ISSN 1870-4069

Eliminación de ruido en sonidos cardíacos mediante técnicas de aprendizaje profundo

Cristóbal González Rodríguez[1], Miguel A. Alonso Arévalo[2], Eloísa García Canseco[1]

[1] Universidad Autónoma de Baja California,
Facultad de Ciencias,
México

[2] Centro de Investigación Científica y de Educación Superior de Ensenada,
Departamento de Electrónica y Telecomunicaciones,
División de Física Aplicada,
México

{a351269, eloisa.garcia}@uabc.edu.mx,
aalonso@cicese.edu.mx

Resumen. Las enfermedades cardiovasculares son la principal causa de mortalidad en todo el mundo. La auscultación cardiaca es un método de diagnóstico prometedor; sin embargo, uno de sus principales inconvenientes es que es altamente propensa al ruido durante la grabación del sonido, lo que dificulta el diagnóstico. En este trabajo proponemos un algoritmo de eliminación de ruido para las señales de audio cardíaco. El ruido se elimina en la representación tiempo–frecuencia de la señal. Específicamente, calculamos la transformada de Fourier de tiempo corto (STFT) de la señal de FCG contaminada y entrenamos una red neuronal de tipo U-Net para que reconozca los sonidos cardíacos, ya sean normales o patológicos, del ruido. En nuestras pruebas, el método propuesto muestra un alto desempeño incluso en escenarios altamente desfavorables, ya que puede eliminar el ruido de una señal FCG contaminada con una relación señal a ruido (SNR) de -5 dB con mejoras promedio del orden de $\approx$15 dB.

Palabras clave: Fonocardiograma, transformada de Fourier, red neuronal convolucional, separación de fuentes.

Noise Removal in Heart Sounds Using Deep Learning Techniques

Abstract. Cardiovascular diseases are the leading cause of mortality worldwide. Cardiac auscultation is a promising diagnostic method; however, one of its main drawbacks is that it is highly susceptible to noise during sound recording, which hinders diagnosis. In this study, we propose a noise removal algorithm for cardiac audio signals. The noise is eliminated in the time-frequency representation of the signal. Specifically, we calculate the Short-Time Fourier Transform (STFT)

of the contaminated FCG signal and train a U-Net neural network to recognize cardiac sounds, whether they are normal or pathological, in the presence of noise. In our tests, the proposed method demonstrates high performance even in highly unfavorable scenarios, as it can remove noise from a contaminated FCG signal with a signal-to-noise ratio (SNR) of -5 dB, with average improvements of $\approx$15 dB.

Keywords: Phonocardiogram, Fourier transform, convolutional neural network, source separation.

1. Introducción

La fonocardiografía (FCG) es la representación gráfica de los sonidos producidos por el corazón y tradicionalmente ha generado mucho interés por el potencial que tiene como herramienta para el diagnóstico clínico. La señal de FCG proporciona información sobre la duración, la frecuencia y otros parámetros importantes de los sonidos cardíacos para determinar la funcionalidad y el estado actual de las válvulas cardíacas [14].

La identificación de síntomas patológicos mediante la auscultación de sonidos cardíacos con la ayuda de un estetoscopio es una gran habilidad y adquirirla es una tarea difícil que puede llevar muchos años de práctica clínica. Además, el oído humano tiene limitaciones fisiológicas para percibir completamente los sonidos producidos por el corazón ya que la mayor parte de la energía del FCG se encuentra por debajo del umbral de audición [7].

Las enfermedades cardiovasculares son la primera causa de mortalidad en México y en el mundo [18, 24]. Aunque existen muchas técnicas modernas de diagnóstico, como el electrocardiograma (ECG), la resonancia magnética (RM) o el ecocardiograma, la auscultación cardiaca es probablemente el método no invasivo más económico, práctico y rápido. Los recientes avances en informática médica en combinación con la cada vez mayor capacidad de procesamiento de los dispositivos electrónicos han impulsado el análisis automático de las señales de sonido cardíaco.

El objetivo principal de este análisis es clasificar con precisión la presencia o ausencia de sonidos patológicos en el ciclo cardíaco [14]. Si se confirma la presencia de un evento de este tipo, lo ideal sería que un sistema automatizado también sea capaz de identificar el tipo de patología. La presencia de ruido en la señal de FCG es uno de los problemas más frecuentes durante la auscultación cardiaca. Este problema es particularmente delicado para un sistema automático ya que no es capaz de discriminar entre verdaderos sonidos cardíacos y espurios.

Las fuentes de ruido pueden ser numerosas y de naturaleza muy distinta: sonidos originados en el entorno, como el habla o el uso de aparatos cercanos al sitio de auscultación; por sonidos fisiológicos, como los gástricos y respiratorios; o por la fricción producida entre el estetoscopio y la piel.

Bajo estos escenarios, realizar un diagnóstico de enfermedades resulta especialmente difícil. La mayoría de los métodos de eliminación de ruido del FCG presentados en la literatura se evalúan contaminando una señal de FCG con ruido blanco Gaussiano aditivo (AWGN por sus siglas en inglés).

