

Diseño y evaluación de un sistema DSP para la rehabilitación vocal post-cordectomía mediante reconstrucción espectral e IA

Alfonso Gamboa Rubén & Flores Montero Edsel Yetlanezi

Resumen— La cordectomía, procedimiento quirúrgico que implica la extirpación parcial o total de los pliegues vocales, compromete severamente la capacidad comunicativa del paciente, afectando su identidad y calidad de vida. Este proyecto presenta el diseño y evaluación de un sistema de procesamiento digital de señales (DSP) orientado a la rehabilitación vocal no invasiva mediante la reconstrucción espectral. La metodología evolucionó a través de tres fases iterativas: una aproximación inicial en el dominio de la frecuencia (FFT global), un modelo adaptativo basado en metadatos y filtrado de intensidad adaptable, y finalmente, la implementación basada en la Transformada de Fourier de Tiempo Corto (STFT) y estimadores estadísticos (MMSE-STSA). Los resultados experimentales demostraron que, si bien la sustracción de ruido estacionario mediante algoritmos de Wiener y Ephraim-Malah es efectiva, la reconstrucción de la voz requiere una intervención más compleja a nivel de las micro-características que forman la voz para preservar la identidad del paciente y evitar artefactos o distorsiones de la voz. El estudio concluye proponiendo una versión adicional de experimentación modular que implementa herramientas de inteligencia artificial.

Index Terms—Procesamiento Digital de Señales (DSP), Transformada de Fourier de Tiempo Corto (STFT), Filtro de Wiener, Filtro Savitzky-Golay, Detección de Actividad de Voz (VAD), Análisis Espectral, Rehabilitación Fónica, Python, Cordectomía, STFT, Ephraim-Malah, Formantes. Inteligencia Artificial (IA), Reinforcement Learning from Human Feedback (RLHF), Red Neuronal, Speech Emotion Recognition (SER), Modelos de Síntesis.

Abstract— Cordectomy, a surgical procedure involving the partial or total removal of the vocal folds, severely compromises a patient's communicative ability, affecting their identity and quality of life. This project presents the design and evaluation of a digital signal processing (DSP) system for non-invasive voice rehabilitation through spectral reconstruction. The methodology evolved through three iterative phases: an initial approach in the frequency domain (global FFT), an adaptive model based on metadata and adaptive intensity filtering, and finally, implementation based on the Short Time Fourier Transform (STFT) and statistical estimators (MMSE-STSA). The experimental results demonstrated that, while stationary noise subtraction using Wiener and Ephraim-Malah algorithms is effective, voice reconstruction requires more complex intervention at the level of the micro-features that constitute the voice to preserve the patient's identity and avoid artifacts or voice distortions. The study concludes by proposing an additional version of modular experimentation that implements artificial intelligence tools.

Index Terms—Digital Signal Processing (DSP), Short Time Fourier Transform (STFT), Wiener Filter, Savitzky-Golay Filter, Voice Activity Detection (VAD), Spectral Analysis, Voice Rehabilitation, Python, Cordectomy, STFT, Ephraim-Malah, Formants, Artificial Intelligence (AI), Reinforcement Learning from Human Feedback (RLHF), Neural Network, Speech Emotion Recognition (SER), Synthesis Models.

ÍNDICE

I. Objetivos del Proyecto	2
I-A. Objetivo General	2
I-B. Objetivos Específicos	2
II. Marco Teórico	2
II-A. Software y Herramientas	2
II-B. Fundamentos Matemáticos	3
II-C. Conceptos Estadísticos y de Programación	3
II-D. Acústica y Características de la Voz . .	3
II-E. Hardware	4
II-F. Inteligencia artificial (suplementario) .	4
III. Metodología	5
III-A. Versión 1.0: Análisis Espectral	5
III-B. Versión 2.0: Metadatos	6
III-C. Versión 3.0: Estimación MMSE	7
III-D. Versión 4.0: IA (Propuesta)	7
IV. Conclusiones Generales	7
Referencias	7

I. OBJETIVOS DEL PROYECTO

I-A. Objetivo General

Desarrollar y evaluar algoritmos de procesamiento digital de señales basado en análisis espectral de tiempo corto y modelado estadístico, en relación a la capacidad de mejorar la calidad de la voz y restaurar parcialmente las características tímbricas en grabaciones de voz de pacientes sometidos a cordectomía.

I-B. Objetivos Específicos

1. **Caracterización Acústica:** Construir una base de datos pareada (pre y post-operatoria) para identificar los patrones de pérdida armónica y deformación espectral en el dominio de la frecuencia causados por la intervención quirúrgica.
2. **Optimización de la Relación Señal-Ruido (SNR):** Implementar y comparar técnicas de sustracción espectral (Noisereduce vs. Ephraim-Malah/VAD) para minimizar el ruido estacionario inherente a la fonación soplada sin degradar los transitorios de la voz.
3. **Reconstrucción Espectral:** Experimentar con algoritmos de transferencia de características que utilicen una máscara espectral diferencial (T_{dB}) para proyectar el timbre e identidad del sonido vocal (envolvente de frecuencia de la voz) sano sobre la señal patológica.
4. **Validación Técnica:** Evaluar mediante espectrogramas y gráficas comparativas, la efectividad de los algoritmos en la rehabilitación de formantes y reducción de artefactos y desfase de frecuencias armónicas.

