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

Alfonso Gamboa Rubén

10 de diciembre de 2025

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 lograr preservar la identidad del paciente y evitar artefactos o distorsiones. El estudio concluye proponiendo una versión adicional de experimentación modular que implementa herramientas de inteligencia artificial.

Palabras Clave: 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, Ephraim-Malah, Formantes, Inteligencia Artificial (IA), RLHF, Red Neuronal, Speech Emotion Recognition (SER).

1. Objetivos del Proyecto

1.1. 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.

1.2. 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.

2. Marco teórico

2.1. Software y Herramientas de Desarrollo

2.1.1. Lenguaje de Programación: Python (v3.12 / v3.16)

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].

2.1.2. Librerías Especializadas

- **Numpy (`numpy`):** Fundamental para la manipulación de arreglos multidimensionales [3]. En el contexto del proyecto, se utiliza para convertir los flujos de bits de audio en arreglos de punto flotante (`float32`), permitiendo operaciones de álgebra lineal [4].

- **Pydub (`pydub`):** Librería de alto nivel que actúa como interfaz para el manejo de archivos de audio (I/O).

- **Scipy (`scipy.signal`):** Proporciona las herramientas matemáticas avanzadas para el procesamiento digital de señales [5, 6].

    - *Filtro Savitzky-Golay:* Utilizado para el suavizado de curvas espectrales [7, 8]. A diferencia de un promedio móvil simple, este filtro ajusta un polinomio de orden k a una ventana de puntos m mediante mínimos cuadrados [9], preservando los momentos espectrales importantes (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 de convolución [11].

2.2. Fundamentos Matemáticos y Procesamiento de Señales

2.2.1. Transformada de Fourier de Tiempo Corto (STFT)

Dado que la voz es una señal no estacionaria, la Transformada de Fourier clásica es insuficiente [14]. La STFT divide la señal en ventanas temporales superpuestas [15]. Matemáticamente:

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

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

2.2.2. Algoritmo de Filtrado: Wiener y Ephraim-Malah

El filtro de Wiener calcula una ganancia de transferencia óptima $W(f)$ basada en la Relación Señal-Ruido (SNR) [22, 23]:

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

El algoritmo atenúa las frecuencias donde la potencia del ruido domina [24].

2.3. Conceptos Estadísticos

2.3.1. Desviación Estándar (σ)

Un píxel espectral se considera "señal"solo si su magnitud supera la media del ruido más n veces su desviación estándar:

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

2.4. Acústica y Características de la Voz

2.4.1. Los Formantes

Los picos de resonancia espectral generados por el filtro del tracto vocal se denominan formantes [28, 29]:

- **F1 y F2 (Vocálicos):** Determinan qué vocal se está pronunciando [30, 31, 32]. La inteligibilidad depende de ellos [33].
- **F3, F4 y F5 (Timbre):** Ubicados en frecuencias superiores ($>$ 2500 Hz), dependen de la anatomía fija [34] y definen la identidad del hablante [35].

2.5. La cordectomía y Teoría Fuente-Filtro

La cordectomía es una intervención quirúrgica para neoplasias laríngeas [36, 37], resultando en disfonía o afonía [38, 39, 40, 41, 42]. El modelo acústico estándar es la **Teoría Fuente-Filtro** (Fant, 1960) [43, 44]:

1. **La Fuente (*Source*):** Vibración de los pliegues vocales [45].
2. **El Filtro (*Filter*):** El tracto vocal que actúa como resonador [46, 47].

3. Metodología

3.1. Versión 1.0: Análisis Espectral y Estadística Descriptiva

La arquitectura se estructura en importación/preprocesamiento y procesamiento DSP. Se utiliza un micrófono Razer Seiren Mini.

