Sintetización de fonemas mediante un modelo AR

Trabajo Práctico 2 - Inferencia y Estimación

Santiago Tobio y Belén Götz

Universidad de San Andrés

1er Semestre 2025

1. Introducción

En este trabajo práctico estudiamos un modelo de síntesis de fonemas utilizando la técnica de *Linear Predictive Coding* (LPC), una metodología reconocida en el campo del procesamiento de voz que permite modelar la producción del habla a través de un sistema Lineal e Invariante en el Tiempo (LTI) basado en un proceso autoregresivo (AR). La técnica LPC fundamenta su efectividad en la capacidad de representar las características en espectrales del tracto vocal mediante un conjunto reducido de coeficientes, aprovechando la naturaleza predictiva inherente en las señales de voz donde cada muestra puede estimarse como una combinación lineal de muestras anteriores.

El modelo propuesto considera la producción del habla como un sistema de dos componentes principales: una *fuente de excitación* que varía según el tipo de fonema (tren de impulsos periódico para sonidos sonoros o ruido blanco gaussiano para sonidos sordos), y un *filtro digital* que modela las resonancias del tracto vocal mediante un sistema IIR autoregresivo de orden $P = 20$. Dada la naturaleza estocástica de las señales de voz, no poseen una transformada de Fourier clásica bien definida, por lo que resulta necesario analizar su contenido frecuencial a través de la *Densidad Espectral de Potencia* (PSD). Esta aproximación permite capturar las características espectrales distintivas de cada fonema, proporcionando una representación matemática precisa de los formantes y estructuras armónicas que definen la percepción auditiva de cada sonido.

Se realizó el análisis y síntesis para un conjunto representativo de fonemas del español: las cinco vocales (a, e, i, o, u) y cuatro consonantes fricativas (j, f, s, sh), modelando las características espectrales de cada uno a partir de coeficientes predefinidos obtenidos mediante técnicas de estimación paramétrica. El objetivo principal era generar señales sintéticas que emulen fidedignamente el sonido de cada fonema, validando la efectividad del modelo AR mediante la comparación entre las densidades espectrales teóricas y empíricas, así como evaluando subjetivamente la calidad perceptual de los fonemas sintetizados.

2. Desarrollo Experimental

El experimento se centra en la implementación de un modelo AR de orden $P = 20$ para la síntesis de los siguientes fonemas: $a, e, i, o, u, j, f, s, sh$. Los coeficientes a_i y la ganancia b son específicos para cada fonema y se utilizan para definir el modelo AR según la siguiente ecuación:

$$X(n) = \sum_{i=1}^{P} a_i X(n-i) + b \cdot U(n) \qquad (1)$$

Donde $X(n)$ es la señal de salida y $U(n)$ representa el proceso de excitación, que puede ser un tren de impulsos (para las vocales) o ruido blanco (para las consonantes).

Inicialmente, se graficaron las señales de cada fonema, considerando un tramo de 200 ms. Posteriormente, se calcularon y graficaron las funciones de autocorrelación para cada fonema, obtenidas directamente de las señales de audio provistas.

Luego, se calculó el periodograma para cada fonema y se comparó con su respectiva PSD teórica, definida por:

$$S_X(f) = \frac{b^2}{\left|1 - \sum_{i=1}^{P} a_i \exp\left(-j\frac{2\pi f i}{f_s}\right)\right|^2} S_U(f) \qquad (2)$$

Finalmente, se realizó una síntesis de los fonemas a partir del modelo AR, reproduciendo

las señales resultantes. Adicionalmente, se generaron señales sintetizadas con diferentes frecuencias de pitch para las vocales.

