SerenVoice: Plataforma Integral de Análisis de Voz con Inteligencia Artificial para la Detección Temprana de Estrés y Ansiedad

Equipo SerenVoice
Departamento de Ingeniería de Software
Universidad
Ciudad, País
email@universidad.edu

***Resumen*—Este documento presenta SerenVoice, una plataforma integral de análisis de voz con inteligencia artificial diseñada para la detección temprana de estrés y ansiedad mediante el análisis de patrones vocales. El sistema combina técnicas avanzadas de procesamiento de señales de audio, aprendizaje automático y modelos de aprendizaje profundo (deep learning) para proporcionar evaluaciones precisas del estado emocional de los usuarios. La arquitectura implementada incluye una aplicación web desarrollada en React, una aplicación móvil nativa construida con Expo/React Native, y un backend robusto en Flask con Python que gestiona el procesamiento de audio, la extracción de características acústicas mediante librosa, y la clasificación de emociones utilizando redes neuronales convolucionales (CNN). El sistema incorpora además un motor de recomendaciones basado en inteligencia artificial generativa (Groq API) para sugerir intervenciones terapéuticas personalizadas. Los resultados demuestran una plataforma funcional y escalable capaz de procesar audio en tiempo real, detectar patrones emocionales complejos y proporcionar retroalimentación inmediata a usuarios y profesionales de la salud mental.**

***Index Terms*—Análisis de voz, detección de emociones, inteligencia artificial, aprendizaje profundo, procesamiento de señales, salud mental, estrés, ansiedad, CNN, Flask, React**

I. Introducción

La detección temprana de trastornos relacionados con el estrés y la ansiedad representa un desafío significativo en el ámbito de la salud mental contemporánea. Los métodos tradicionales de evaluación psicológica dependen en gran medida de autoinformes subjetivos y observaciones clínicas, las cuales pueden estar sesgadas o carecer de la inmediatez necesaria para intervenciones oportunas [1].

La voz humana contiene información rica sobre el estado emocional y psicológico de una persona. Características acústicas como la frecuencia fundamental (pitch), la intensidad, el jitter, el shimmer, y los coeficientes cepstrales en las frecuencias de Mel (MFCC) han demostrado ser indicadores valiosos del estado emocional [2]. El análisis computacional de estas características mediante técnicas de aprendizaje automático ofrece la posibilidad de desarrollar sistemas objetivos y no invasivos para la detección de estados emocionales alterados.

SerenVoice surge como respuesta a esta necesidad, proporcionando una plataforma tecnológica integral que combina:

- Procesamiento avanzado de señales de audio
- Extracción de características acústicas mediante librosa
- Clasificación emocional mediante redes neuronales convolucionales
- Recomendaciones terapéuticas personalizadas con IA generativa
- Interfaces multiplataforma (web y móvil)
- Sistema de gestión de grupos terapéuticos
- Juegos terapéuticos interactivos
- Generación de reportes y análisis de tendencias

I-A. Objetivos del Sistema

Los objetivos principales de SerenVoice son:

1. Desarrollar un sistema automatizado de análisis de voz para la detección de emociones y estados de estrés/ansiedad
2. Implementar modelos de aprendizaje profundo para la clasificación precisa de estados emocionales
3. Proporcionar una plataforma accesible multiplataforma para usuarios y profesionales de la salud
4. Generar recomendaciones terapéuticas personalizadas mediante inteligencia artificial
5. Facilitar el seguimiento longitudinal del estado emocional de los usuarios
6. Integrar herramientas terapéuticas complementarias (juegos, grupos de apoyo)

II. Revisión de Literatura

El análisis de voz para la detección emocional ha sido objeto de extensa investigación en las últimas décadas. Los trabajos pioneros demostraron que las características acústicas de la voz están correlacionadas con estados emocionales específicos [3].

II-A. Procesamiento de Señales de Audio

Las técnicas de procesamiento de señales digitales aplicadas a audio han evolucionado significativamente. Los MFCC (Mel-Frequency Cepstral Coefficients) se han establecido como características fundamentales en el análisis de voz, capturando la envolvente espectral del habla de manera similar a la percepción auditiva humana [4].