II. MARCO TEÓRICO

II-A. Software y Herramientas

II-A1. Lenguaje de Programación: Python: Para la implementación de los algoritmos de procesamiento de audio, se seleccionó Python como lenguaje núcleo. Esta elección se basa en la extensa documentación y la robustez de su ecosistema de librerías científicas (*SciPy Stack*, etc.), que permiten prototipar y desplegar soluciones complejas matemáticas, estadísticas y procesamiento de señales con alta eficiencia [1], [2].

II-A2. Librerías Especializadas:

- **Numpy (numpy):** Fundamental para la manipulación de arreglos multidimensionales [3]. Se utiliza para convertir los flujos de bits de audio en arreglos de punto flotante (`float32`) [4].
- **Pydub (pydub):** Interfaz de alto nivel para el manejo de archivos de audio (I/O).
- **Scipy (scipy.signal):** Proporciona herramientas matemáticas avanzadas [5], [6].
 - *Filtro Savitzky-Golay:* Utilizado para el suavizado de curvas espectrales [7], [8]. A diferencia de un promedio móvil simple que tiende a aplanar los picos, este filtro ajusta un polinomio de orden k a una ventana de puntos m mediante el método de mínimos cuadrados [9]. Esto permite reducir el ruido de alta frecuencia preservando los momentos espectrales importantes de la voz (formantes), manteniendo la

identidad tímbrica del paciente [10]. Su formulación discreta es:

$$Y_j = \sum_{i=-(m-1)/2}^{(m-1)/2} C_i \cdot y_{j+i} \quad (1)$$

Donde Y_j es el valor suavizado, y los datos crudos, m el tamaño de la ventana (impar) y C_i los coeficientes de convolución derivados del polinomio ajustado [11].

- *STFT / iSTFT*: Funciones base para la transformación desde el dominio del tiempo al dominio de la frecuencia y viceversa (ver sección 2.1) [12], [13].

II-B. Fundamentos Matemáticos

II-B1. Transformada de Fourier de Tiempo Corto (STFT): Dado que la voz es una señal no estacionaria (los valores de frecuencia varían en el tiempo), la Transformada de Fourier clásica es insuficiente [14]. La STFT divide la señal en "ventanas" temporales superpuestas y aplica la FFT a cada una de ellas [15].

Matemáticamente, para una señal $x(n)$ y una ventana $w(n)$:

$$X(m, k) = \sum_{n=0}^{N-1} x(n + mH)w(n)e^{-j\frac{2\pi}{N}kn} \quad (2)$$

Donde m es el índice temporal, k el índice de frecuencia y H el tamaño del salto o ventana [16]. Esto genera una matriz compleja que representa la magnitud y fase de la señal a lo largo del tiempo [17].

II-B2. Transformada Inversa de Fourier de Tiempo Corto (iSTFT): Es el proceso de reconstrucción de la señal de audio al dominio del tiempo a partir de la matriz espectral modificada [18]. Utiliza un método de superposición y suma para reintegrar las ventanas procesadas, compensando la modulación introducida por la función de ventana $w(n)$ para evitar artefactos de discontinuidad [19].

II-B3. Algoritmo de Filtrado: Wiener y Ephraim-Malah: Si bien el algoritmo Ephraim-Malah (MMSE-STSA) es el estándar de oro estadístico para la estimación de amplitud logarítmica [20], el código implementado utiliza una variante basada en el Filtro de Wiener.

Este filtro opera bajo el principio de minimizar el Error Cuadrático Medio (MSE) entre la señal estimada y la señal real [21]. A diferencia de la sustracción espectral simple, el filtro de Wiener calcula una ganancia de transferencia óptima $W(f)$ para cada frecuencia basada en la Relación Señal-Ruido (SNR) [22], [23]:

$$W(f) = \frac{P_{señal}(f)}{P_{señal}(f) + P_{ruido}(f)} \quad (3)$$

Donde P representa la densidad espectral de potencia. El algoritmo atenúa las frecuencias donde la potencia del ruido domina sobre la señal, y preserva aquellas donde la señal es fuerte [24].

II-B4. Detección Activa de Voz (VAD) Estadística: El VAD (Voice Activity Detection) es un mecanismo para distinguir segmentos de voz útil de segmentos de silencio o ruido de fondo [25].

En la implementación actual, se utiliza un VAD basado en energía:

1. Se calcula la energía total de cada trama espectral.
2. Se establece un umbral dinámico (ej. percentil 10 de la energía global).
3. Las tramas por debajo del umbral se clasifican como "perfil de ruido", permitiendo al sistema aprender las características estadísticas del ruido estacionario para sustraerlo posteriormente sin afectar la voz.

II-C. Conceptos Estadísticos y de Programación

II-C1. Expresiones Regulares (Regex): Las Expresiones Regulares (*Regular Expressions*) son secuencias de caracteres que forman patrones de búsqueda. En este proyecto, se utilizan para el análisis sintáctico de los nombres de archivo, permitiendo la extracción automatizada de metadatos incrustados (género, identificación del paciente, calidad de la grabación) para adaptar los parámetros del algoritmo de procesamiento dinámicamente.