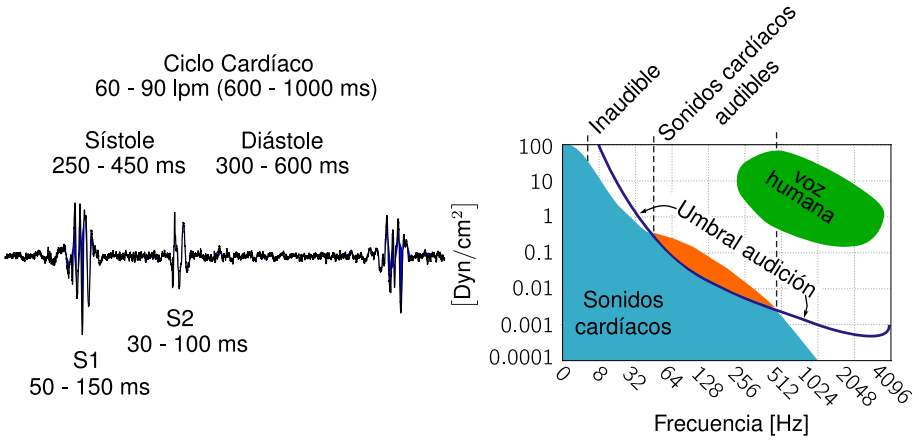

Fig. 1. Características principales de la señal de fonocardiograma (FCG).

Algunos otros trabajos también consideran la contaminación de la señal utilizando ruido rosa Gaussiano aditivo. En ambos casos, una vez eliminado el ruido de la señal de FCG contaminada, se procede a compararla con la señal original.

Tradicionalmente, la métrica más utilizada en la literatura para medir el rendimiento de un sistema de eliminación de ruido es la relación señal a ruido (SNR por sus siglas en inglés). Sin embargo la SNR tiene el problema de ser altamente susceptible a problemas de escalamiento debido a la energía de la señal.

Dicho escalamiento puede aumentar artificialmente la SNR resultante, abriendo así la posibilidad de obtener resultados engañosos que no representen adecuadamente el verdadero rendimiento de un método de eliminación de ruido. Un análisis a profundidad sobre este tema y posibles soluciones se presenta en [13].

1.1. Antecedentes

Una señal de FCG normal consta de dos sonidos cardíacos fundamentales llamados S1 y S2. Estos sonidos se producen durante el ciclo cardíaco cuando las válvulas semilunar y auriculoventricular se cierran [2]. La sístole (contracción), que marca el inicio de S1, y la diástole (relajación), que marca el inicio de S2, comprenden el comportamiento cuasi periódico del ciclo cardíaco. Existen dos periodos de silencio en los individuos sanos, ya que estos sonidos sólo comprenden una pequeña parte de cada sección del ciclo cardíaco: los silencios sistólico (s-Sys) y diastólico (s-Dia).

Lo más frecuente es que los ciclos cardíacos de individuos con afecciones cardíacas contengan ruidos adicionales, como sonidos S3 y S4, soplos, fricciones o chasquidos. Los sonidos cardíacos S1 y S2 suelen durar entre 25 y 150 ms, y la mayor parte de su contenido espectral se sitúa entre 24 y 144 Hz [5]. La duración de las patologías (chasquidos, fricciones y soplos) varía significativamente a lo largo del ciclo cardíaco, y su contenido espectral se sitúa entre 25 Hz y 700 Hz [6]. En la Figura 1 se resumen las características principales de la señal de FCG y en la Figura 2 se presentan dos ejemplos correspondientes a a) un sonido cardíaco normal y b) un sonido cardíaco anormal.

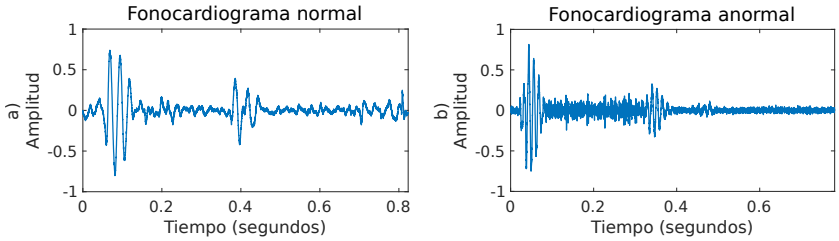

Fig. 2. Dos ejemplos de sonidos cardíacos, a) presenta el sonido de un corazón normal, mientras que b) presenta un sonido correspondiente a una persona con una patología cardíaca.

Se ha buscado eliminar el ruido en de la señal de FCG abordando el problema de múltiples maneras. Uno de los enfoques más populares es el uso de la transformada ondeleta discreta (DWT por sus siglas en inglés), el cual consiste en descomponer la señal en coeficientes ondeleta. A lo largo del tiempo se han propuesto variantes y mejoras de este método [11, 3, 10, 9, 17].

Este enfoque parte de la idea de que el ruido puede reordenarse en coeficientes en distintas bandas de frecuencia y que es posible aislar el sonido FCG mediante el uso de umbrales que eliminen los coeficientes correspondientes al ruido [16]. Se han realizado varias mejoras a este método básico, como probar diferentes ondeletas madre así como niveles de descomposición [16, 11, 3, 9]; evaluando los modelos propuestos con diferentes tipos de ruido [11, 10].