3. Resultados y Análisis

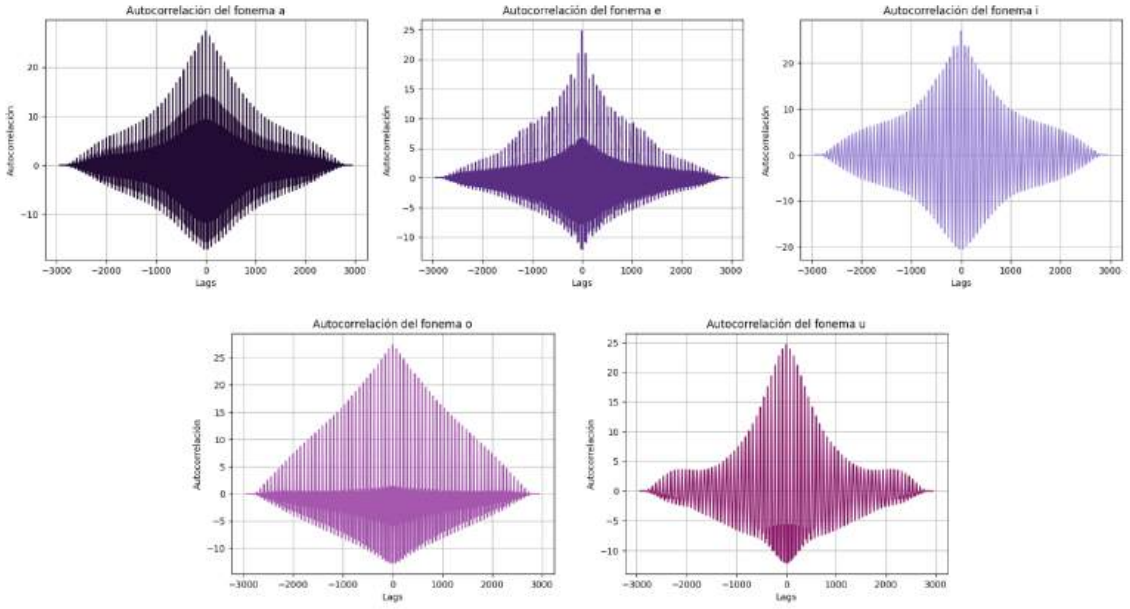

Figura 1: Funciones de autocorrelacción de vocales.

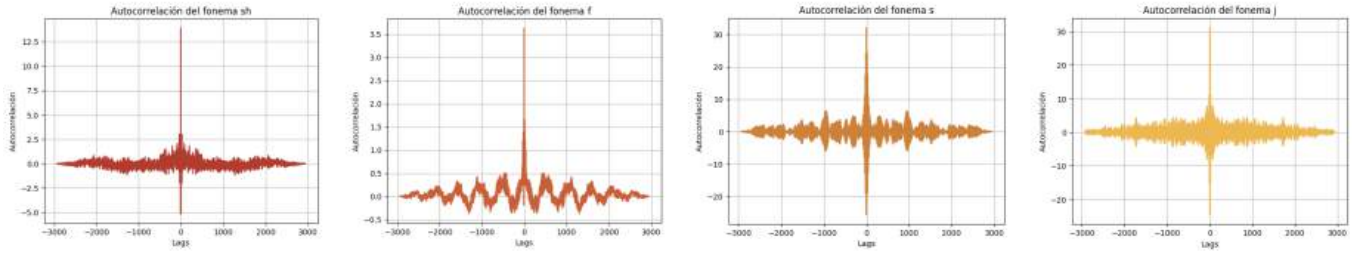

Figura 2: Funciones de autocorrelación de consonantes.

3.1. Análisis cualitativo de las autocorrelaciones

Vocales: Las funciones de autocorrelación de las vocales (/a/, /e/, /i/, /o/, /u/) nos revelan características propias de señales sonoras con fuerte componente periódica. Esta periodicidad marcada es consecuencia directa de la excitación producida por las cuerdas vocales durante la producción del habla, manifestándose como un tren

de impulsos regular que atraviesa el tracto vocal. Se aprecian algunas diferencias sutiles entre vocales: las cerradas como /i/ y /u/ exhiben oscilaciones más regulares y definidas, mientras que vocales abiertas como la /a/ presentan una estructura más densa con mayor concentración de energía.

Consonantes fricativas: Las funciones de autocorrelación de las consonantes fricativas (/s/, /sh/, /f/, /j/) exhiben patrones claramente diferenciados de las vocales, característicos de señales sordas generadas mediante excitación con ruido blanco. El rasgo más distintivo es la concentración de la energía principalmente en torno al lag cero, manifestada como un pico central prominente, con un decaimiento mucho más rápido al alejarse del origen. Esta estructura refleja la naturaleza más aleatoria y menos periódica de estos sonidos, producidos por turbulencia del aire al pasar por constricciones en el tracto vocal sin participación de las cuerdas vocales.