II-C2. Desviación Estándar (σ): Medida de dispersión que indica qué tan extendidos están los valores de un conjunto de datos respecto a su media. En la reducción de ruido, se utiliza para definir umbrales de tolerancia (`n_std_thresh`) en la librería 'Scypy'. Un píxel espectral se considera "señal" solo si su magnitud supera la media del ruido más n veces su desviación estándar:

$$\text{Umbral}(f) = \mu_{\text{ruido}}(f) + (n \cdot \sigma_{\text{ruido}}(f)) \quad (4)$$

II-D. Acústica y Características de la Voz

II-D1. Timbre: Calidad psicoacústica que permite distinguir dos sonidos de igual frecuencia fundamental e intensidad. El timbre está determinado por la envolvente espectral y la distribución de energía en los armónicos superiores [26]. Preservar el timbre es el objetivo principal del uso de filtros conservadores como Savitzky-Golay.

II-D2. Frecuencia Fundamental (F_0): Corresponde a la frecuencia de vibración de las cuerdas vocales y determina la altura tonal (pitch) percibida de la voz [27].

II-D3. Armónicos: Son los múltiplos enteros de la frecuencia fundamental ($2F_0, 3F_0, \dots$). La riqueza y amplitud de estos armónicos definen la claridad y la resonancia de la voz. En patologías laríngeas, los armónicos suelen perderse o mezclarse con ruido turbulento.

II-D4. Los Formantes: Los picos de resonancia espectral generados por el filtro del tracto vocal se denominan formantes [28], [29]. Son esenciales para la inteligibilidad y la identidad del hablante:

- **F1 y F2 (Formantes Vocálicos)**: Son los dos primeros picos de energía y determinan qué vocal se está pronunciando [30], [31]. Por ejemplo, la vocal /a/ tiene un F1 alto y un F2 bajo, mientras que la /i/ tiene un F1 bajo y un F2 muy alto [32]. La inteligibilidad del mensaje depende casi

exclusivamente de la preservación de estos dos formantes [33].

- F3, F4 y F5 (Formantes de Timbre): Ubicados en frecuencias superiores (generalmente por encima de 2500 Hz), estos formantes son estáticos y dependen de la anatomía fija del paciente [34]. Son responsables del timbre personal y la identidad del hablante [35].

En el contexto de este proyecto, el objetivo de la reconstrucción espectral no es solo recuperar el volumen, sino restaurar la estructura de los formantes superiores (F3-F5) que suelen perderse en la señal ruidosa post-cordectomía.

II-D5. La cordectomía: La cordectomía es una intervención quirúrgica indicada principalmente para el tratamiento de neoplasias laríngeas, que consiste en la resección total o parcial de las cuerdas vocales [36], [37]. Dependiendo de la extensión del tejido extirpado (desde una cordectomía subepitelial hasta una resección transmuscular o total), las consecuencias fonatorias varían desde una disfonía leve hasta una afonía severa [38], [39].

La alteración anatómica impide el cierre glótico completo, generando un escape de aire excesivo (voz soplada) y reduciendo la capacidad de vibración mucosa necesaria para generar una frecuencia fundamental (F_0) estable [40]. Clínicamente, esto se traduce en una reducción drástica de la intensidad vocal, fatiga al hablar y pérdida de definición armónica [41], [42].

II-D6. Teoría Fuente - Filtro de la Producción Vocal: El modelo acústico estándar para describir la generación de la voz es la **Teoría Fuente-Filtro** (Fant, 1960) [43], [44]. Este modelo descompone el proceso en dos etapas independientes pero interconectadas:

1. La Fuente (*Source*): Proporcionada por la vibración de los pliegues vocales en la laringe, que genera un sonido complejo rico en armónicos (el "zumbido" base) [45]. En pacientes con cordectomía, esta fuente es ruidosa y aperiódica debido a la irregularidad del tejido cicatricial.
2. El Filtro (*Filter*): Constituido por el tracto vocal (faringe, cavidad oral y nasal) [46]. Este conducto actúa como un resonador acústico que amplifica ciertas frecuencias y atenúa otras, esculpiendo el sonido final [47].

II-D7. Geometría del Aparato Fonador y Resonancia: La laringe y el tracto vocal pueden modelarse físicamente como un tubo acústico de sección variable, cerrado en un extremo (glotis) y abierto en el otro (labios) [48]. La geometría de este "tubo" es determinante para la voz:

- Longitud del Tracto: Determina las frecuencias de resonancia base. Un tracto más largo (típico en hombres) resuena a frecuencias más bajas, mientras que uno más corto (mujeres y niños) lo hace a frecuencias más altas [49].
- Configuración Transversal: Los movimientos de la lengua, mandíbula y labios modifican el área de sección transversal del tubo a lo largo de su longitud [50]. Estas constricciones cambian las frecuencias de resonancia del sistema, permitiendo la articulación de diferentes fonemas a pesar de que la fuente sonora sea la misma.

II-E. Hardware

II-E1. Micrófono: Razer Seiren Mini: Es un dispositivo de transducción electroacústica de tipo condensador (cápsula de 14mm).

