# Diseño de Sistema de Captura de Voz Direccional con Filtrado de Ruido para Reconocimiento de Voz

Renzo Tassara
renzotassara98@gmail.com

Control y Sistemas, Facultad de Ingeniería,
Universidad Nacional de Cuyo,
Mendoza, Argentina

Diciembre de 2024, versión v002

**Resumen**

En este informe se aborda la solución al problema de optimización de micrófonos integrados en parlantes con inteligencia artificial, como Alexa o Siri, para mejorar la interpretación de las consignas emitidas por los usuarios. La solución propuesta utiliza ambos micrófonos para cancelar el ruido ambiente, se optimiza la ganancia de cada uno y, posteriormente, se realiza un postprocesamiento de la señal mediante un filtro pasa bajos FIR (Finite Impulse Response).

## 1. Introducción

La utilización de parlantes con inteligencia artifial (Ejemplo fig. 1) es cada día más frecuente, por lo que se busca que la interpretación de las consignas del usuario sean precisas para que la inteligencia artificial pueda responder según lo interpretado. Es por eso que se busca diseñar lo mejor posible el hardware y software del parlante para que el procesamiento de señal sea lo óptimo.

Figura 1: Parlante Amazon Echo Dot 4° generación

Estos dispositivos suelen incorporar múltiples micrófonos, que se orientan hacia el hablante y atenúan el ruido ambiental presente en su entorno. Sin embargo, esta configuración no siempre es suficiente para obtener una señal limpia, debido a que las membranas de los micrófonos suelen tener un bajo amortiguamiento. Como resultado, es común que las membranas entren en resonancia, amplificando las frecuencias altas y deteriorando la calidad de la señal. Por esta razón, además de las mejoras en hardware, es necesario implementar un postprocesamiento de la señal, aplicando filtros específicos, como el filtro pasa bajos FIR, que permiten reducir las interferencias y mejorar la interpretación de los comandos por parte del sistema.

# 2. Desarrollo

En esta sección se detallan los pasos realizados para obtener como resultado final una señal más "limpia" del habla de la persona. El procesamiento de la señal sigue la estructura mostrada a continuación:

1. Generación de sonido y movimiento de la persona.
2. Sistema del micrófono.
3. Control.
4. Filtrado.

## 2.1. Generación de sonido y movimiento de la persona

Para la generación de sonido, se descargaron dos archivos formato ".mp3", uno de los cuales contiene el habla de una persona (discurso de graduación del tenista Roger Federer en Dartmouth College[1]), y el otro audio incluye ruido ambiente (sonido simulando el interior de una cafetería[2]). Ambos sonidos se suman para generar una sola señal resultante que ingresa a cada micrófono.

Además, se simuló el movimiento de una persona caminando alrededor de uno de los micrófonos. Para esto, se utilizó el bloque MATLAB Function en Simulink, cuya entrada es un bloque Clock, que simula el transcurso del tiempo, junto con un bloque retenedor de orden cero (ZOH), para retener el valor cada 0.1 s. Para realizarlo, se tuvieron que definir las constantes que muestra la tabla 1:

Tabla 1: Constantes caminata

| | Valor | Unidad | Descripción |
|---|---|---|---|
| $r_1$ | 0.5 | m | Distancia de la persona al primer micrófono |
| d | 0.08 | m | Distancia entre micrófonos |
| $v_c$ | 0.66 | $\frac{m}{s}$ | Velocidad tangencial de la persona |
| $\alpha_0$ | $\pi$ | rad | Ángulo inicial que se encuentra la persona respecto del eje que pasa por ambos parlantes |

Este bloque calcula la distancia del hablante al segundo micrófono cada 0.1s utilizando el teorema del coseno (Ec. 1) y recalculando el ángulo en que se encuentra la persona (Ec. 2)

$$r_2{}^2 = r_1{}^2 + d^2 - 2r_1 d\cos(\alpha) \tag{1}$$

$$\alpha = \alpha_0 + \omega t = \alpha_0 + \frac{v_c}{r_1}t \tag{2}$$

Por lo tanto, las distancias $r_1$ y $r_2$ son las salidas del bloque y sirven para calcular la fuerza que genera el sonido en las membranas de los micrófonos.

El bloque resultante se muestra en la siguiente figura (fig. 2)

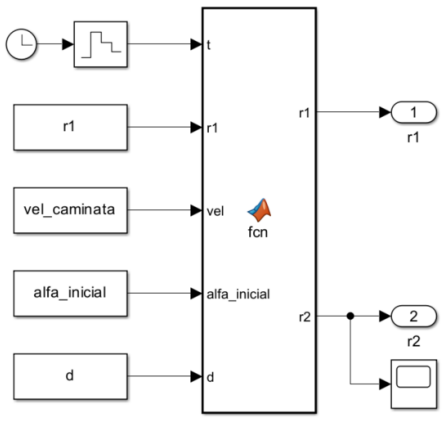


Figura 2: Bloque caminata

## 2.2. Comportamiento del sonido en el sistema

Se realizó un modelo el cual convierte la señal sonora en señales de tensión, y para ello se tuvieron que hacer tres subsistemas:

1. Subsistema mecánico sonoro: convierte la onda de sonido en fuerza que impacta sobre la membrana;
2. Subsistema mecánico de la membrana: sistema mecánico de la membrana, que se modela como un sistema masa-resorte [3] y convierte la fuerza aplicada en la membrana a movimiento de esta;
3. Subsistema eléctrico: Sistema eléctrico de la membrana, que convierte el movimiento de la membrana en variaciones de tensión.