En otros casos, implementando redes neuronales para reconstruir la señal con un umbral adaptativo según los coeficientes obtenidos [10]. Más recientemente hay variantes que utilizan la transformada ondeleta síncrona [8] (SSWT por sus siglas en inglés) como la propuesta de [17]. Otro método reciente se basa en el algoritmo de mínimos cuadrados (LMS por sus siglas en ingés) [19]. Este enfoque utiliza un filtro adaptativo y consigue eliminar la mayor parte del ruido introducido en la señal.

Algunos ejemplos mencionados anteriormente implementan redes neuronales para mejorar el rendimiento de sus algoritmos [10], pero la aplicación de la inteligencia artificial en la eliminación del ruido presente en la señal de FCG aún puede ser explorada más a fondo, como se muestra a continuación.

1.2. Enfoque propuesto

La inteligencia artificial (IA) es una herramienta moderna que ha dado grandes resultados en numerosos ámbitos. Tal vez una de las aplicaciones más sobresalientes de la IA ha sido en el procesamiento de imágenes. En particular, la implementación de redes neuronales convolucionales (CNN por sus siglas en inglés) ha mostrado ventajas significativas sobre los métodos tradicionales para la eliminación de ruido de imágenes [25]. Algunas técnicas desarrolladas para imágenes han sido adaptadas para procesar señales de audio.

Un ejemplo notable de una arquitectura de CNN fue propuesta en [20], y se denomina U-Net. Este método fue originalmente propuesto para la segmentación de imágenes biomédicas, sin embargo variantes de la arquitectura U-Net han sido exitosamente utilizadas en la separación de fuentes de sonido.

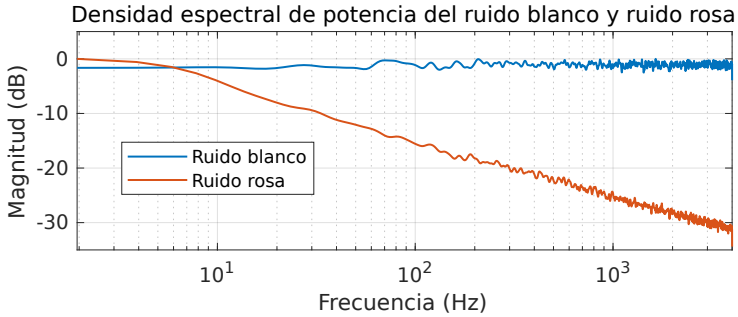

Fig. 3. Densidad espectral de potencia normalizada de los dos tipos de ruido utilizados en este trabajo.

Específicamente, en desagregar una señal grabada mediante un solo micrófono en las múltiples fuentes que la conforman, como se muestra en [4, 12]. En estos dos trabajos, las voces y los diferentes instrumentos de una canción se separan en diferentes formas de onda. Estos métodos funcionan obteniendo primero los espectrogramas del audio original mediante la transformada de Fourier de tiempo corto (STFT por sus siglas en inglés).

Los espectrogramas se utilizan después para entrenar una U-Net que identifique un tipo específico de sonido; esto se hace procesando los espectrogramas para crear una máscara. La máscara contendrá la información sobre qué partes del espectrograma corresponden a la señal que se desea aislar y cuáles al ruido; multiplicando la máscara por el espectrograma inicial se eliminará el ruido.

Dado que la U-Net está entrenada para identificar un tipo específico de sonido, el tipo de ruido con el que está contaminada la señal es irrelevante para la separación. En el presente trabajo mostramos cómo adaptar esta técnica de separación de fuentes al contexto de eliminación del ruido aditivo Gaussiano y rosa en la señal de FCG.

2. Metodología

Esta sección describe la metodología para la eliminación de ruido en señales de FCG y está divida en tres partes. Primeramente se describe la base de datos de sonidos cardíacos utilizados y los tipos de ruido utilizados para contaminarlas. En la segunda parte se describe brevemente el algoritmo de la Transformada de Fourier de Tiempo Corto (STFT por sus siglas en inglés) mediante el cual se convierten los sonidos en imágenes. En la tercera parte se describe la arquitectura de la U-Net y el proceso de entrenamiento de la red.

2.1. Base de datos y tipos de ruido

Como fuente de sonidos cardíacos se utilizó la base de datos propuesta en [22]. Esta base de datos pública es altamente popular en el área de análisis del FCG y se caracteriza por tener señales de sonidos cardíaco particularmente limpias. El uso de señales de sonido limpias permite un mejor entrenamiento de la red.