- Patrón Polar: Supercardioides. Crucial para el proyecto ya que maximiza la rechazo al ruido incidente desde los laterales y la parte trasera, capturando prioritariamente la fuente directa (paciente).
- Respuesta en Frecuencia: 20 Hz - 20 kHz, cubriendo el espectro vocal completo.
- Especificaciones Digitales: Muestreo a 44.1/48 kHz y profundidad de 16 bits. Aunque presenta un ruido propio (self-noise) mayor que equipos de estudio de gama alta, su relación costo-beneficio y conectividad USB directa lo hacen viable para entornos clínicos no especializados.

II-E2. Sistema de Monitoreo: KZ ZSN Pro: Audífonos tipo *In-Ear Monitor* (IEM) de arquitectura híbrida.

- Drivers: Combina un driver dinámico (para graves/medios) y un driver *Balanced Armature* (para agudos).
- Respuesta: 7 Hz - 40 kHz.
- Firma Sonora: Ecualización en "V". Esta característica resalta los agudos y los graves. Para fines de evaluación cualitativa en este proyecto, la acentuación de agudos del driver *Balanced Armature* facilita la detección de ruidos sibilantes, soplos y artefactos digitales (glitches) introducidos durante el procesamiento de la señal.

II-F. Inteligencia artificial (suplementario)

II-F1. Representación y Caracterización de la Voz:

II-F1a. Red Neuronal Artificial (ANN): Modelo computacional inspirado en la biología que constituye la base del aprendizaje profundo (Deep Learning). En el procesamiento de audio, estas redes aprenden comportamientos no lineales complejos entre señales de entrada (audio crudo o espectrogramas) y representaciones latentes (representación condensada de las características esenciales), superando las limitaciones de los filtros lineales tradicionales [51].

II-F1b. Speaker Embedding (Incrustación del Hablante): Es una representación vectorial compacta de longitud fija que captura las características acústicas únicas de la identidad de un hablante (timbre, entonación, estilo), independientemente del contenido lingüístico. En sistemas de conversión de voz, este vector permite condicionar al modelo para que genere audio con la "identidad" de un sujeto específico, actuando como una firma biométrica digital [52].

II-F1c. Tokens Semánticos y Unidades de Contenido: A diferencia de los fonemas tradicionales, los tokens semánticos son representaciones discretas derivadas de modelos auto-supervisados (como HuBERT) que capturan y separan la información lingüística y características complejas de la voz. [53].

II-F1d. Unidades de Contenido: Son los clústeres discretos resultantes de cuantizar estas representaciones latentes (representación comprimida de puntos de datos que conserva solo las características esenciales). Es decir, Permiten que el sistema manipule el "qué se dice" sin alterar el "quién lo dice", siendo fundamentales para la conversión de voz *zero-shot* (sin entrenamiento previo extensivo).

II-F2. Arquitecturas de Aprendizaje Auto-Supervisado (SSL):

II-F2a. HuBERT (Hidden Unit BERT): Modelo de aprendizaje auto-supervisado que aprende representaciones de habla mediante la predicción enmascarada de unidades ocultas. A diferencia de otros modelos, HuBERT utiliza un paso de agrupamiento (clustering) offline (generalmente K-means) para generar etiquetas objetivo discretas, obligando al modelo a aprender tanto la estructura acústica como lingüística continua [54].

II-F2b. ContentVec: Es una evolución de la arquitectura HuBERT diseñada específicamente para tareas de conversión de voz. Su innovación radica en la capacidad de "desenmarañar" (disentangle) la información del hablante de la información del contenido. ContentVec impone restricciones para que las representaciones aprendidas sean invariantes al hablante, mejorando drásticamente el rendimiento en la conversión de identidad [55].

II-F2c. Information Bottleneck (Cuello de Botella de Información): Principio teórico aplicado en Deep Learning que postula que una red neuronal óptima debe comprimir la entrada X en una representación compacta Z que retenga solo la información relevante para la tarea objetivo Y , descartando el ruido y detalles irrelevantes (como el ruido de fondo o variaciones intra-hablante no deseadas) [56].

II-F3. Modelos de Síntesis y Conversión de Voz:

II-F3a. Modelos de Síntesis (Generativos): Sistemas de IA capaces de generar formas de onda de audio a partir de representaciones intermedias (texto o espectrogramas). En la actualidad, predominan los modelos probabilísticos y adversariales que pueden generar habla con alta fidelidad y naturalidad perceptiva [57].

II-F3b. Arquitectura RVC (Retrieval-based Voice Conversion): Modelo de conversión de voz que combina la extracción de características de contenido (vía HuBERT) con un sistema de recuperación de información. RVC utiliza una base de datos de incrustaciones del hablante objetivo y busca los vectores más similares para fusionarlos con la fuente, preservando el timbre con alta fidelidad incluso con pocos datos de entrenamiento [58].

II-F3c. "Centroide" del Vector: En el contexto de RVC y búsqueda vectorial (Faiss), se refiere al promedio de los vectores de características dentro de un clúster específico. Durante la inferencia, el modelo busca el centroide más cercano en el espacio latente del hablante objetivo para reemplazar o suavizar las características de la voz fuente, garantizando una conversión más estable.

II-F3d. So-VITS-SVC (SoftVC VITS Singing Voice Conversion): Arquitectura especializada en la conversión de voz cantada. Combina un codificador de contenido suave (SoftVC) que preserva la entonación original, con el modelo generativo VITS (Inferencia Variacional con aprendizaje adversarial). Permite transferir el timbre de un cantante a otro manteniendo la melodía y expresividad original [59].