### 2.2.1. Subsistema Mecánico Sonoro

Para el Subsistema Mecánico Sonoro, se utilizó una ecuación [4] la cual relaciona la amplitud del sonido (I) con la fuerza que ejerce dicha amplitud sobre la membrana (Ec. 3) y las constantes que se definieron para dicha ecuación son las que se muestran en la tabla 2

$$F = \sqrt{\rho_{aire} v_s I} \frac{A_m}{r} \tag{3}$$

Tabla 2: Constantes Fuerza sonora

| | Valor | Unidad | Descripción |
|---|---|---|---|
| $\rho_{aire}$ | 1.2 | $\frac{kg}{m^3}$ | Densidad del aire |
| $v_s$ | 343.02 | $\frac{m}{s}$ | Velocidad del sonido en el aire |
| $r_m$ | 0.0004 | m | Radio de la membrana[5] |
| $r_{amb1}$ | 0.8 | m | Distancia virtual de sonido ambiente a la membrana 1 |
| $r_{amb2}$ | 0.7 | m | Distancia virtual de sonido ambiente a la membrana 2 |

En la siguiente imagen, se puede observar el bloque se simulink (fig. 3)

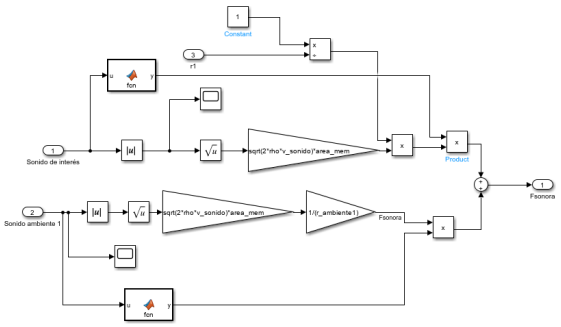


Figura 3: Bloque Subsistema Mecánico Sonoro

### 2.2.2. Subsistema Mecánico de la Membrana

Luego, para el Subsistema Mecánico de la Membrana, se realizó la ecuación dinámica del sistema.(Ec. 4).

$$F(t) = m\ddot{x} + c\dot{x} + kx \tag{4}$$

Para obtener los valores de la masa, rigidez y amortiguamiento, se tuvieron en cuenta materiales, dimensiones, sensibilidad mecánica promedio de micrófonos de membrana y el factor de amortiguamiento [6]. Dichos valores se muestran en la tabla 3

Tabla 3: Prop de membrana

| | Valor | Unidad | Descripción |
|---|---|---|---|
| h | 330e-6 | m | Espesor de membrana |
| $\rho_m$ | 2267 | $\frac{kg}{m^3}$ | Densidad del grafeno apilado |
| $S_{mec}$ | 2.e-9 | $\frac{m}{Pa}$ | Sensibilidad mecánica promedio |
| $\zeta$ | 0 | Adimensional | Factor de amortiguamiento |

Para calcular la masa simplemente se multiplicó el volumen de la membrana por el densidad (Ec. 5). Luego se calculó la rigidez dividiendo el área de la membrana por la sensibilidad mecánica (Ec. 6). Este cálculo contempla la rigidez debida a la mecánica del sistema y a la electroestática. Finalmente se calcula el amortiguamiento, el cual se utiliza el valor del amortiguamiento crítico, la masa y la rigidez previamente calculados (Ec. 7). Aunque el amortiguamiento se considera despreciable, igualmente se incluye en la ecuaciones.

$$m = V\rho_m = A_m h \rho_m = \pi r_m^2 h \rho_m \tag{5}$$

$$k = \frac{A_m}{S_{mec}} = \frac{\pi r_m^2}{S_{mec}} \tag{6}$$

$$c = \zeta c_c = \zeta 2\sqrt{mk} \tag{7}$$

La tabla 4 muestra los valores que se obtuvieron a partir de las ecuaciones.

Tabla 4: Constantes sistema dinámico

| | Valor | Unidad | Descripción |
|---|---|---|---|
| m | 3.7604e-7 | kg | Masa de membrana |
| c | 0 | $\frac{Ns}{m}$ | Amortiguamiento de membrana |
| k | 251.33 | $\frac{N}{m}$ | Rigidez de membrana |

La siguiente figura (fig. 4)) muestra el subsistema mecánico de la membrana:

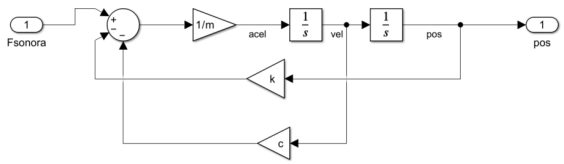


Figura 4: Bloque Subsistema Mecánico de la Membrana

### 2.2.3. Subsistema Eléctrico

Finalmente, para modelar el Subsistema Eléctrico, se tuvo en cuenta que el micrófono es:

- **Micrófono capacitivo**: Consiste en un capacitor con una placa fija y otra acoplada a la membrana, cuya posición variable genera un cambio en la tensión[6];
- **Micrófono omnidireccional**: Capta el sonido de todas las direcciones con igual sensibilidad[7].