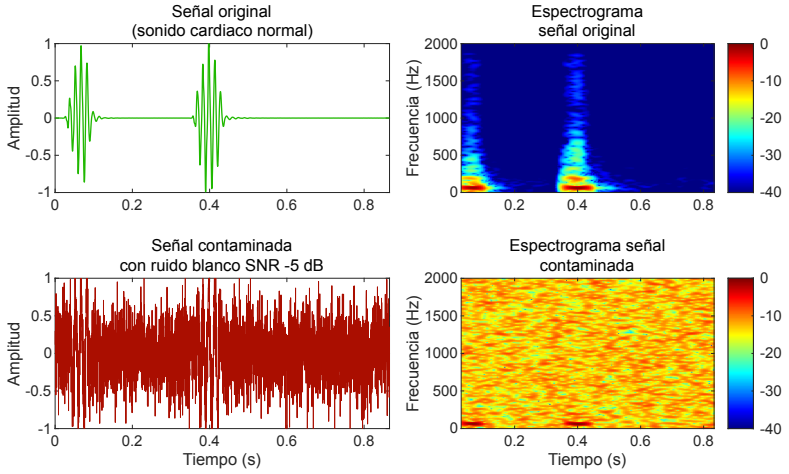

Fig. 4. Ejemplo de las imágenes de espectrograma que se utilizan para el entrenamiento de la red.

Este conjunto de datos tiene cinco categorías de señales: señales con sonidos de FCG normales (N) y cuatro señales con sonidos de FCG patológicos, que incluyen estenosis aórtica (EA), regurgitación mitral (RM), estenosis mitral (EM) y soplo en sístole (MVP). Estas cinco categorías contienen 200 grabaciones cada una; todas las señales se componen de tres ciclos cardíacos completos con una duración del FCG de aproximadamente tres segundos.

En total, la base de datos contiene 1,000 sonidos cardíacos muestreados a 8 kHz. En procesamiento de señales, es relativamente común utilizar ruido blanco aditivo como modelo para imitar el efecto de muchos procesos aleatorios que se dan en la naturaleza y que afectan el rendimiento de un sistema.

El concepto de ruido blanco se refiere a la idea de que tiene una potencia uniforme en toda la banda de frecuencias del sistema de información. En este trabajo se propone como primer tipo de señal de interferencia el uso de ruido blanco Gaussiano aditivo, porque tiene una distribución normal en el dominio del tiempo con una media de cero.

El segundo tipo de señal de interferencia utilizado en este trabajo es el ruido rosa. Este tipo de ruido es útil para aplicaciones de sonido y sistemas de audio ya que muchos sonidos musicales y naturales tienen espectros que contienen todas las frecuencias audibles, pero que disminuyen en intensidad a una razón de $\approx$3 dB por octava en función de la frecuencia (f) siguiendo un comportamiento de la forma $1/f$. La Figura 3 muestra la densidad espectral de potencia normalizada de los dos tipos de ruido utilizados en este trabajo.

2.2. Transformada de Fourier de tiempo corto

Para entrenar la red neuronal lo mejor es proporcionar tanta información sobre el comportamiento de la señal como sea posible. En [4, 12], los autores sugieren utilizar la transformada de Fourier de tiempo corto (STFT) para generar los espectrogramas de las señales de audio y utilizarlos como imágenes para entrenar el modelo.

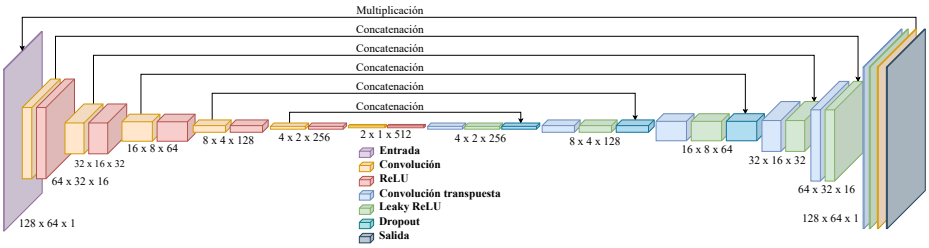

Fig. 5. Arquitectura de la U-Net con las dimensiones obtenidas a través de cada capa.

Este enfoque ha demostrado un excelente rendimiento en problemas de separación de fuentes, ya que contiene una representación precisa de la información temporal y frecuencial de la señal; en el presente trabajo, adoptamos el mismo enfoque.

Existen representaciones tiempo-frecuencia más precisas que la STFT. Sin embargo, en la práctica la STFT tiene una alta popularidad por múltiples razones, entre ellas su sencillez, velocidad de cálculo y un rendimiento altamente satisfactorio. En la práctica, las grabaciones de FCG que analizamos son de tiempo discreto, y los datos a transformar, que se muestrean en el tiempo, se procesan mediante la STFT discreta.

Un grupo de muestras forma un segmento o bloque de tiempo, y para cada segmento se calcula la transformada discreta de Fourier (DFT). El resultado es un vector complejo que se añade como columna a una matriz, que registra la magnitud y la fase de todos los segmentos de tiempo y bins de frecuencia. La STFT discreta de una señal s se calcula como sigue [21]:

$$X\left(k,m\right)=\sum_{n=-\infty}^{\infty}x\left(n\right)w\left(n-mH\right)e^{-\imath 2\pi nk/N}, \tag{1}$$

donde m representa cada marco temporal a lo largo de la señal de la que se calcula la STFT, y como se ha mencionado anteriormente, para cada marco temporal m se calcula localmente la transformada discreta de Fourier (DFT por sus siglas en inglés) a través de una señal $x\left(n\right)$, obteniendo un vector de frecuencias $X\left(k\right)$.