3.1.1. 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 TU DIAGRAMA MERMAID 1.1.0 COMO IMAGEN]
graph TD
    A[Inicio] --> B[/Cargar Audio Pre A y Post B/]
    B --> C[Aplicar Reducción de Ruido Noisereduce]
    ...
```

Figura 1: Diagrama de flujo del Algoritmo 1.1.0

3.1.2. Algoritmo 1.2.0: Rehabilitación por Amplificación

Se calcula un factor de ganancia $G(k)$:

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

La señal rehabilitada $Y_{reh}(k)$ es:

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

Finalmente se aplica la IFFT:

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

3.2. Versión 2.0: Optimización mediante Metadatos

Se implementa un promedio ponderado para el Perfil Espectral Ideal:

$$S_{ideal}(k) = \frac{\sum_{i=1}^{N}(X_i(k) \cdot w_i)}{\sum_{i=1}^{N} w_i} \tag{9}$$

Reconstrucción híbrida con filtro Savitzky-Golay:

$$Y_{reh}(k) = \text{SavGol}\left(\begin{cases} X_{post}(k) \cdot G_{amp} & \text{si } 500 \leq f \leq 2500 \\ S_{ideal}(k) & \text{si } 3500 \leq f \leq 4500 \\ X_{post}(k) & \text{resto} \end{cases}\right) \tag{10}$$

3.3. Versión 3.0: Estimación Espectral MMSE

Se utiliza la STFT. El espectrograma se calcula como:

$$S(m,k) = 10 \cdot \log_{10}(|X(m,k)|^2) \tag{11}$$

Algoritmo 3.3.0: Rehabilitación por Máscara de Transferencia Espectral (T_{dB}):

$$T_{dB}(k) = \mu_{PRE,dB}(k) - \mu_{POST,dB}(k) \tag{12}$$

$$|Y_{reh}(m,k)| = |X_{post}(m,k)| \cdot 10^{\frac{T_{dB}(k)}{20}} \tag{13}$$

3.4. Versión 4.0: Implementación de Inteligencia Artificial (Propuesta)

Se propone una arquitectura ASR-TTS. Matemáticamente, buscamos maximizar la probabilidad de la onda de salida Y dado el texto T y el vector de identidad S:

$$P(Y|T,S) = \prod_{n} P(y_n|y_{<n}, T, S) \tag{14}$$

4. 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,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: A structure for efficient numerical computation,"*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, pp. 263-283.

[5] P. Virtanen et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python,"*Nature Methods*, vol. 17, no. 3, pp. 261-272, 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 by simplified least squares procedures,"*Analytical Chemistry*, vol. 36, no. 8, pp. 1627-1639, 1964.

[8] R. W. Schafer, "What is a Savitzky-Golay filter? [lecture notes],"*IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 111-117, 2011.

[9] W. H. Press and S. A. Teukolsky, "Savitzky-Golay smoothing filters,"*Computers in Physics*, vol. 4, no. 6, pp. 669-672, 1990.

[10] M. Schmid, D. Rath, and U. Diebold, "Why and how Savitzky–Golay filters should be replaced,"*ACS Measurement Science Au*, vol. 2, no. 2, pp. 185-196, 2022.

[11] H. H. Madden, Çomments on the Savitzky-Golay convolution method for least-squares-fit smoothing and differentiation of digital data,"*Analytical Chemistry*, vol. 50, no. 9, pp. 1383-1386, 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,.Englewood Cliffs, NJ: Prentice-Hall, Inc., 1975.

[14] J. B. Allen and L. R. Rabiner, .ª unified approach to short-time Fourier analysis and synthesis,"*Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558-1564, 1977.

[15] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis,"*IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 55-69, 1980.

[16] M. Dolson, "The phase vocoder: A tutorial,"*Computer Music Journal*, vol. 10, no. 4, pp. 14-27, 1986.

[17] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform,"*IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236-243, 1984.

[18] B. Sharpe, Ïnvertibility of overlap-add processing,"https://gauss256.github.io/blog/cola.html, accessed July 2019.

[19] L. R. Rabiner and R. W. Schafer, "Digital processing of speech signals,.Englewood Cliffs, NJ: Prentice Hall, 1978.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,"*IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109-1121, 1984.

[21] N. Wiener, .Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications,"MIT Press, 1949.