II-F3e. XTTS v2: Modelo de síntesis de voz (TTS) de última generación desarrollado por Coqui AI. Utiliza una arquitectura basada en transformadores tipo GPT-2 para predecir tokens de audio ("quantized audio tokens") condicionados por

un vector de hablante, permitiendo clonación de voz multilingüe de alta calidad con solo unos segundos de audio de referencia [60].

II-F3f. Whisper: Modelo de reconocimiento automático del habla (ASR) desarrollado por OpenAI, basado en la arquitectura Transformer. Entrenado con 680,000 horas de audio multilingüe con supervisión débil, es capaz de generar transcripciones robustas y marcas de tiempo precisas, sirviendo a menudo como codificador "para extraer texto o fonemas en tuberías de conversión" [61].

II-F4. Componentes de Generación de Audio y Control:

II-F4a. Vocoder y HiFi-GAN: El vocoder es el componente final en la cadena de síntesis que transforma las representaciones acústicas intermedias en la forma de onda audible. HiFi-GAN es un vocoder neuronal basado en Redes Generativas Adversariales (GAN), considerado el estándar actual en síntesis de alta fidelidad debido a su eficiencia computacional y capacidad para generar audio libre de artefactos metálicos, superando a métodos autoregresivos más lentos como WaveNet [62].

II-F4b. RLHF (Reinforcement Learning from Human Feedback): Técnica de aprendizaje automático donde el modelo es ajustado utilizando retroalimentación humana directa. En síntesis de voz, se usa para alinear la prosodia o el estilo emocional del modelo con la preferencia humana, utilizando un "modelo de recompensa" entrenado para juzgar la naturalidad del audio generado [63].

II-F4c. SER (Speech Emotion Recognition): Tecnología orientada a detectar y clasificar el estado emocional del hablante (alegría, tristeza, ira) a partir de la señal de audio. En rehabilitación vocal, puede utilizarse para evaluar si la voz sintetizada logra transmitir la intención emocional correcta del paciente [64].

III. METODOLOGÍA

III-A. Versión 1.0: Análisis Espectral

III-A1. Generalidades: La arquitectura del algoritmo se ha estructurado en dos fases fundamentales. La primera fase tiene como objetivo la importación, preprocesamiento, optimización y exportación de recursos (datos y metadatos). La segunda fase se centra en el procesamiento digital de señales (DSP) para la rehabilitación de la voz mediante la reconstrucción de archivos de audio.

Los datos de para la versión post-cordectomía consisten en simulaciones de lectura del mismo dialogo en una versión pre-cordectomía, realizadas por el paciente en un entorno controlado con bajo ruido estacionario. Para la captura se utilizó un micrófono Razer Seiren Mini con las siguientes especificaciones técnicas:

III-A1a. Nomenclatura de Archivos: Para organizar los datos de manera eficiente, se estableció el siguiente protocolo de nomenclatura: '{CódigoPaciente}-{Origen}{Número}.{Extensión}'

- Código del paciente: Iniciales del nombre (ej. GFA).
- Origen del archivo:
 - A: Audio pre-cordectomía.
 - B: Audio post-cordectomía.

- Número identificador: Valor numérico consecutivo. Un prefijo '0' indica que el archivo es independiente y carece de una contraparte recíproca de origen pre-cordectomía.
- Identificador del archivo (ID): '{Origen}{Número}'
- Formato utilizado:
 - Audio: '.mp3', '.opus', '.ogg', '.wav', etc.
 - Datos: '.txt', '.csv'.
 - Gráficos: '.pdf'.

Ejemplo: 'GFA-B1.ogg (Paciente G.F.A., primer audio de origen post-cordectomía, formato OGG).

III-A2. Descripción de Algoritmos (v1.0):

III-A2a. Algoritmo 1.1.0: Preprocesamiento y Análisis Comparativos: Esta fase gestiona la importación de señales de audio pareadas (pre y post-cordectomía). Se aplica una reducción de ruido mediante el algoritmo de sustracción espectral incluido por la librería 'Noisereduce'. Posteriormente, se transforma la señal al dominio de la frecuencia utilizando la Transformada Rápida de Fourier (FFT) para obtener la magnitud del espectro.

Sea $x(n)$ la señal de entrada, su representación en frecuencia $X(k)$ se define como:

$$X(k) = \sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}kn} \quad (5)$$

III-A3. Algoritmo 1.2.0: Rehabilitación: Se calcula un factor de ganancia $G(k)$:

$$G(k) = \frac{|X_{pre}(k)|}{|X_{post}(k)|} \quad (6)$$

Señal rehabilitada $Y_{reh}(k)$:

$$Y_{reh}(k) = X_{post}(k) \cdot G(k) \quad (7)$$

Aplicación de la IFFT:

$$y(n) = \frac{1}{N} \sum_{k=0}^{N-1} Y_{reh}(k)e^{j\frac{2\pi}{N}kn} \quad (8)$$

III-A4. Algoritmo 1.1.1: Archivos Independientes: Iteración para procesamiento unilateral.