Por lo tanto, la tensión dependerá de la carga del capacitor, de la distancia entre placas (variable), del área de superficie (supuesta del mismo tamaño que la membrana) y de la permitividad del medio. Esta relación se puede observar en la siguiente ecuación (Ec. 8)

$$V = \frac{Q}{C} = \frac{Q}{A_m \frac{\epsilon_{aire}}{d-x}} = \frac{Q}{\pi r_m^2 \frac{\epsilon_{aire}}{d-x}} \tag{8}$$

Los valores que se utilizaron en la ecuación anterior (Ec. 8) se muestran en la tabla 5

Tabla 5: Subsistema eléctrico

| | Valor | Unidad | Descripción |
|---|---|---|---|
| Q | 1e-12 | C | Carga del capacitor |
| $\epsilon_{aire}$ | 8.854e-12 | $\frac{F}{m}$ | Amortiguamiento de membrana |
| $s_g$ | 2.2e-6 | m | Distancia entre membrana y placa fija |

La siguiente figura (fig. 5) muestra el subsistema eléctrico:

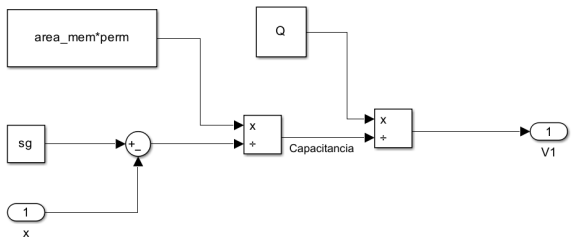


Figura 5: Bloque Subsistema Eléctrico

Con los tres subsistemas, podemos obtener la tensión leída por el procesador a partir de un audio. En la siguiente figura (fig. 6) se observa el acoplamiento entre subsistemas.

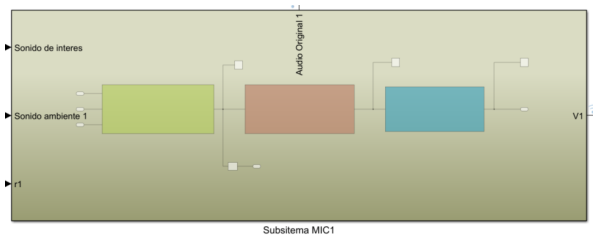


Figura 6: Sistema Completo

## 2.3. Control

El bloque de control calcula ganancias proporcionales en función de la diferencia de amplitudes entre los dos micrófonos en el rango de 10Hz a 1800Hz (espectro en común entre el habla humana y el ruido generado). La metodología es la siguiente:

1. Se define la ventana de tiempo en la cual se analizarán las frecuencias (Ec. 9);

$$\tau_v = 0{,}1s \tag{9}$$

2. Se define el tiempo que retendrá un dato el bloque Zero-Order Hold (ZOH) (Ec. 10);

$$\tau_{zoh} = \frac{1}{f_s} = \frac{1}{44100Hz} = 2{,}2676e-05s \tag{10}$$

3. Se calcula la cantidad de elementos que deberá retener el bloque Buffer tal que corresponda con la ventana de tiempo definida y la tasa de muestreo (Ec. 11);

$$N = \frac{\tau_v}{\tau_{zoh}} = \frac{0{,}1s}{\frac{1}{44100Hz}} = 4410 \tag{11}$$

4. En Simulink, se colocan los bloques en serie y se agrega un bloque de Transformada Rápida de Fourier (FFT) para poder obtener el espectro de frecuencias en forma compleja de la ventana analizada;

5. Se agrega un bloque Submatrix para recortar la matriz y analizar las frecuencias deseadas (10-1800Hz). Para esto se calcula la posición correspondiente a 10Hz y 1800Hz dentro del Buffer(Ec. 12 y Ec- 13):

$$p_{10Hz} = \frac{f_{corte}}{\Delta f} = \frac{f_{corte}}{\frac{f_s}{N}} = \frac{10HZ}{\frac{44100Hz}{4410}} = 1 \tag{12}$$

$$p_{1800Hz} = \frac{f_{corte}}{\Delta f} = \frac{f_{corte}}{\frac{f_s}{N}} = \frac{1800HZ}{\frac{44100Hz}{4410}} = 1800 \tag{13}$$

6. Con el bloque de Complex to Magnitude-Angle, se calcula la amplitud de cada frecuencia de la sub-matriz(fig. 7).

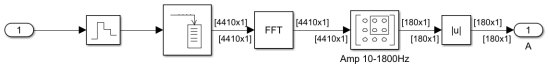


Figura 7: Sistema de Transformada de Fourier

;

7. Se coloca un buffer a la salida para retener el resultado. Se retiene una cantidad de 10 elementos con un solapamiento de 9 elementos.