El conjunto de intervalos de frecuencia se define como $f_k = kf_s/N$, para $k = 0,1,2,\ldots,N/2$, siendo f_s la frecuencia de muestreo de la señal y N el número de muestras temporales en la DFT; n representa cada muestra a través de la señal local $x\left(n\right)$; H la longitud de salto entre dos segmentos temporales consecutivos m y $m+1$ que delimitan $x\left(n\right)$, con $m = 0,1,2,\ldots,M-1$, donde M es el número total de segmentos temporales en que se divide la señal original. Cada DFT es ponderada con la función de ventana de Hann:

$$w\left(n'\right)=0.5-0.5\cos\left(\frac{2\pi n'}{L_w-1}\right),\ 0\leq n'\leq L_w-1. \tag{2}$$

De longitud L_w muestras. La forma de la matriz resultante $X\left(k,m\right)$ sería entonces de dimensiones $K\times M$, con $K = N/2+1$ y $M = \lceil\left(S+R+1\right)/H\rceil$, donde $R=\left(-\left(S-L_w\right)\ \%H\right)\ \%L_w$, $\lceil\cdot\rceil$ representa la función techo (ceil en inglés) y $\%$ el residuo de la división; siendo $S = Tf_s$ el número total de muestras que conforman la señal original.

Tabla 1. Desempeño promedio del algoritmo de eliminación de ruido. La métrica utilizada es la SI-SDR para una validación cruzada de 10 iteraciones, los tipos de ruido considerados son blanco y rosa.

Ruido de entrenamiento	Tipo/Nivel de ruido (dB)	-5	0	5	10
Blanco	**Blanco**	9.93	13.41	17.21	20.71
	Rosa	-0.57	4.38	9.20	13.21
Rosa	**Blanco**	1.76	9.37	15.29	19.15
	Rosa	7.28	11.28	14.86	18.19

Los espectrogramas de las señales limpias y contaminadas fueron necesarios para entrenar la red. Además, para ajustar las imágenes (espectrogramas) a las dimensiones de entrada de la red, pero también para reducir la carga computacional de procesar miles de espectrogramas, hubo que ajustar su tamaño, eligiendo una FFT de tamaño de 256 muestras, longitud de salto de 32 muestras y una longitud de ventana de 128 muestras.

Utilizando estos parámetros obtuvimos los espectrogramas que se muestran en la Figura 4. Las gráficas presentadas en dicha figura también ilustran el tipo de imágenes PCG ruidosas que se alimentan a la red neuronal, en este ejemplo en particular, con señales contaminadas a -5 dB de relación señal a ruido (SNR por sus siglas en inglés).

Durante el entrenamiento de la red, solamente se utiliza la magnitud de la STFT. La fase de cada espectrograma obtenida durante el cálculo de la STFT se almacena para su reconstrucción posterior. A continuación, se procede a procesar los espectrogramas con la red entrenada para obtener la máscara. Luego se multiplican elemento por elemento con los espectrogramas originales. Finalmente se utiliza la fase para reconstruir el audio mediante la transformada de Fourier inversa.

2.3. Arquitectura de la U-Net

En este trabajo, utilizamos una red neuronal convolucional con una arquitectura U-Net similar a la que se propuso originalmente en [12]. La Figura 5 ilustra la arquitectura U-Net que utilizamos. Las imágenes de entrada y salida de la red tienen un tamaño de $(128, 64, 1)$.

Cada capa de convolución 2D y de transposición 2D utiliza un tamaño de núcleo de 5×5 y pasos de 2×2, rellenando con ceros después de cada convolución. Para la capa Leaky ReLU, se utilizó un parámetro $\alpha = 0.2$. Antes de multiplicar la máscara de salida por la entrada, sigue una última capa de convolución 2D con un tamaño de núcleo de 4×4, una tasa de dilatación de (2,2) y una función de activación de sigmoide.

Para entrenar la red, es necesario un conjunto de imágenes de entrada lo más amplio y diverso posible. En nuestro caso, las imágenes utilizadas son las obtenidas a partir del valor absoluto de la STFT (i.e., el espectrograma) de las señales de FCG de tamaño $K \times M$.

En procesamiento de imágenes, cuando se consideran matrices multidimensionales, las imágenes suelen tener un parámetro de tercera dimensión que representa el número de canales de color de la imagen. El valor típico es tres cuando nos referimos a imágenes de color verdadero o RGB (rojo, verde y azul).