3.2. Análisis del espectro de frecuencias de los fonemas

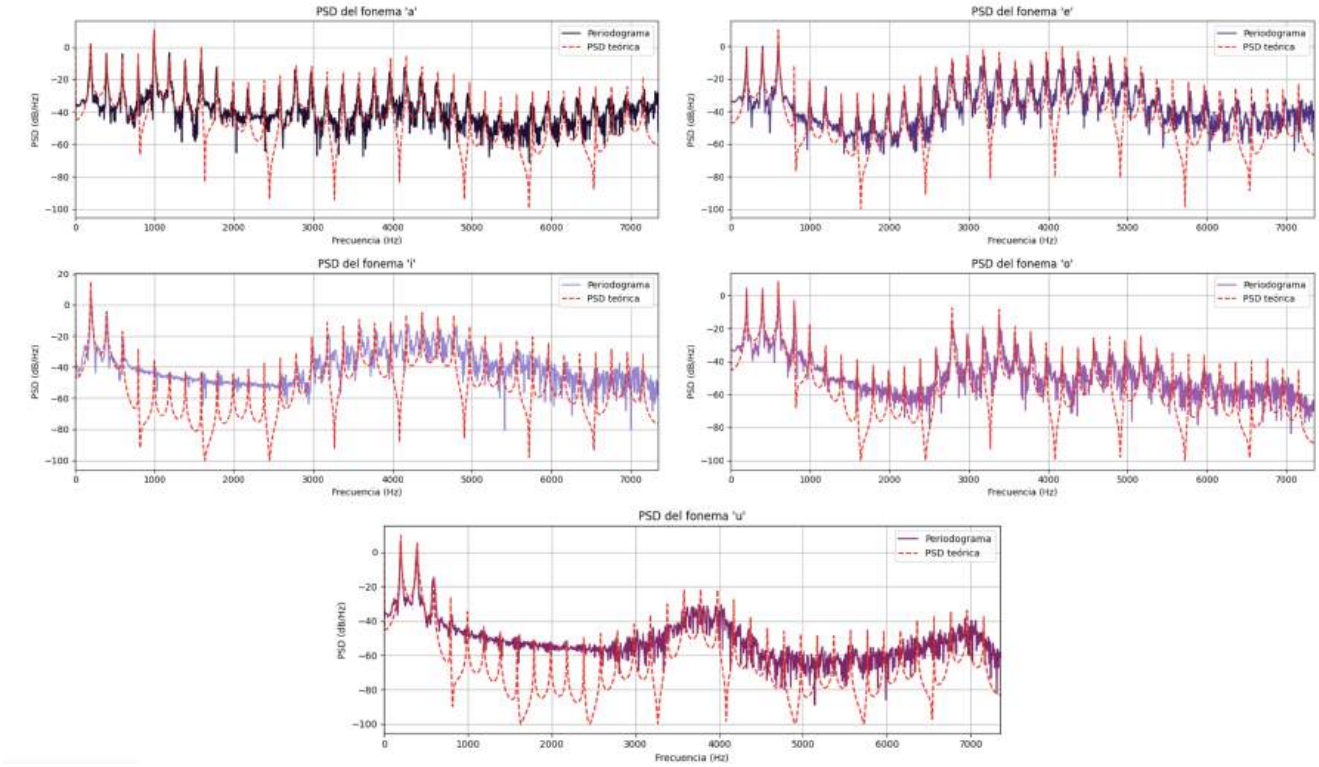

Figura 3: PSD y periodograma de vocales.

La densidad espectral de potencia de las vocales muestra características distintivas relacionadas con la configuración del tracto vocal durante su articulación. En los gráficos se observa claramente la comparación entre el periodograma empírico (líneas violetas continuas) y la PSD teórica del modelo AR (líneas rojas punteadas). Las vocales presentan una estructura de forman-

tes bien definida, con concentraciones de energía en bandas específicas de frecuencia que actúan como resonancias del tracto vocal. El primer formante (F1) es prominente en todas las vocales, visible como un pico significativo por debajo de los 1000 Hz. El modelo AR teórico captura exitosamente la ubicación de los formantes principales, aunque tiende a subestimar la energía en los valles entre formantes, mostrando caídas más pronunciadas (hasta -100 dB) que no se observan en el periodograma empírico (generalmente limitado a -60 dB). La estructura armónica es evidente en todas las vocales, manifestándose como picos regularmente espaciados que corresponden a múltiplos de la frecuencia fundamental de excitación, validando el modelo de excitación periódica para estos fonemas sonoros.