III-A5. Algoritmo 1.2.1: Suma Diferencial: Compensación aditiva basada en promedios espectrales.

$$\Delta_{media}(k) = \mu_{pre}(k) - \mu_{post}(k) \quad (9)$$

$$|Y_{rehab}(k)| = |X_{post}(k)| + \Delta_{media}(k) \quad (10)$$

III-A6. Algoritmo 1.2.2: Inyección Proyectada: Control estadístico para evitar distorsión usando μ y σ .

$$G_{corr}(k) = \frac{|X_{post}(k)| + I_{proy}(k)}{\mu_{pre}(k)} \quad (11)$$

III-B. Versión 2.0: Metadatos

III-B1. Algoritmo 2.1.0: Filtrado Selectivo: Uso de REGEX para configuración dinámica de filtros.

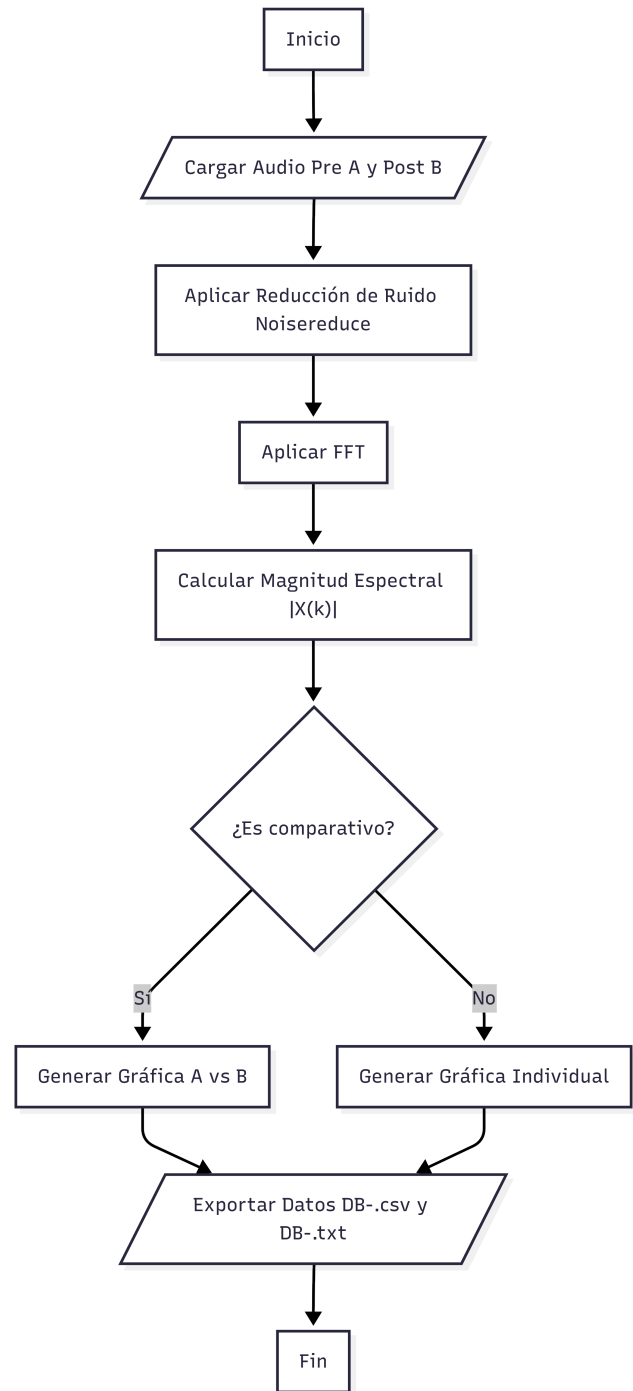


Figura 1. Diagrama de flujo del Algoritmo 1.1.0

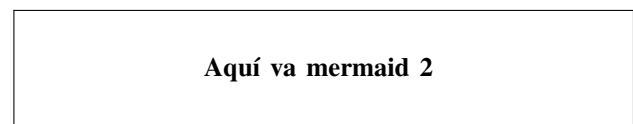


Figura 2. Diagrama de flujo del Algoritmo 1.2.0

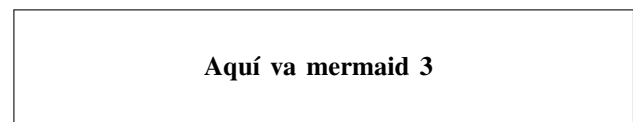


Figura 3. Diagrama de flujo del Algoritmo 1.1.1

Aquí va mermaid 4

Figura 4. Diagrama de flujo del Algoritmo 1.2.1

Aquí va mermaid 5

Figura 5. Diagrama de flujo del Algoritmo 1.2.2

III-B2. Algoritmo 2.2.1.0: Modelo Espectral: Promedio ponderado (w) según calidad de grabación.

$$S_{ideal}(k) = \frac{\sum_{i=1}^N (X_i(k) \cdot w_i)}{\sum_{i=1}^N w_i} \quad (12)$$

III-B3. Algoritmo 2.2.2.0: Reconstrucción Híbrida: Sustitución espectral en banda alta y amplificación en media, suavizado con Savitzky-Golay.

III-C. Versión 3.0: Estimación MMSE

III-C1. Algoritmo 3.1.0: MMSE-STSA con VAD: Estimador de Ephraim-Malah con Detección de Actividad de Voz.