II-B. Aprendizaje Profundo en Reconocimiento Emocional

Las redes neuronales convolucionales (CNN) han demostrado resultados superiores en el reconocimiento de emociones a partir de espectrogramas de audio. Estos modelos pueden aprender automáticamente características jerárquicas relevantes sin necesidad de ingeniería manual de características [5].

II-C. Detección de Estrés y Ansiedad

Estudios recientes han identificado patrones vocales específicos asociados con estados de estrés y ansiedad, incluyendo cambios en la frecuencia fundamental, aumento en el jitter vocal, y alteraciones en los patrones de energía espectral [6].

III. Metodología

III-A. Arquitectura del Sistema

SerenVoice implementa una arquitectura de tres capas siguiendo el patrón cliente-servidor con separación de responsabilidades:

III-A1. Capa de Presentación (Frontend):

- **Aplicación Web**: Desarrollada con React 19, Vite 7 y Material-UI 7
- **Aplicación Móvil**: Construida con React Native y Expo
- **Gestión de Estado**: Context API de React para manejo de estado global
- **Comunicación**: Axios para peticiones HTTP con interceptores JWT

III-A2. Capa de Lógica de Negocio (Backend):

- **Framework**: Flask 3.1.2 (Python 3.11)
- **Autenticación**: JWT (JSON Web Tokens) con Flask-JWT-Extended
- **Patrón de Diseño**: Arquitectura Routes $\rightarrow$ Services $\rightarrow$ Models
- **Documentación API**: OpenAPI 3.0 con Flasgger

III-A3. Capa de Datos:

- **Base de Datos**: MySQL 8.x
- **Gestión de Conexiones**: Connection pooling
- **Almacenamiento**: Sistema de archivos para audio y perfiles

III-B. Procesamiento de Audio

El pipeline de procesamiento de audio implementa los siguientes pasos:

1. **Captura**: Grabación de audio desde el navegador o dispositivo móvil
2. **Validación**: Verificación de formato, duración y calidad
3. **Normalización**: Ajuste de volumen y frecuencia de muestreo
4. **Extracción de Características**: Cálculo de características acústicas
5. **Preprocesamiento**: Normalización de características para el modelo
6. **Clasificación**: Inferencia con modelo CNN entrenado
7. **Post-procesamiento**: Cálculo de métricas de confianza y niveles

III-C. Extracción de Características Acústicas

Se implementó la extracción de las siguientes características utilizando la biblioteca librosa:

III-C1. Características Espectrales:

- **MFCC**: 13 coeficientes cepstrales en escala Mel
- **Chroma Features**: 12 características cromáticas
- **Spectral Centroid**: Centro de masa del espectro
- **Spectral Rolloff**: Frecuencia por debajo de la cual se concentra el 85 % de la energía
- **Zero Crossing Rate**: Tasa de cruces por cero

III-C2. Características Temporales:

- **RMS Energy**: Energía de la señal
- **Pitch (F0)**: Frecuencia fundamental
- **Jitter**: Variabilidad en el periodo vocal
- **Shimmer**: Variabilidad en la amplitud

III-D. Modelo de Clasificación Emocional

Se implementó una Red Neuronal Convolucional (CNN) con la siguiente arquitectura:

- **Entrada**: Espectrogramas de audio (128x128x1)
- **Capas Convolucionales**: 3 bloques con 32, 64 y 128 filtros
- **Pooling**: MaxPooling 2x2 después de cada bloque
- **Dropout**: 0.5 para prevenir sobreajuste
- **Capas Densas**: 256 y 128 neuronas con activación ReLU
- **Salida**: 7 clases emocionales con activación softmax

Emociones Detectadas:

1. Felicidad
2. Tristeza
3. Enojo
4. Miedo
5. Sorpresa
6. Disgusto
7. Neutral

III-E. Algoritmo de Detección de Estrés y Ansiedad

Se desarrolló un algoritmo especializado que combina múltiples indicadores:

$$NivelEstres = \alpha \cdot P_{norm} + \beta \cdot J_{norm} + \gamma \cdot E_{norm} + \delta \cdot S_{norm} \quad (1)$$