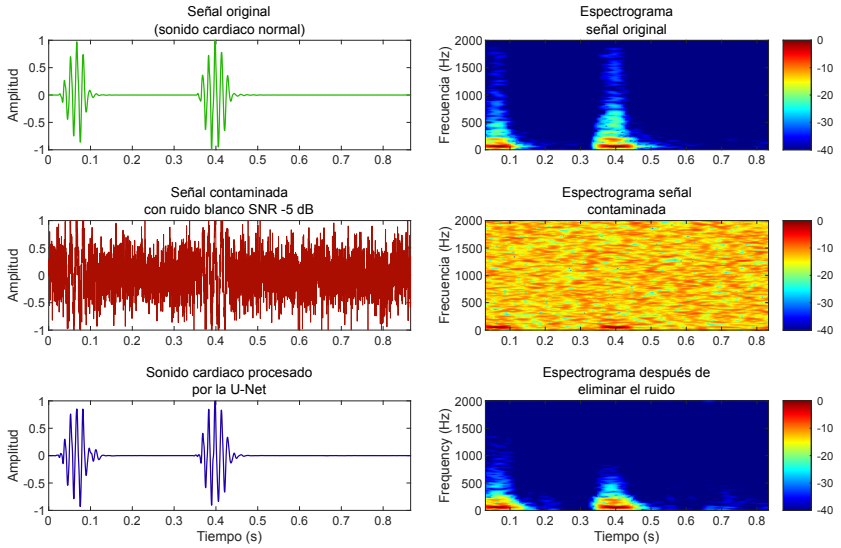

Fig. 6. Ejemplo del funcionamiento de la red para limpiar una señal de FCG normal contaminada con ruido blanco a -5 dB de SNR.

Sin embargo, como estamos utilizando espectrogramas compuestos por una matriz bidimensional, podemos considerarlos como si fueran imágenes con un solo canal de color; es decir, el eje de tercera dimensión tiene una longitud de uno.

La forma de cada imagen de espectrograma se fijó entonces de la siguiente manera $(128, 64, 1)$, como se muestra en la Figura 5. La red original propuesta en [12] se implementó considerando imágenes más grandes generadas por señales de audio de alta fidelidad (Hi-Fi).

Sin embargo, nuestra versión modificada también muestra un rendimiento notable para la eliminación de ruido en las señales de FCG que tienen un contenido espectral de tipo de pasa-bajas. Dado que las dimensiones de entrada de la red son 128×64, cada espectrograma se dividió en bloques de tiempo que coinciden con las dimensiones de entrada de la red.

De los mil sonidos de FCG, 900 audios fueron seleccionados aleatoriamente a partir de las de las cinco clases (una de FCG normal y cuatro tipos de FCG patológicos) para el entrenamiento de la red y se generaron más de cinco mil bloques de espectrograma utilizados únicamente para el entrenamiento. Internamente, las bibliotecas de inteligencia artificial que utilizamos dividieron las mil señales de la siguiente manera:

El 72 % de las señales se utilizaron para el entrenamiento de la red, el 18 % para la validación durante el entrenamiento, y el último 10 % se utilizó para la evaluación final de prueba de la red entrenada, ya que se trataba de datos que no había sido previamente vistos por la red.

Seleccionamos ADAM como algoritmo de optimización, y los parámetros de la red que mostraron el mejor rendimiento de convergencia para la eliminación de ruido fueron una tasa de aprendizaje de 0,0001 y un tamaño de lote de 64.

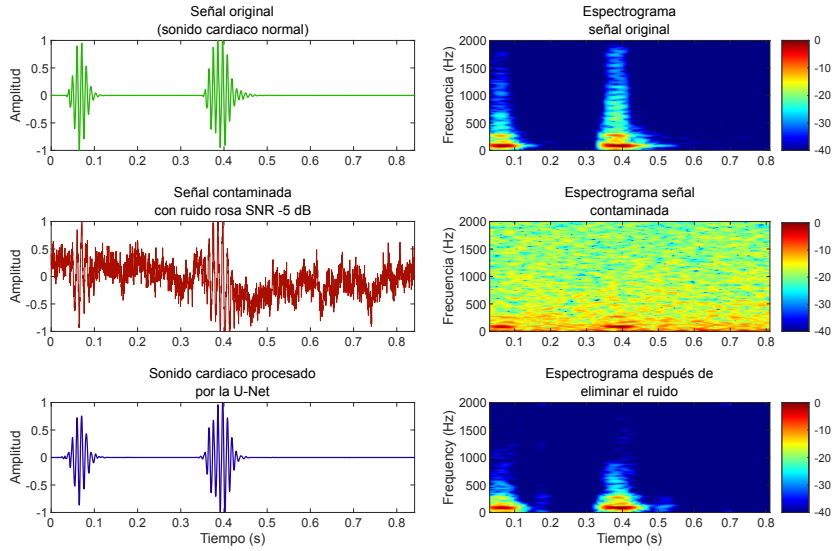

Fig. 7. Ejemplo del funcionamiento de la red para limpiar una señal de FCG normal contaminada con ruido rosa a -5 dB de SNR.

3. Resultados

El desempeño de la red propuesta fue evaluado mediante una validación cruzada de 10 iteraciones, es decir se entrenaron diez redes para cada uno de los dos tipos de ruido (blanco y rosa). Las redes entrenadas se evaluaron contaminando las señales del conjunto de prueba con una relación señal-ruido (SNR) de -5, 0, 5 y 10 dB. Estos valores de SNR se seleccionaron ya que cubren el intervalo en el que es más necesaria la eliminación de ruido.