III-C2. Algoritmo 3.2.0: Visualización: Generación de espectrogramas logarítmicos.

$$S(m, k) = 10 \cdot \log_{10}(|X(m, k)|^2) \quad (13)$$

III-C3. Algoritmo 3.3.0: Máscara de Transferencia: Definición de Función de Transferencia Objetivo (T_{dB}):

$$T_{dB}(k) = \mu_{PRE,dB}(k) - \mu_{POST,dB}(k) \quad (14)$$

Aplicación a la señal post-operatoria:

$$|Y_{reh}(m, k)| = |X_{post}(m, k)| \cdot 10^{\frac{T_{dB}(k)}{20}} \quad (15)$$

III-D. Versión 4.0: IA (Propuesta)

III-D1. Fase 1: Reconstrucción Offline: Arquitectura ASR-TTS (Whisper + XTTS).

$$P(Y|T, S) = \prod_n P(y_n | y_{<n}, T, S) \quad (16)$$

Aquí va mermaid 6

Figura 6. Diagrama de flujo del Algoritmo 2.1.0

Aquí va mermaid 7

Figura 7. Diagrama de flujo del Algoritmo 2.2.1.0

Aquí va mermaid 8

Figura 8. Diagrama de flujo del Algoritmo 2.2.2.0

Aquí va mermaid 9

Figura 9. Diagrama de flujo del Algoritmo 3.1.0

III-D2. Fase 2: Optimización de Preferencias: Ajuste de vectores de estilo (RLHF simplificado).

III-D3. Fase 3: Streaming Baja Latencia: Conversión RVC/So-VITS-SVC con Information Bottleneck.

$$Y_{str} = Dec(Content(X_{post}), F0_{smooth}, S_{pre}) \quad (17)$$

III-D4. Fase 4: Modulación Emocional: Integración de Speech Emotion Recognition (SER).

IV. CONCLUSIONES GENERALES

La evolución del proyecto permite establecer hallazgos críticos. La voz humana no puede tratarse como un fenómeno estático; la aproximación global resulta insuficiente. El éxito de la rehabilitación espectral depende de la capacidad de ajustar la señal en una escala temporal micro-segmentada.

Se evidenció una dicotomía entre limpieza de señal y fidelidad tímbrica. Los algoritmos de sustracción espectral (Wiener+VAD) están limitados a ruido estacionario. Un hallazgo fundamental fue la criticidad de la alineación temporal; cualquier desviación rítmica genera incoherencias de fase. El futuro apunta hacia la caracterización multidimensional empleando tensores y matrices de mayores dimensiones.

REFERENCIAS

- [1] T. E. Oliphant, A guide to NumPy, USA: Trelgol Publishing, vol. 1, 2006.
- [2] W. McKinney, "Python for data analysis: Data wrangling with Pandas, NumPy, and IPython," O'Reilly Media, Inc., 2012.
- [3] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22-30, 2011.
- [4] J. M. Kizza, "Python for scientific computing," in *Guide to Computer Network Security*, Springer, 2017.
- [5] P. Virtanen et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, 2020.

Aquí va mermaid 10

Figura 10. Diagrama de flujo del Algoritmo 3.2.0

Aquí va mermaid 11

Figura 11. Diagrama de flujo del Algoritmo 3.3.0

Aquí va mermaid 12 (Diagrama Fase 1)

Figura 12. Diagrama de flujo de la Fase 1: Reconstrucción Offline

Aquí va mermaid 13 (Diagrama Fase 2)