8. Finalmente se calcula el área bajo la curva de la amplitud en función de la frecuencia entre 10Hz y 300Hz sumando el producto de cada amplitud calculada por el $\Delta f$ (Ec. 14)

$$A_{10-1800Hz} = \sum A_i \Delta f \tag{14}$$

Este bloque completo se puede ver en la siguiente figura(fig. 8);

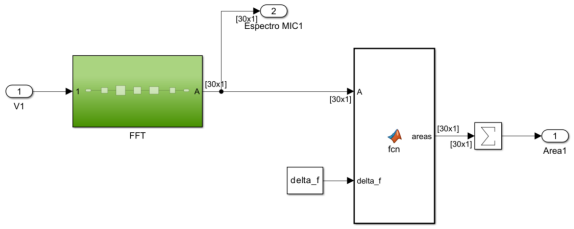


Figura 8: Bloque análisis de área

;

9. Se restan los resultados de cada micrófono.

10. Se asigna una ganancia proporcional a cada micrófono, priorizando aquel con mayor amplitud de frecuencias bajas (10-1800Hz). Para asignar las ganancias se propuso analizar la diferencia entre ambos, y según sea el resultado, las ganancias toman el valor de la siguiente ecuación (Ec. 15)

$$Kp_i = bias \pm |\Delta A| \frac{bias}{A_{max}} \tag{15}$$

Como se puede ver, se le asigna un bias ($bias = 10$) a la ganancia, y el valor que tomarán las ganancias de cada micrófono serán en torno al bias. Sin embargo, las ganancias tomarán dicho valor si y solo si el valor absoluto de la diferencia entre áreas sea mayor a 50 y además que ninguna de las 9 diferencias calculadas anteriormente sea menor a 50. Cuando sucede alguna de estas condiciones, las ganancias tomarán un valor como se muestra en la ecuación (Ec. **??**)

$$Kp_1 = Kp_2 = bias \tag{16}$$

Esta igualdad sirve para evitar darle prioridad a algún micrófono cuando el ruido ambiente predomina.

11. Se multiplican las señales con sus respectivas ganancias. Cabe destacar que las señales que se multiplican son las resultantes del primer filtro aplicado (ver sección 2.4.1)

12. Por último, se realiza la diferencia entre ambas señales para obtener una señal resultante.

## 2.4. Filtrado de señal

Se implementaron dos métodos para realizar el filtrado de señal. El primero consta de restar ambas señales ya que tienen en común el ruido ambiente pero con diferentes amplitudes. El segundo es aplicando un filtro FIR a la señal resultante luego de haber sumado ambas señales y haberles aplicado sus ganancias correspondientes.

### 2.4.1. Resta de señales

Para realizar la resta de señales, se implementó una lógica la cual evita que se atenúe en exceso la señal de interés, ya que si la persona se encuentra a una distancia similar a cada micrófono y luego se restaran ambas señales, entonces la señal de interés sería casi nula o desaparecería. Dicha lógica tiene en cuenta la diferencia entre ambas señales tal que si la diferencia es menor a 0,00005 entonces las señales no se verán modificadas y si es mayor entonces las tensiones nuevas serán las mostradas en las siguientes ecuaciones (Ec. 17 y Ec. 18):

$$v_1 = V_1 - V_2 \tag{17}$$

$$v_2 = V_2 - V_1 \tag{18}$$

Siendo $V_1$ y $V_2$ las señales originales de la salida del subsistema eléctrico del micrófono 1 y del micrófono 2 respectivamente. Como se dijo anteriormente, $v_1$ y $v_2$ serán las señales que se multiplicarán con las ganancias calculadas.

Como se dijo anteriormente, esta etapa sirve para atenuar el ruido ambiente. A continuación se muestra el bloque correspondiente al filtro junto con la implementacion de ganancias (fig. 9)

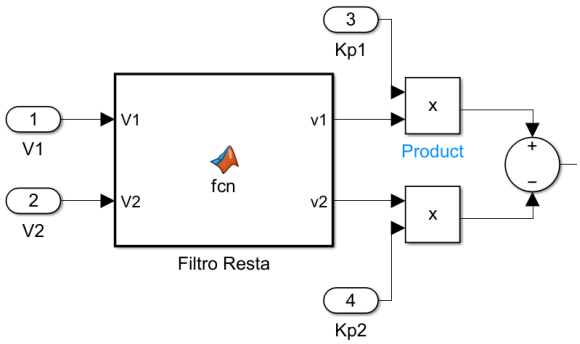


Figura 9: Bloque de resta de señales junto con ganancias

;

### 2.4.2. Filtro FIR

Se aplica un filtro FIR con el objetivo de atenuar el ruido de alta frecuencia (frecuencias mayores a 2000Hz). Para realizarlo, se diseñó un filtro pasa bajos FIR con ventana de Hamming de orden 100 y con una frecuencia de corte de 1800Hz. Este filtro atenúa las frecuencias de gran amplitud ocacionadas por la poca amortiguación que tiene el sistema. Para diseñarlo se utilizo "filterDesigner"de MATLAB y luego se creó el bloque correspondiente en Simulink. La señal filtrada es la diferencia entre las señales luego de aplicarles el primer filtro y ya multiplicadas por sus ganancias. A continuación se muestran los bloques utilizados implementar dicho filtro (fig. 10)

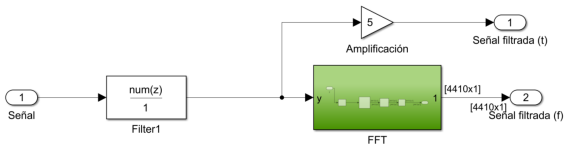


Figura 10: Filtro FIR

;
En la figura siguiente se presentan todos los bloques del sistema (fig. 11)