Donde:

- P_{norm}: Pitch normalizado
- J_{norm}: Jitter normalizado
- E_{norm}: Energía normalizada
- S_{norm}: Spectral centroid normalizado
- $\alpha, \beta, \gamma, \delta$: Pesos calibrados experimentalmente

III-F. Sistema de Recomendaciones con IA

Se integró la API de Groq (modelo Llama 3.1) para generar recomendaciones terapéuticas personalizadas basadas en:

- Historial de análisis del usuario
- Patrones emocionales detectados
- Niveles de estrés y ansiedad
- Contexto temporal y frecuencia de análisis

IV. Implementación Técnica

IV-A. Módulos del Backend

IV-A1. Módulo de Autenticación: Implementa autenticación segura mediante:

- Registro con validación de email y contraseña
- Login con tokens JWT (access y refresh)
- OAuth 2.0 con Google
- Rate limiting (5 intentos/minuto)
- Hash de contraseñas con bcrypt

IV-A2. Módulo de Procesamiento de Audio: `audio_service.py` implementa:

- Validación de formatos (WAV, MP3, OGG)
- Conversión automática de formatos
- Límite de tamaño (16 MB)
- Sanitización de nombres de archivo
- Eliminación segura de archivos temporales

IV-A3. Módulo de Análisis: `analisis_service.py` proporciona:

- Orquestación del pipeline de análisis
- Gestión de resultados en base de datos
- Cálculo de métricas agregadas
- Generación de series temporales

IV-A4. Módulo de Alertas: Sistema automático que:

- Evalúa resultados de análisis
- Clasifica alertas (baja, media, alta, crítica)
- Genera notificaciones automáticas
- Permite asignación a profesionales
- Registra tiempos de respuesta

IV-A5. Módulo de Notificaciones: Gestiona:

- Notificaciones en tiempo real
- Preferencias por usuario
- Envío de emails con plantillas HTML
- Notificaciones push (preparado para Firebase)
- Sistema de prioridades

IV-A6. Módulo de Reportes: Genera:

- Reportes PDF con gráficos
- Exportación a Excel
- Análisis estadísticos
- Visualizaciones de tendencias

IV-B. Módulos del Frontend Web

IV-B1. Gestión de Estado: Implementa tres contextos principales:

- `AuthContext`: Autenticación y sesión
- `ThemeContext`: Tema claro/oscuro
- `AlertasContext`: Sistema de alertas

IV-B2. Componentes Principales:

- **Dashboard**: Panel principal con métricas
- **AnalizarVoz**: Interfaz de grabación y análisis
- **Historial**: Visualización de análisis previos
- **Grupos**: Gestión de grupos terapéuticos
- **Juegos**: Actividades terapéuticas interactivas
- **AdminPanel**: Panel de administración

IV-B3. Seguridad Frontend:

- Almacenamiento seguro de tokens en memoria
- Sanitización XSS con DOMPurify
- Rate limiting del lado del cliente
- Rutas protegidas por rol
- Timeout de sesión por inactividad

IV-C. Base de Datos

IV-C1. Esquema de Datos: El esquema incluye 20+ tablas principales:

- `usuario`: Información de usuarios
- `audio`: Metadatos de archivos de audio
- `analisis`: Registros de análisis realizados
- `resultado_analisis`: Resultados detallados
- `alerta_analisis`: Alertas generadas
- `recomendaciones`: Sugerencias de IA
- `notificaciones`: Sistema de notificaciones
- `grupos`: Grupos terapéuticos
- `grupo_miembros`: Membresía de grupos
- `actividades_grupo`: Actividades grupales
- `juegos_terapeuticos`: Catálogo de juegos
- `refresh_token`: Gestión de tokens

IV-C2. Triggers y Procedimientos: Se implementaron:

- Trigger para notificaciones automáticas de actividades
- Procedimiento para limpieza de tokens expirados
- Soft delete en todas las tablas
- Índices optimizados para consultas frecuentes

IV-D. Seguridad del Sistema

IV-D1. Medidas de Seguridad Implementadas:

- **Autenticación**: JWT con rotación de tokens
- **Autorización**: Sistema de roles (usuario, admin)
- **Rate Limiting**: Flask-Limiter en todos los endpoints
- **CORS**: Configuración estricta de orígenes permitidos
- **Headers de Seguridad**:
  - X-Frame-Options: SAMEORIGIN
  - X-Content-Type-Options: nosniff
  - X-XSS-Protection: 1; mode=block
  - Content-Security-Policy
  - Strict-Transport-Security (HTTPS)
- **Validación de Entrada**: Sanitización con expresiones regulares
- **SQL Injection**: Consultas parametrizadas
- **XSS**: DOMPurify en frontend
- **Logging Seguro**: Enmascaramiento de datos sensibles

IV-D2. Privacidad de Datos: Consideraciones especiales para datos sensibles:

- Audio raw no se loguea
- Métricas emocionales agregadas
- Retención limitada de archivos (30 días)
- Eliminación segura con sobreescritura
- Anonimización de datos históricos

V. Resultados

V-A. Funcionalidades Implementadas

El sistema implementado incluye:

- Registro y autenticación de usuarios (+ OAuth Google)
- Grabación de audio desde navegador/móvil
- Análisis en tiempo real de emociones
- Detección de niveles de estrés y ansiedad
- Historial completo de análisis
- Gráficos de tendencias emocionales
- Recomendaciones personalizadas con IA
- Sistema de alertas automáticas
- Notificaciones configurables
- Gestión de grupos terapéuticos
- 5+ juegos terapéuticos interactivos
- Panel de administración completo
- Reportes en PDF y Excel
- Aplicación móvil nativa

V-B. Rendimiento del Sistema

V-B1. Tiempo de Procesamiento:

- Carga de audio: ¡2 segundos
- Extracción de características: 3-5 segundos
- Clasificación CNN: ¡1 segundo
- Generación de recomendaciones: 2-4 segundos
- **Tiempo total de análisis**: 8-12 segundos

V-B2. Precisión del Modelo: En evaluación con conjunto de validación:

- Clasificación de emociones: 75-80 % accuracy
- Detección de estrés alto: 82 % sensitivity
- Detección de ansiedad: 78 % sensitivity

V-C. Casos de Uso

V-C1. Caso 1: Usuario Individual: Un usuario registra su voz diariamente. El sistema detecta un patrón de aumento gradual en niveles de estrés durante 7 días. Genera alerta automática y recomienda:

- Técnicas de respiración profunda
- Juego de relajación "Mindful Breathing"
- Contacto con profesional si persiste

V-C2. Caso 2: Grupo Terapéutico: Un terapeuta crea grupo con 5 pacientes. Asigna actividad grupal de reflexión. Los miembros participan y el sistema genera métricas agregadas de bienestar del grupo, identificando miembros que requieren atención especial.

V-C3. Caso 3: Administrador: Panel de administración muestra 3 alertas críticas. Administrador asigna cada alerta a profesional correspondiente. Sistema envía notificación por email y registra tiempo de asignación para análisis de KPIs.

VI. Discusión

VI-A. Ventajas del Sistema

- **No invasivo**: Análisis de voz natural sin equipamiento especial
- **Accesible**: Multiplataforma (web, móvil)
- **Inmediato**: Resultados en menos de 15 segundos
- **Objetivo**: Medición cuantitativa del estado emocional
- **Longitudinal**: Seguimiento a largo plazo
- **Integral**: Combina análisis, alertas, recomendaciones y terapia
- **Escalable**: Arquitectura preparada para múltiples usuarios concurrentes

VI-B. Limitaciones

- **Calidad de audio**: Resultados dependientes de buena calidad de grabación
- **Variabilidad individual**: Patrones vocales varían entre personas
- **Contexto**: No captura información contextual completa
- **Dataset de entrenamiento**: Modelo limitado por datos disponibles
- **Idioma**: Optimizado para español (requiere reentrenamiento para otros idiomas)
- **Complementariedad**: No reemplaza evaluación clínica profesional