Figura 13. Diagrama de flujo de la Fase 2: Optimización de Preferencias

Aquí va mermaid 14

Figura 14. Diagrama de flujo: Fase 3

Aquí va mermaid 15

Figura 15. Diagrama de flujo: Fase 4

- [6] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open source scientific tools for Python," 2001.
- [7] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data," *Analytical Chemistry*, vol. 36, no. 8, 1964.
- [8] R. W. Schafer, "What is a Savitzky-Golay filter?," *IEEE Signal Processing Magazine*, vol. 28, no. 4, 2011.
- [9] W. H. Press and S. A. Teukolsky, "Savitzky-Golay smoothing filters," *Computers in Physics*, vol. 4, no. 6, 1990.
- [10] M. Schmid et al., "Why and how Savitzky-Golay filters should be replaced," *ACS Measurement Science Au*, vol. 2, no. 2, 2022.
- [11] H. H. Madden, "Comments on the Savitzky-Golay convolution method," *Analytical Chemistry*, vol. 50, no. 9, 1978.
- [12] J. O. Smith, "Spectral audio signal processing," W3K Publishing, 2011.
- [13] L. R. Rabiner and B. Gold, "Theory and application of digital signal processing," Prentice-Hall, 1975.
- [14] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis," *Proc. IEEE*, vol. 65, 1977.
- [15] M. R. Portnoff, "Time-frequency representation of digital signals," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, 1980.
- [16] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, 1986.
- [17] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. ASSP*, vol. 32, 1984.
- [18] B. Sharpe, "Invertibility of overlap-add processing," accessed July 2019.
- [19] L. R. Rabiner and R. W. Schafer, "Digital processing of speech signals," Prentice Hall, 1978.
- [20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error STSA estimator," *IEEE Trans. ASSP*, vol. 32, 1984.
- [21] N. Wiener, "Extrapolation, interpolation, and smoothing of stationary time series," MIT Press, 1949.
- [22] J. Chen et al., "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, 2006.
- [23] P. C. Loizou, "Speech enhancement: theory and practice," CRC Press, 2013.
- [24] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol. 27, 1979.
- [25] J. Sohn et al., "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, 1999.
- [26] A. J. M. Houtsma, "Pitch and timbre," *Journal of New Music Research*, vol. 26, 1997.
- [27] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms," in *Speech Production*, Springer, 1990.
- [28] P. Ladefoged, "Vowels and consonants," Blackwell Publishers, 2001.
- [29] G. Fant, "Acoustic theory of speech production," Mouton, 1960.
- [30] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Am.*, vol. 24, 1952.
- [31] J. Hillenbrand et al., "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, vol. 97, 1995.
- [32] K. N. Stevens, "Acoustic phonetics," MIT Press, 1998.
- [33] D. H. Whalen and A. G. Levitt, "The universality of intrinsic F0 of vowels," *Journal of Phonetics*, vol. 23, 1995.
- [34] I. R. Titze, "Principles of voice production," National Center for Voice and Speech, 2000.
- [35] M. Hirano, "Clinical examination of voice," Springer, 2013.
- [36] C. E. Silver et al., "Current trends in initial management of laryngeal cancer," *Eur. Arch. Otorhinolaryngol.*, vol. 266, 2009.
- [37] M. Remacle et al., "Endoscopic cordectomy. A proposal for a classification," *Eur. Arch. Otorhinolaryngol.*, vol. 257, 2000.
- [38] E. V. Sjögren et al., "Voice outcome in T1a midcord glottic carcinoma," *Arch. Otolaryngol.-Head Neck Surg.*, vol. 134, 2008.
- [39] T. Yilmaz et al., "Voice after cordectomy type I or type II," *Otolaryngol.-Head Neck Surg.*, vol. 168, 2023.
- [40] L. M. Aaltonen et al., "Voice quality after treatment of early vocal cord cancer," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 90, 2014.
- [41] H. S. Lee et al., "Voice outcome according to surgical extent," *The Laryngoscope*, vol. 126, 2016.
- [42] A. K. Fouad et al., "Laryngeal compensation for voice production," *Clin. Exp. Otorhinolaryngol.*, vol. 8, 2015.
- [43] G. Fant, "Acoustic theory of speech production: with calculations," Mouton, 1960.
- [44] T. Chiba and M. Kajiyama, "The vowel: Its nature and structure," Tokyo-Kaiseikan, 1941.
- [45] K. N. Stevens, "Acoustic phonetics," MIT Press, 1999.
- [46] J. L. Flanagan, "Speech analysis synthesis and perception," Springer, 1972.
- [47] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2733-2749, 2008.
- [48] P. Birkholz, D. Jackèl, and B. J. Kröger, "Construction and control of a three-dimensional vocal tract model," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, IEEE, vol. 1, 2006.
- [49] B. H. Story, "A parametric model of the vocal tract area function for vowel and consonant simulation," *The Journal of the Acoustical Society of America*, vol. 117, no. 5, pp. 3231-3254, 2005.
- [50] W. J. Hardcastle, J. Laver, and F. E. Gibbon, *The Handbook of Phonetic Sciences*, 2nd ed. Oxford: Wiley-Blackwell, 2010.
- [51] I. Goodfellow, Y. Bengio, y A. Courville, *Deep Learning*. MIT Press, 2016.

- [52] J. Wang, K. Chin, y H. Wang, "Speaker-informed speech enhancement and separation,"^{en} *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
- [53] Y. Fathullah *et al.*, "Neural Speech Synthesis using Semantic Tokens,"*arXiv preprint arXiv:2305.xxxx*, 2023.
- [54] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, y A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units,"^{en} *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451-3460, 2021.
- [55] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, y S. Chang, ÇontentVec: An Improved Self-Supervised Speech Representation by Disentangling Speakers.,^{en} *Proc. of the 39th International Conference on Machine Learning (ICML)*, 2022.
- [56] N. Tishby y N. Zaslavsky, "Deep learning and the information bottleneck principle,"^{en} *IEEE Information Theory Workshop (ITW)*, 2015.
- [57] X. Tan *et al.*, ^ A Survey on Neural Speech Synthesis,"*arXiv preprint arXiv:2106.15561*, 2021.
- [58] RVC-Project, Retrieval-based Voice Conversion WebUI,"GitHub repository, 2023. [En línea]. <https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI>
- [59] C. Kavin (svc-develop-team), "So-VITS-SVC: SoftVC VITS Singing Voice Conversion,"GitHub repository, 2023. [En línea]. <https://github.com/svc-develop-team/so-vits-svc>
- [60] E. Gölge *et al.*, Çoqui XTTS: Open-Source Text-to-Speech Model,Çoqui AI, 2023.
- [61] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, y I. Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision,"*OpenAI Technical Report*, 2022.
- [62] J. Kong, J. Kim, y J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,"^{en} *Proc. NeurIPS*, 2020.
- [63] P. Christiano *et al.*, "Deep Reinforcement Learning from Human Feedback,"*Advances in Neural Information Processing Systems*, 2017.
- [64] R. A. Khalil *et al.*, "Speech Emotion Recognition Using Deep Learning Techniques: A Review,"*IEEE Access*, vol. 7, pp. 117327-117345, 2019.