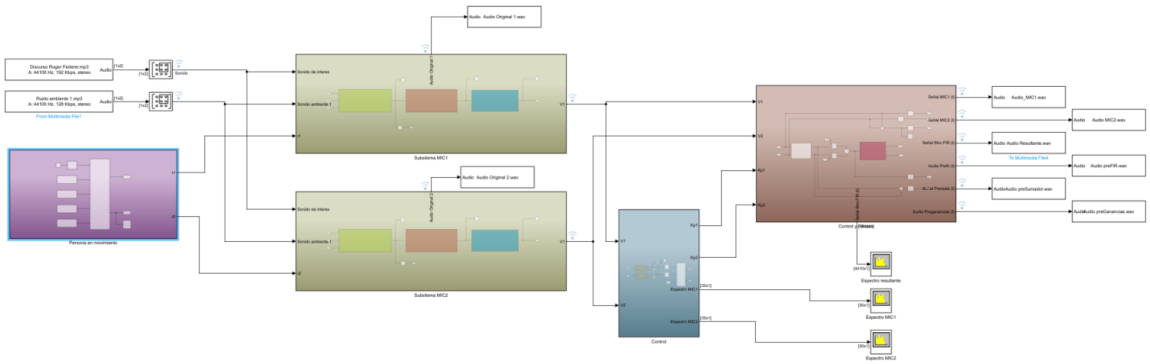

Figura 11: Sistema Completo

;

# 3. Ensayos

Se tuvieron que realizar una serie de ensayos para determinar parámetros tales como espectro de frecuencias deseado, proporciones de ganancias para señales, lógica para suma entre señales y recorrido de la persona.

## 3.1. Caminata de la persona

Para simular la caminata de la persona, se realizó un gráfico el cual representa la posición de la persona respecto a los dos micrófonos. Se realizó solamente un ensayo de una persona caminando de forma radialmente equidistante al primer micrófono (fig. 12)

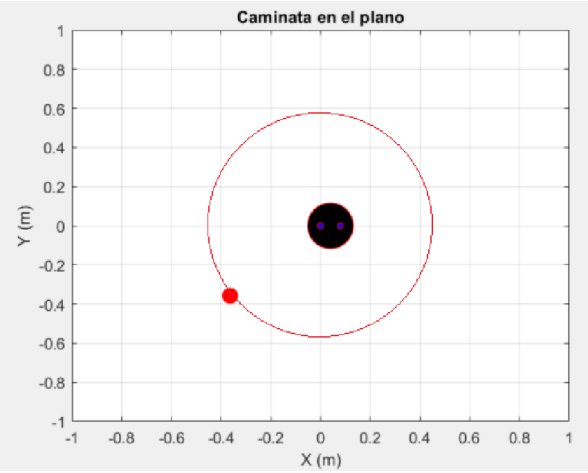


Figura 12: Caminata de la persona

;

## 3.2. Análisis de espectro de frecuencias

Para analizar el espectro de frecuencias del habla humana deseada, se tuvo que analizar, mediante transformadas de fourier, el espectro a la salida del subsistema eléctrico, ya que el del habla de la señal original con respecto al de la tensión son diferentes debido a que la señal pasa por los diferentes subsistemas y dicho espectro cambia. Para lograr el análisis, se realizó en Simulink la primera parte de la simulación (hasta la salida del Subsistema Eléctrico) y se analizó el espectro de la señal de ruido ambiente y la señal de interés (fig 13)

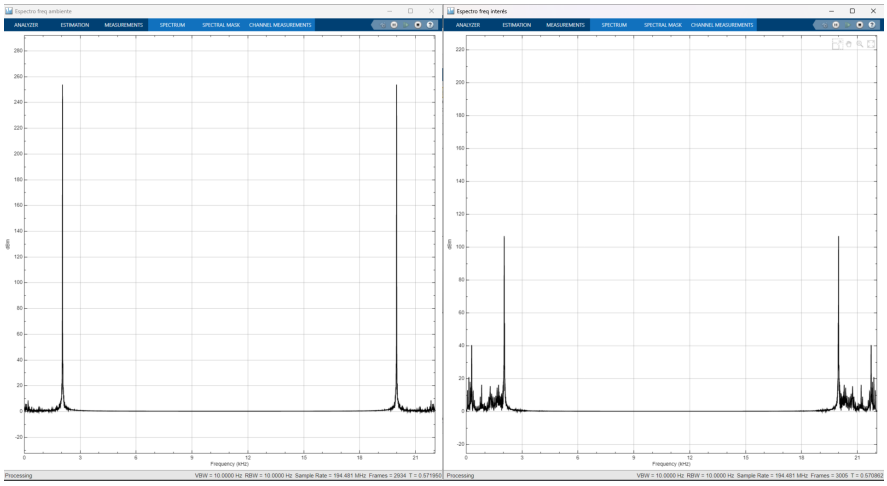

Figura 13: Espectro de ruido ambiente y de señal de interés

Como se puede ver, el espectro del habla humana es aproximadamente de 0Hz a 1800Hz, la del ruido ambiente se encuentra entre 0Hz y 300Hz y la del ruido ocacionada por el propio sistema supera los 2000Hz. Este análisis sirvió para:

- El diseño del filtro;
- Saber que rangos de frecuencia analizar para el cálculo de las ganancias.