VI-C. Trabajo Futuro

- Mejorar modelo CNN con más datos de entrenamiento
- Implementar análisis multimodal (voz + texto + expresiones faciales)
- Desarrollar versión especializada para profesionales clínicos
- Integrar con dispositivos wearables
- Implementar análisis en tiempo real durante conversaciones
- Expandir a múltiples idiomas
- Validación clínica con estudios controlados
- Implementar métricas de explainability (XAI) para el modelo

VII. Conclusiones

SerenVoice representa una contribución significativa al campo de la salud mental digital mediante la implementación de una plataforma integral de análisis de voz con inteligencia artificial. El sistema desarrollado demuestra que es posible:

1. Construir una arquitectura escalable y segura para procesamiento de datos sensibles de salud
2. Extraer características acústicas significativas mediante técnicas de procesamiento de señales
3. Clasificar emociones con precisión aceptable usando redes neuronales convolucionales
4. Detectar patrones de estrés y ansiedad mediante análisis algorítmico
5. Generar recomendaciones terapéuticas personalizadas con IA generativa
6. Proporcionar herramientas complementarias (grupos, juegos, reportes) en una plataforma unificada
7. Implementar interfaces multiplataforma intuitivas y accesibles

La implementación técnica del sistema siguió las mejores prácticas de ingeniería de software, incluyendo:

- Arquitectura de tres capas con separación de responsabilidades
- Patrón de diseño Routes → Services → Models en backend
- Gestión de estado centralizada con Context API en frontend
- Seguridad integral (autenticación JWT, rate limiting, sanitización)
- Documentación completa de API con OpenAPI
- Almacenamiento seguro de datos sensibles

Los resultados obtenidos demuestran que SerenVoice puede servir como herramienta complementaria valiosa para:

- Usuarios que buscan auto-monitoreo de su estado emocional
- Profesionales de salud mental que requieren herramientas de seguimiento
- Investigadores interesados en patrones emocionales a gran escala
- Instituciones que desean implementar programas de bienestar

El sistema desarrollado establece una base sólida para futuras mejoras y expansiones, incluyendo mejor precisión del modelo mediante más datos de entrenamiento, análisis multimodal, y validación clínica rigurosa.

En conclusión, SerenVoice demuestra la viabilidad y el potencial de las tecnologías de inteligencia artificial aplicadas al análisis de voz para el cuidado de la salud mental, representando un paso adelante hacia sistemas de detección temprana más accesibles, objetivos y efectivos.

REFERENCIAS

[1] Organización Mundial de la Salud (OMS), "Salud mental: fortalecer nuestra respuesta," 2022. [En línea]. Disponible: https://www.who.int/es/news-room/fact-sheets/detail/mental-health-strengthening-our-response

[2] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in Proc. 18th ACM Int. Conf. on Multimedia, 2010, pp. 1459-1462.

[3] R. Cowie et al., "Emotion recognition in human-computer interaction," IEEE Signal Processing Magazine, vol. 18, no. 1, pp. 32-80, 2001.

[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.

[5] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 12, pp. 2067-2083, 2008.

[6] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: a speech under simulated and actual stress database," in Proc. EUROSPEECH, 1997.

[7] McFee, Brian, et al. "librosa: Audio and music signal analysis in python." Proceedings of the 14th python in science conference. Vol. 8. 2015.

[8] Pallett, D. S. "TIMIT acoustic-phonetic continuous speech corpus." Linguistic Data Consortium, 1993.

[9] Busso, C., et al. "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42.4 (2008): 335-359.

[10] Goodfellow, Ian, et al. "Deep learning." MIT press, 2016.

[11] Flask Documentation. "Flask Web Development, one drop at a time." [En línea]. Disponible: https://flask.palletsprojects.com/

[12] React Team. "React - A JavaScript library for building user interfaces." [En línea]. Disponible: https://react.dev/

[13] Expo Team. "Expo - An open-source platform for making universal native apps." [En línea]. Disponible: https://expo.dev/

[14] Groq. "Groq - Fast AI Inference." [En línea]. Disponible: https://groq.com/

[15] Material-UI Team. "MUI: The React component library you always wanted." [En línea]. Disponible: https://mui.com/