Las redes se entrenaron para los cuatro niveles de SNR, pero en nuestras pruebas observamos que el entrenamiento a 0 dB SNR muestra los mejores resultados para cualquier nivel de ruido. Durante la evaluación, cada conjunto de prueba contenía un total de 100 señales, 20 para cada uno de los sonidos cardíacos disponibles en la base de datos [22]. Los programas para llevar a cabo el algoritmo propuesto fueron desarrollados en Python. Para implementar la U-Net se utilizaron las librerías de software TensorFlow y Keras [1].

Los bloques de procesamiento de la señal se implementaron utilizando SciPy [23] y Librosa [15]. La métrica para medir el desempeño de la red es una versión ligeramente modificada de la SNR, la cual fue originalmente propuesta en [13]. Esta métrica da lugar a una medida más sencilla y robusta denominada relación señal a distorsión invariante a la escala (SI-SDR por sus siglas en inglés) y que está definida como:

$$\text{SI-SDR} = 10\log_{10}\left(\frac{\|\frac{\hat{s}^T s}{\|s\|^2}s\|^2}{\|\frac{\hat{s}^T s}{\|s\|^2}s-\hat{s}\|^2}\right), \tag{3}$$

donde s es la señal objetivo (original) y $\hat{s}$ es la señal obtenida a la salida del algoritmo de eliminación de ruido. Esta redefinición de la SNR tiene en cuenta un posible reescalado de la señal y es invariante frente a variaciones de amplitud, evitando así resultados engañosos que pueden obtenerse utilizando únicamente la SNR [13].

La Tabla 1 muestra los resultados obtenidos usando la metodología propuesta en la sección anterior. La capacidad de la red para eliminar el ruido en escenarios difíciles es bastante notable, brindando mejoras de $\approx$15 dB para señales contaminadas con ruido blanco a -5dB de SNR y de $\approx$12 dB para señales contaminadas con ruido rosa.

Como se esperaba, los resultados muestran que la red elimina mejor el tipo de ruido para el que fue entrenada. Sin embargo, el desempeño es más alto para ruido blanco que para ruido rosa. Este fenómeno también tiene sentido, ya que al ser la señal de FCG de tipo pasa-bajas, el ruido rosa (que tiene mucho mayor energía en las bajas frecuencias que en las altas frecuencias) tiene un efecto más importante en la disminución de la calidad de la señal.

Conforme el nivel de ruido presente en la señal disminuye, también disminuye ligeramente el desempeño de la red neuronal. Este mismo comportamiento es visible para ambos tipos de ruido. Las Figuras 6 y 7 muestran dos ejemplos particulares del desempeño de la red para limpiar señales contaminadas con un nivel de -5 dB y utilizando ruido blanco y rosa respectivamente. Del lado izquierdo de la figura se presentan las formas de onda en el dominio del tiempo y del lado derecho los respectivos espectrogramas.

La forma de onda de la señal original es prácticamente indistinguible en la versión contaminada de la gráfica (segunda fila). En las gráficas de la tercera fila es posible apreciar el gran trabajo de limpieza que realiza el algoritmo propuesto. En los espectrogramas es posible apreciar que es en las altas frecuencias ($>$500 Hz) donde la red tiene más dificultades para reconstruir la señal original.

4. Conclusiones

Las enfermedades cardiovasculares son una de las principales causas de mortalidad en todo el mundo; su detección precoz es fundamental para mejorar los resultados en materia de salud a largo plazo. El análisis automático de los sonidos cardíaco es un método de diagnóstico prometedor, pero es altamente susceptible al ruido durante la grabación de audio.

En este trabajo propusimos una metodología robusta para la eliminación de ruido en sonidos cardíacos, la cual está basada en el análisis de tiempo-frecuencia (mediante la transformada de Fourier de tiempo corto) y de una arquitectura de red neuronal de tipo U-Net. La metodología propuesta fue evaluada usando una base de datos pública que contiene mil sonidos cardíacos, incluyendo sonidos normales y patológicos.

Llevamos a cabo una evaluación exhaustiva del algoritmo propuesto para distintos valores de relación señal a ruido (SNR), que van desde calidad de sonido altamente desagradable (-5 dB) hasta niveles de calidad de audio aceptables (10 dB). La metodología propuesta presenta un alto desempeño ya que puede eliminar el ruido de una señal FCG contaminada a -5 dB de SNR con mejoras promedio del orden de $\approx$15 dB en el caso de ruido blanco y de $\approx$12 dB para ruido rosado.

Consideramos que el método propuesto tiene un gran potencial para mejorar significativamente el rendimiento de los algoritmos de clasificación automática de sonidos cardíacos en entornos ruidosos, pero también podría utilizarse en estetoscopios electrónicos. Versiones posteriores de este trabajo deberán enfocarse en entrenar y evaluar el desempeño de la red utilizando otras fuentes de ruido.

Específicamente, es importante entrenar la red utilizando fuentes de ruido adquiridas en condiciones reales de auscultación, tales como ruidos ambientales y señales de voz. También sería altamente deseable incrementar el número de audio limpios para entrenamiento y prueba de la red.