## 3.3. Constante para ganancias proporcionales

Para saber el rango de valores que podian tener las ganancias, se debía hacer pruebas de forma iterativa hasta verificar que el rango sea óptimo. Para esto, como primera prueba, se ejecutó el código de tal forma que las ganancias pudieran tomar valores entre 0 y 2 ($bias = 0$). Sin embargo, este rango no fue bueno, ya que no eran suficientemente grandes para luego poder guardar las señales como archivos .mp3 y ser escuchados y además porque la diferencia de ganancias entre micrófonos no era notable. Por lo tanto, para solucionar el primer problema, se agregó un sesgo($bias = 10$), el cual proporcione ganancias alrededor de este, y para solucionar el segundo problema, el valor que se suma o resta al sesgo es proporcional a este.

## 3.4. Resta entre señales

Se analizaron diferentes lógicas para que la señal resultante no amplifique los ruidos, sino que los atenúe. Cabe destacar que para realizar dicho filtro, se supuso que no hay un desfasaje de ruido entre micrófonos pero si una diferencia de amplitudes. Para proporcionar la amplitud de la señal de ruido, se impuso una distancia virtual entre el ruido generado por el archivo y el micrófono ($r_{amb1}$ y $r_{amb2}$), los cuales se han probando de forma iterativa hasta que el audio resultante, luego del subsistema mecánico sonoro, sea coherente y no haya una señal prevaleciente sobre otra (señal se interés sobre la ambiente o viceversa).

# 4. Resultados

En esta sección se mostrarán todos los resultados que se obtuvieron a partir de la simulación.

## 4.1. Audio original

A continuación se muestran la señal que llega a la membrana del primer micrófono (fig. 14)

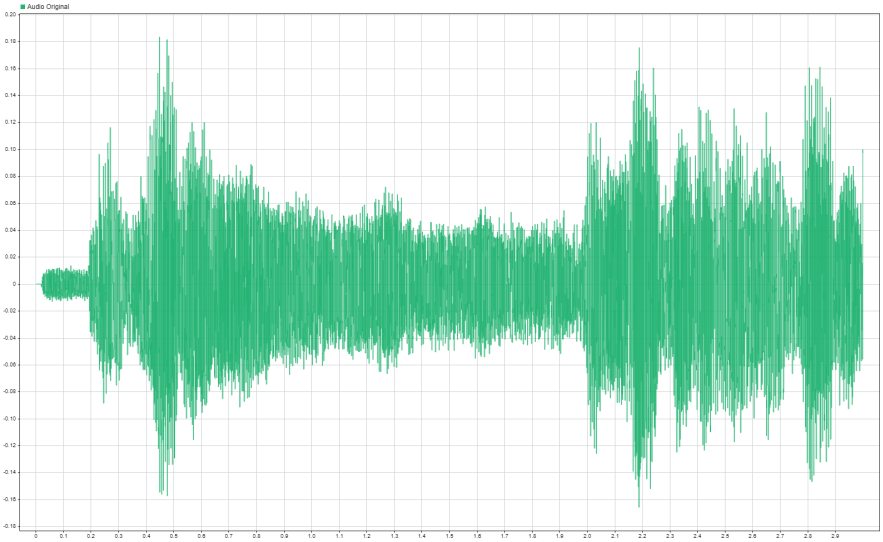


Figura 14: Audio original

;

## 4.2. Señal de tensión para cada micrófono

Cada micrófono generó su señal de tensión dependiendo del audio de entrada (fig. 15)

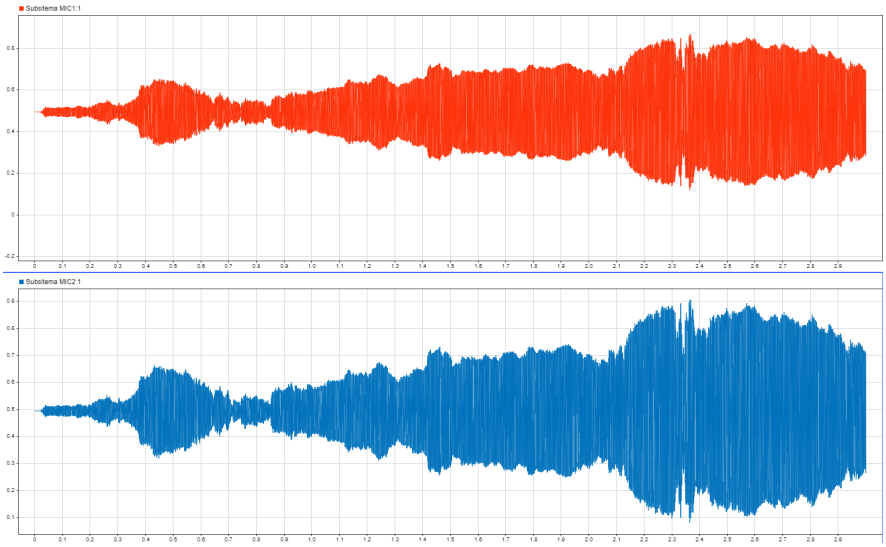


Figura 15: Tensiones originales

;

## 4.3. Diferencia de señales

Se realizó una comparativa entre las diferencias de señales antes y después de aplicar el primer filtro (fig. 16)