[22] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter,"*IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218-1234, 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 Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

[25] J. Sohn, N. S. Kim, and W. Sung, .A statistical model-based voice activity detection,"*IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, 1999.

[26] A. J. M. Houtsma, "Pitch and timbre: Definition, meaning and use,"*Journal of New Music Research*, vol. 26, no. 2, pp. 104-115, 1997.

[27] H. M. Teager and S. M. Teager, .Evidence for nonlinear sound production mechanisms in the vocal tract,ïn *Speech Production and Speech Modelling*, Springer, 1990, pp. 241-261.

[28] P. Ladefoged, "Vowels and consonants: An introduction to the sounds of languages,"Malden, MA: Blackwell Publishers, 2001.

[29] G. Fant, .Acoustic theory of speech production,"The Hague: Mouton, 1960.

[30] G. E. Peterson and H. L. Barney, Çontrol methods used in a study of the vowels,"*The Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175-184, 1952.

[31] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, .Acoustic characteristics of American English vowels,"*The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3099-3111, 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, no. 3, pp. 349-366, 1995.

[34] I. R. Titze, "Principles of voice production,Ïowa City: National Center for Voice and Speech, 2000.

[35] M. Hirano, Çlinical examination of voice,"Springer Science & Business Media, 2013.

[36] C. E. Silver et al., Çurrent trends in initial management of laryngeal cancer,"*European Archives of Oto-Rhino-Laryngology*, vol. 266, no. 9, pp. 1333-1352, 2009.

[37] M. Remacle et al., .Endoscopic cordectomy. A proposal for a classification,"*European Archives of Oto-Rhino-Laryngology*, vol. 257, no. 4, pp. 227-231, 2000.

[38] E. V. Sjögren et al., "Voice outcome in T1a midcord glottic carcinoma,"*Archives of Otolaryngology–Head & Neck Surgery*, vol. 134, no. 9, pp. 965-972, 2008.

[39] T. Yılmaz et al., "Voice after cordectomy type I or type II or radiation therapy,"*Otolaryngology–Head and Neck Surgery*, vol. 168, no. 3, pp. 559-568, 2023.

[40] L. M. Aaltonen et al., "Voice quality after treatment of early vocal cord cancer,"*International Journal of Radiation Oncology Biology Physics*, vol. 90, no. 2, pp. 255-270, 2014.

[41] H. S. Lee et al., "Voice outcome according to surgical extent of transoral laser microsurgery,"*The Laryngoscope*, vol. 126, no. 9, pp. 2051-2056, 2016.

[42] A. K. Fouad et al., "Laryngeal compensation for voice production after CO2 laser cordectomy,"*Clinical and Experimental Otorhinolaryngology*, vol. 8, no. 4, pp. 340-346, 2015.

[43] G. Fant, .Acoustic theory of speech production: with calculations based on X-ray studies,"The Hague: Mouton, 1960.

[44] T. Chiba and M. Kajiyama, "The vowel: Its nature and structure,"Tokyo-Kaiseikan Publishing Co., 1941.

[45] K. N. Stevens, .Acoustic phonetics,Çurrent Studies in Linguistics Series, vol. 30, MIT Press, 1999.

[46] J. L. Flanagan, "Speech analysis synthesis and perception,"Berlin: Springer-Verlag, 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, Çonstruction and control of a three-dimensional vocal tract model,ïn *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,"*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 et al., "HuBERT: Self-Supervised Speech Representation Learning,.[en] *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451-3460, 2021.

[55] K. Qian et al., ÇontentVec: An Improved Self-Supervised Speech Representation,.[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.

[59] C. Kavin (svc-develop-team), "So-VITS-SVC: SoftVC VITS Singing Voice Conversion,"GitHub repository, 2023.

[60] E. Gölge et al., Çoqui XTTS: Open-Source Text-to-Speech Model,Çoqui AI, 2023.

[61] A. Radford et al., 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.