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

Alfonso Gamboa Rubén y Flores Montero Edsel Yetlanezi

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].

Los autores son estudiantes de Ingeniería Biomédica.
Manuscrito generado el 10 de diciembre de 2025.

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]. Ajusta un polinomio de orden k a una ventana de puntos m mediante mínimos cuadrados [9], preservando los formantes [10]. Su formulación discreta es:

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

Donde Y_j es el valor suavizado, y los datos crudos, m el tamaño de la ventana y C_i los coeficientes [11].

II-B. Fundamentos Matemáticos

II-B1. STFT: La voz es una señal no estacionaria, por lo que la Transformada de Fourier clásica es insuficiente [14]. La STFT divide la señal en ventanas temporales [15]:

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

Donde m es el índice temporal, k el índice de frecuencia y H el tamaño del salto [16], [17].

II-B2. Algoritmo de Filtrado: Wiener: El filtro de Wiener calcula una ganancia óptima $W(f)$ basada en la SNR [22], [23]:

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

II-C. Conceptos Estadísticos

II-C1. Desviación Estándar (σ): Un píxel espectral se considera "señal"si supera un umbral dinámico:

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

II-D. Acústica de la Voz

II-D1. Los Formantes: Picos de resonancia espectral [28], [29]:

- **F1 y F2:** Determinan la vocal [30].
- **F3, F4 y F5:** Determinan el timbre e identidad [34], [35].

II-E. Cordectomía y Modelo Fuente-Filtro

La cordectomía reseca las cuerdas vocales por neoplasias [36], [37], causando disfonía [38]. El modelo acústico estándar (Fant, 1960) [43] separa la Fuente (vibración) del Filtro (tracto vocal) [46].

III. Metodología

III-A. Versión 1.0: Análisis Espectral

III-A1. Algoritmo 1.1.0: Preprocesamiento: Sea $x(n)$ la señal de entrada, su representación en frecuencia $X(k)$ se define como:

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

Aquí va mermaid 1

Figura 1. Diagrama de flujo del Algoritmo 1.1.0

III-A2. Algoritmo 1.2.0: Rehabilitación: Se calcula un factor de ganancia $G(k)$:

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

Señal rehabilitada $Y_{reh}(k)$:

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

Aplicación de la IFFT:

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

Aquí va mermaid 2

Figura 2. Diagrama de flujo del Algoritmo 1.2.0

III-A3. Algoritmo 1.1.1: Archivos Independientes: Iteración para procesamiento unilateral.

Aquí va mermaid 3

Figura 3. Diagrama de flujo del Algoritmo 1.1.1

III-A4. Algoritmo 1.2.1: Suma Diferencial: Compensación aditiva basada en promedios espectrales.

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

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

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-A5. Algoritmo 1.2.2: Inyección Proyectada: Control estadístico para evitar distorsión usando μ y σ.

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

III-B. Versión 2.0: Metadatos

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

Aquí va mermaid 6

Figura 6. Diagrama de flujo del Algoritmo 2.1.0

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

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

Aquí va mermaid 7

Figura 7. Diagrama de flujo del Algoritmo 2.2.1.0

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.

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

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

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

Aplicación a la señal post-operatoria:

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

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-D. Versión 4.0: IA (Propuesta)

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

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

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.

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

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.

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

Referencias

[1] T. E. Oliphant, .A guide to NumPy,ÜSA: 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,ïn *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.
[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, Çomments 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, Ïnvertibility 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,ÇRC 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,ïn *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, Çontrol 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, Çlinical examination of voice,"Springer, 2013.
[36] C. E. Silver et al., Çurrent 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. Yılmaz 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.

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

[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.