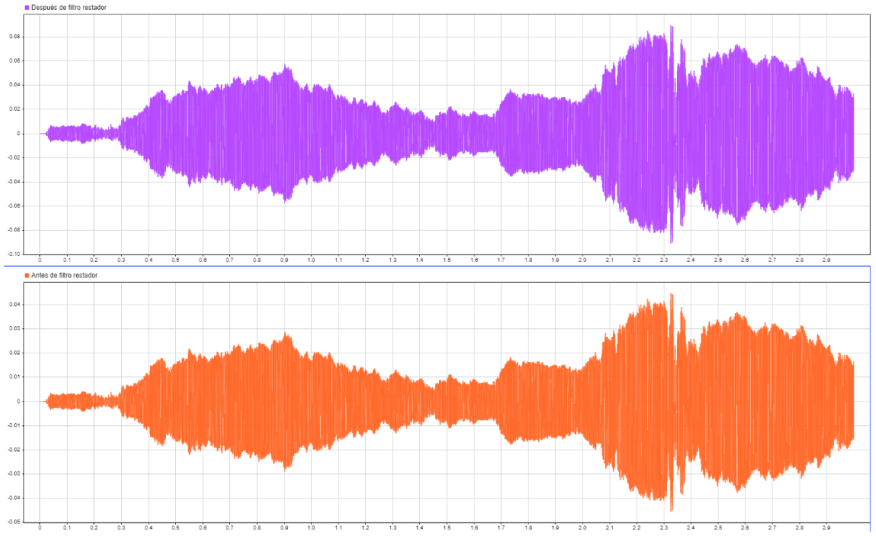


Figura 16: Resta de señales

;

## 4.4. Diferencias entre integrales de A(f) entre 1Hz y 30Hz de ambos micrófonos

La diferencia de las integrales entre 1Hz y 30Hz de ambos micrófonos es escencial para determinar as ganancias de cada micrófono (fig. 17)

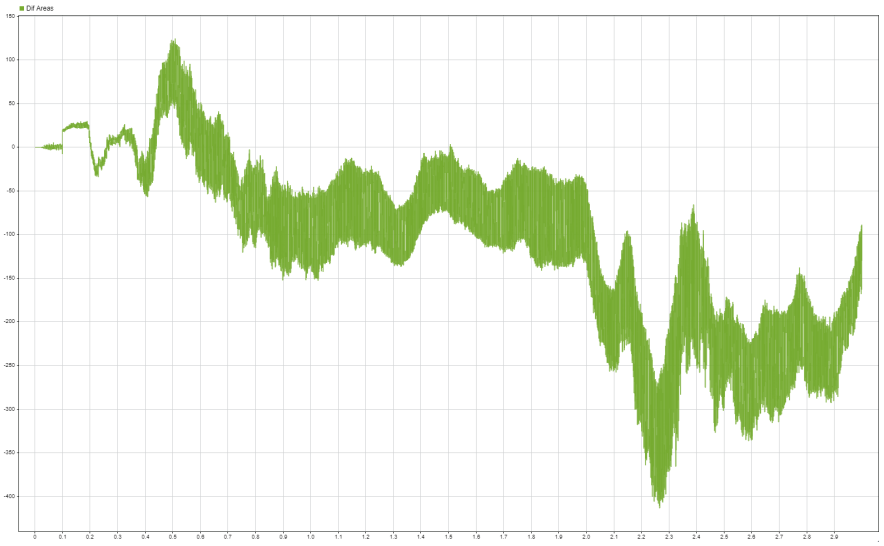


Figura 17: Diferencias entre áreas

;

## 4.5. Valores de $Kp_1$ y $Kp_2$

A continuación se muestra el gráfico de $Kp_1$ y $Kp_2$ (fig. 18). Como se puede observar, cuando la persona no está hablando y solo hay ruido ambiente, las ganancias se igualan a 10.

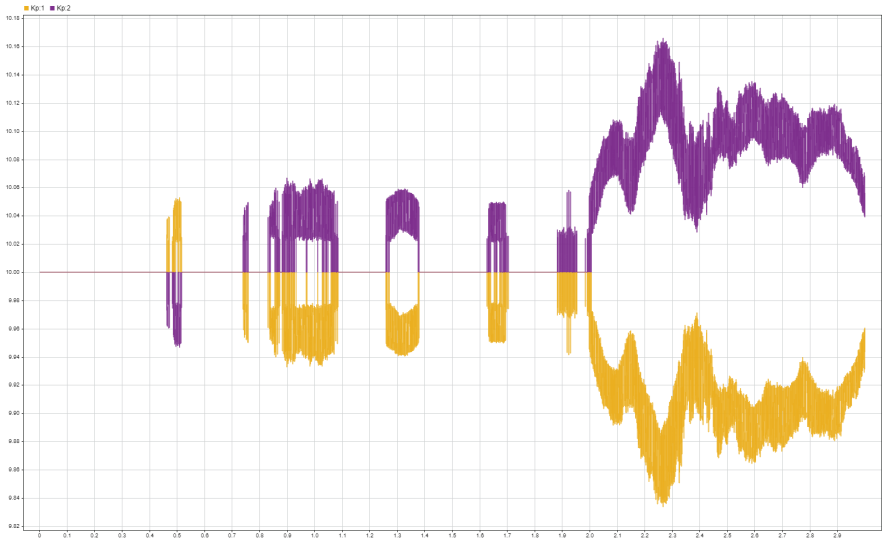


Figura 18: Ganancias

;

## 4.6. Resultado final

Finalmente se muestra la señal de tensión antes y después de aplicar el filtro pasa bajo (fig. 19)