Referencias

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2016) doi: 10.48550/ARXIV.1603.04467
2. Abbas, A. K., Bassam, R.: Phonocardiography signal processing. Synthesis Lectures on Biomedical Engineering, vol. 4, no. 1, pp. 1–194 (2009) doi: 10.1007/978-3-031-01637-0
3. Ali, M. N., El-Dahshan, E. S. A., Yahia, A. H.: Denoising of heart sound signals using discrete wavelet transform. Circuits, Systems, and Signal Processing, vol. 36, no. 11, pp. 4482–4497 (2017) doi: 10.1007/s00034-017-0524-7
4. Andreas, J., Eric, H., Nicola, M., Rachel, B., Aparna, K., Tillman, W.: Singing voice separation with deep U-Net convolutional networks. In: 18th International Society for Music Information Retrieval Conference, pp. 23–27 (2017)
5. Arnott, P., Pfeiffer, G., Tavel, M.: Spectral analysis of heart sounds: Relationships between some physical characteristics and frequency spectra of first and second heart sounds in normals and hypertensives. Journal of Biomedical Engineering, vol. 6, no. 2, pp. 121–128 (1984) doi: 10.1016/0141-5425(84)90054-2
6. Choi, S., Jiang, Z.: Cardiac sound murmurs classification with autoregressive spectral analysis and multi-support vector machine technique. Computers in Biology and Medicine, vol. 40, no. 1, pp. 8–20 (2010) doi: 10.1016/j.compbiomed.2009.10.003
7. Cruz-Gutiérrez, A.: Segmentación robusta de audio cardíaco mediante análisis tiempo-frecuencia y métodos de optimización. Master's thesis, Centro de Investigación Científica y de Educación Superior de Ensenada (2016)
8. Daubechies, I., Lu, J., Wu, H. T.: Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. Applied and Computational Harmonic Analysis, vol. 30, no. 2, pp. 243–261 (2011) doi: 10.1016/j.acha.2010.08.002
9. Ghosh, S. K., Tripathy, R. K., Ponnalagu, R.: Evaluation of performance metrics and denoising of PCG signal using wavelet based decomposition. In: IEEE 17th India Council International Conference, pp. 1–6 (2020) doi: 10.1109/INDICON49873.2020.9342464
10. Gradolewski, D., Magenes, G., Johansson, S., Kulesza, W. J.: A wavelet transform-based neural network denoising algorithm for mobile phonocardiography. Sensors, vol. 19, no. 4, pp. 957 (2019) doi: 10.3390/s19040957
11. Gradolewski, D., Redlarski, G.: Wavelet-based denoising method for real phonocardiography signal recorded by mobile devices in noisy environment. Computers in Biology and Medicine, vol. 52, pp. 119–129 (2014) doi: 10.1016/j.compbiomed.2014.06.011
12. Hennequin, R., Khlif, A., Voituret, F., Moussallam, M.: Spleeter: A fast and efficient music source separation tool with pre-trained models. Journal of Open Source Software, vol. 5, no. 50, pp. 2154 (2020) doi: 10.21105/joss.02154

13. Le-Roux, J., Wisdom, S., Erdogan, H., Hershey, J. R.: SDR–half-baked or well done? In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 626–630 (2019) doi: 10.48550/arXiv.1811.02508
14. Mahnke, C. B.: Automated heart sound analysis/computer-aided auscultation: A cardiologist's perspective and suggestions for future development. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3115–3118 (2009) doi: 10.1109/IEMBS.2009.5332551
15. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., Nieto, O.: Librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, vol. 8, pp. 18–25 (2015) doi: 10.25080/Majora-7b98e3ed-003
16. Messer, S. R., Agzarian, J., Abbott, D.: Optimal wavelet denoising for phonocardiograms. Microelectronics Journal, vol. 32, no. 12, pp. 931–941 (2001) doi: 10.1016/S0026-2692(01)00095-7
17. Mohan, N., Kumar, S., Soman, K.: Group sparsity assisted synchrosqueezing approach for phonocardiogram signal denoising. In: 11th International Conference on Computing, Communication and Networking Technologies, pp. 1–5 (2020) doi: 10.1109/ICCCNT49239.2020.9225320
18. Organisation for economic cooperation and development: Obesity update (2017) www.oecd.org/health/health-systems/Obesity-Update-2017.pdf
19. Pauline, S. H., Dhanalakshmi, S.: A robust low-cost adaptive filtering technique for phonocardiogram signal denoising. Signal Processing, vol. 201, pp. 108688 (2022) doi: 10.1016/j.sigpro.2022.108688
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241 (2015) doi: 10.48550/arXiv.1505.04597
21. Smith, J. O.: Spectral audio signal processing, W3K (2011)
22. Son, G. Y., Kwon, S.: Classification of heart sound signal using multiple features. Applied Sciences, vol. 8, no. 12, pp. 2344 (2018) doi: 10.1016/j.procs.2015.08.045
23. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., et al.: SciPy 1.0: Fundamental algorithms for scientific computing in python. Nature Methods, vol. 17, pp. 261–272 (2020) doi: 10.1038/s41592-019-0686-2
24. World health organization: World health statistics 2021: Monitoring health for the SDGs, sustainable development goals (2021) https://www.who.int/data/gho/publications/world-health-statistics
25. Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. Advances in Neural Information Processing Systems, vol. 25 (2012)