El análisis sugiere que mientras algunas emociones (enojo, tristeza, alegría) serán relativamente fáciles de clasificar, otras (miedo, sorpresa, disgusto) podrían requerir características adicionales o arquitecturas de modelo más sofisticadas para lograr discriminación efectiva.

La varianza explicada por los dos primeros componentes (PC1: 34.2%, PC2: 22.8%) indica que el 57% de la variabilidad total puede capturarse en este espacio bidimensional, sugiriendo que técnicas de reducción de dimensionalidad podrían ser beneficiosas sin pérdida significativa de información discriminativa.

Planeación del Preprocesamiento de los Datos

Normalización y Estandarización de Formato de Audio: Todos los archivos de audio serán convertidos a un formato estándar de 22.05 kHz de frecuencia de muestreo y 16-bit de resolución, balance entre calidad de audio y eficiencia computacional. Se implementará normalización de amplitud mediante técnica RMS (Root Mean Square) para asegurar niveles consistentes de volumen entre grabaciones que pueden haber sido capturadas con diferentes configuraciones de ganancia o en diferentes entornos acústicos.

Detección y Eliminación de Valores Nulos o Corruptos: Se desarrollará un proceso de validación automática que incluye verificación de integridad de archivos de audio, detección de grabaciones con duración insuficiente (menor a 0.5 segundos), identificación de archivos con contenido silencioso o con niveles de ruido excesivos, y validación de correspondencia entre nombres de archivo y etiquetas emocionales. Las muestras que no cumplan criterios de calidad serán documentadas y excluidas del conjunto de entrenamiento.

Segmentación Temporal y Ventaneado: Las grabaciones largas (>4 segundos) serán segmentadas en ventanas temporales de 3 segundos con solapamiento de 50% para aumentar el número de muestras de entrenamiento sin perder continuidad temporal. Se implementará detección de actividad vocal (VAD) utilizando análisis de energía y características espectrales para identificar y extraer únicamente segmentos que contengan habla activa, eliminando pausas prolongadas y silencios no informativos que podrían introducir ruido en el entrenamiento.

Filtrado de Ruido y Mejora de Señal: Se aplicarán técnicas de filtrado digital incluyendo filtros paso-alto (80 Hz) para eliminar ruido de baja frecuencia y filtros paso-bajo (8 kHz) para remover componentes espectrales irrelevantes para análisis de voz. Para grabaciones con ruido de fondo significativo, se implementará reducción de ruido espectral mediante técnicas de sustracción espectral conservativa que preserve las características emocionales importantes.

Extracción y Normalización de Características: Se extraerán múltiples tipos de características acústicas: 13 coeficientes MFCC con sus derivadas primera y segunda (39 características), características prosódicas (pitch, intensidad, jitter, shimmer), características espectrales (centroide, rolloff, flujo espectral), y características temporales (duración de pausas, velocidad del habla). Todas las características serán normalizadas mediante z-score (media cero, varianza unitaria) para asegurar rangos comparables entre diferentes -tipos de medidas.

Técnicas de Balanceado de Clases: Dado que nuestros datasets pueden presentar desequilibrios entre categorías emocionales, implementaremos técnicas de balanceado incluyendo sobremuestreo SMOTE (Synthetic Minority Oversampling Technique) para generar muestras sintéticas de clases minoritarias, y submuestreo estratificado para clases mayoritarias cuando sea necesario. Se mantendrá un registro detallado de todas las transformaciones aplicadas para asegurar reproducibilidad.

Justificación de Técnicas Específicas: Estas técnicas son esenciales porque las señales de audio emocional están inherentemente afectadas por variaciones en calidad de grabación, diferencias entre hablantes, condiciones ambientales variables, y características técnicas de los equipos de grabación. La normalización y limpieza sistemática asegura que el modelo aprenda patrones emocionales genuinos en lugar de artefactos técnicos irrelevantes, maximizando la capacidad de generalización hacia nuevas muestras y contextos de uso.

Planeación de la Reducción de Dimensionalidad

Justificación para la Reducción de Dimensionalidad: En nuestro contexto de reconocimiento emocional por voz, la extracción comprehensiva de características acústicas genera espacios de alta dimensionalidad (estimamos >150 características por muestra) que pueden contener información redundante, ruido irrelevante, y correlaciones espurias. La reducción de dimensionalidad se vuelve crítica para mitigar el overfitting en nuestro dataset de tamaño moderado (~5,100 muestras), mejorar la eficiencia computacional durante entrenamiento e inferencia, y facilitar la visualización e interpretación de patrones emocionales subyacentes.

Técnicas Seleccionadas y Aplicación Temporal:

Análisis de Componentes Principales (PCA) - Fase de Exploración: Durante la etapa de análisis exploratorio, aplicaremos PCA para identificar las direcciones de máxima varianza en nuestro espacio de características acústicas. Esta técnica nos permitirá determinar cuánta información se puede retener con un número reducido de componentes (objetivo: 95% de varianza explicada) y visualizar la separabilidad natural entre categorías emocionales. PCA se aplicará después de la normalización z-score pero antes de cualquier técnica de balanceado de clases.

Análisis Discriminante Lineal (LDA) - Fase de Entrenamiento: LDA será nuestra técnica principal para reducción de dimensionalidad en el modelo final, ya que está específicamente diseñada para problemas de clasificación supervisada. LDA buscará direcciones que maximicen la separación entre nuestras siete clases emocionales mientras minimizan 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: Como enfoque complementario, implementaremos selección de características basada en pruebas estadísticas (ANOVA F-test) para identificar las características individuales más discriminativas. Esta técnica se aplicará en paralelo con LDA para validar la importancia de diferentes tipos de 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 comprensión), y segundo, inmediatamente antes del entrenamiento del modelo final (LDA para optimización). Esta implementación 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 y manteniendo la validez de la evaluación.

Validación de Efectividad: Se evaluará el impacto de la reducción de dimensionalidad comparando el rendimiento del modelo con y sin estas técnicas, utilizando validación cruzada estratificada para asegurar resultados robustos. Se documentará el trade-off entre reducción de dimensionalidad y pérdida de información discriminativa, estableciendo el punto óptimo para nuestro caso específico.

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 durante los próximos meses.

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.

000000000007