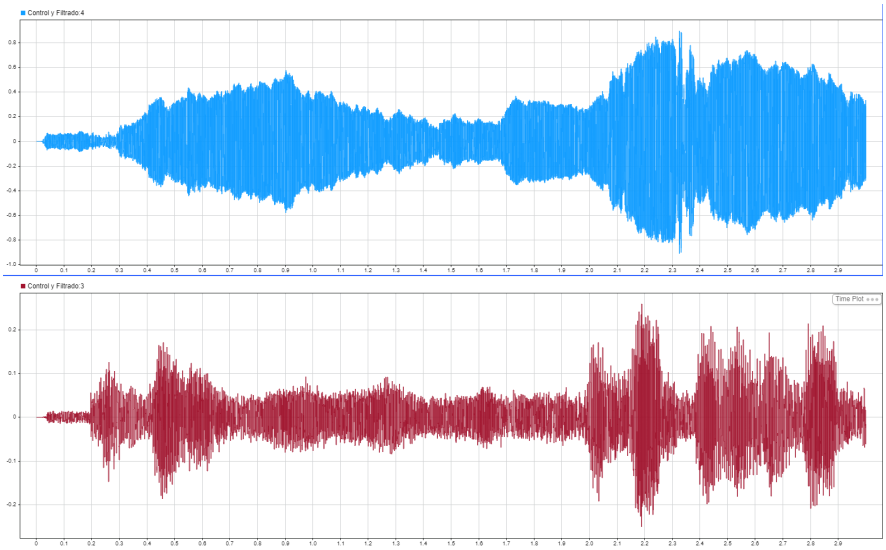


Figura 19: Resultados antes y después de filtro FIR

Además a continuación se pueden ver los espectros de frecuencias correspondientes a la salida de la señal de cada micrófono, sin hacer modificaciones, y el espectro de frecuencias de la señal resultante (fig. 20):

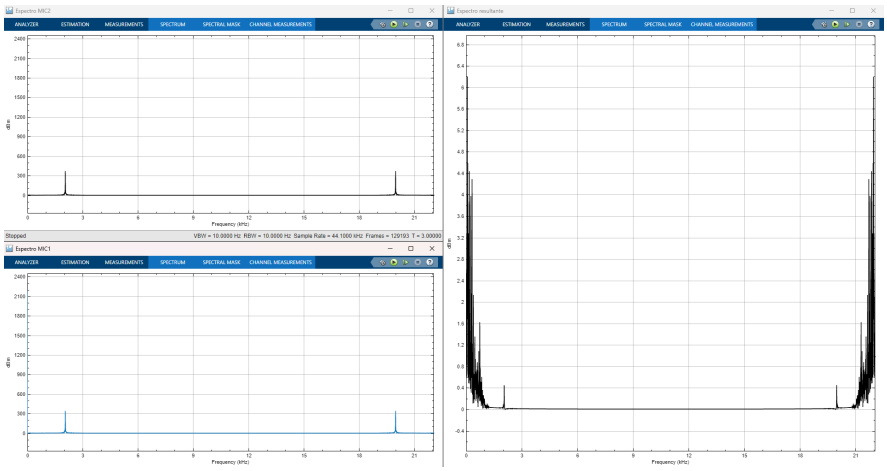

Figura 20: Espectros de frecuencias

;

# 5. Conclusiones

El desarrollo del sistema de captura de voz direccional con filtrado de ruido permitió realizar un buen procesamiento de la señal para su posterior análisis. Gracias a la simulación, se lograron resultados en los siguientes aspectos:

- Cancelación de ruido ambiental: La implementación de ganancias proporcionales ajustadas en función de las diferencias de amplitudes entre los dos micrófonos permitió priorizar la señal del micrófono más cercano al hablante, atenuando eficazmente el ruido ambiental.
- Filtrado de alta frecuencia: El filtro FIR pasa bajos diseñado cumplió su objetivo de reducir las interferencias ocasionadas por las frecuencias superiores a 1800 Hz, generadas por la resonancia de las membranas de los micrófonos, lo que contribuyó a una señal más limpia.
- Eficiencia del sistema de control: El análisis de las frecuencias en el rango de 10 Hz a 1800 Hz demostró ser un buen rango para determinar las ganancias.
- Aunque no se haya lograda filtrar por completo la señal, una mejora que se puede realizar es agregar más cantidad de micrófonos para tener más información del ruido ambiente. Además, agregando más micrófonos, se podrían orientar mucho mejor los micrófonos aplicando las ganancias correspondientes.

# Referencias


[1] Darthmouth. 2024 commencement address by roger federer at dartmouth, 2024. Video en YouTube.

[2] Atmo Sphere. Sonidos de cafeterÍa para concentrarse o relajarse, August 2021. Video en YouTube.

[3] Ramazan-Ali. Jafari-Talookolaei Hamidreza. Habibi, Bahram. Azizollah Ganji. A novel high performance mems capacitive microphone. 2024.

[4] Alan B. Coppens James V. Sanders Lawrence E. Kinsler, Austin R. Frey. *Fundamentals of Acoustics*. John ¡Wiley Sons, Inc, edición 4 edition.

[5] Infineon. High performance digital xensivtm mems microphone, 2017.

[6] Duck-Gyu Lee Shin Hur Muhammad Ali Shah, Ibrar Ali Shah. Design approaches of mems microphones for enhanced performance. *Journal of Sensors*, 2019:26, 2019.

[7] Juan Carlos López. Amazon echo input, análisis: así rinde el dispositivo que lleva a alexa a cualquier altavoz, 2019.