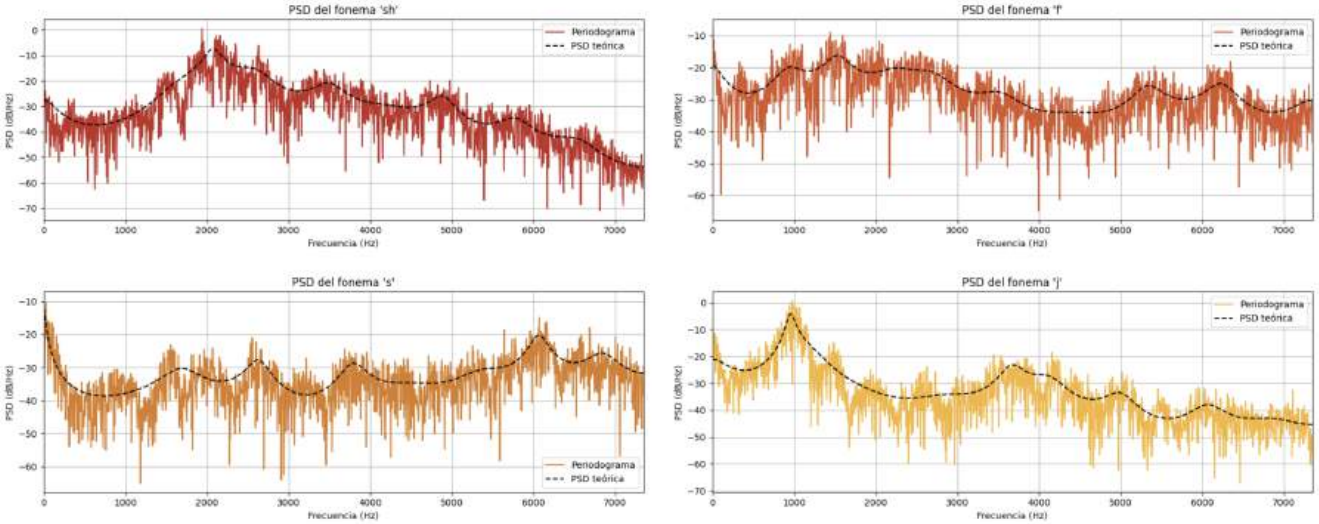

Figura 4: PSD y periodograma de consonantes frictivas

Las consonantes fricativas, como (/s/, /sh/, /f/, /j/) se distinguen por generar un espectro de ruido amplio y aperiódico, originado por la turbulencia del aire al pasar por constricciones en el tracto vocal. Esto las diferencia claramente de las vocales, cuyos espectros exhiben formantes bien definidos y una estructura armónica regular. La PSD teórica, basada en el modelo autoregresivo (AR), suaviza estas fluctuaciones y ofrece una visión general de la tendencia espectral, pero tiende a subestimar la energía en las regiones de menor intensidad, mostrando caídas más marcadas, como -60 dB, en comparación con los -40 dB del periodograma. Esta diferencia, aunque también presente en vocales, se acentúa en fricativas debido a su componente de ruido continuo, lo que pone en evidencia las limitaciones del modelo AR para capturar detalles finos en fonemas no periódicos.

3.2.1. Análisis cuantitativo del ajuste del modelo

Fonema	Tipo	RMSE Lin.	RMSE (dB)
s	Fric.	3.54e-3	6.78
f	Fric.	9.26e-3	6.88
j	Fric.	4.82e-2	6.99
sh	Fric.	4.42e-2	7.16
Media fric.		**2.49e-2**	**6.95**
a	Vocal	3.00e-1	13.14
e	Vocal	2.26e-1	13.88
o	Vocal	1.22e-1	14.81
i	Vocal	6.21e-1	19.03
u	Vocal	2.16e-1	20.10
Media voc.		**3.05e-1**	**16.19**

Tabla 1: Error cuadrático medio entre periodograma empírico y PSD teórica.

Para evaluar la precisión del modelo AR en la representación de fonemas, se calculó el error cuadrático medio (RMSE) entre el periodograma empírico y la PSD teórica, en escalas lineal y logarítmica. Los resultados se presentan en la Tabla 3.2.1.

El análisis muestra que las consonantes fricativas tienen errores menores (RMSE promedio de 6.95 dB) frente a las vocales (RMSE promedio de 16.19 dB). Esto se debe a que las fricativas, con su excitación de ruido blanco y espectros suaves sin estructura armónica compleja, son más fáciles de modelar con el filtro AR.

Las vocales, en cambio, presentan mayores desafíos por su estructura armónica rica y formantes definidos. Las vocales cerradas /i/ (19.03 dB) y /u/ (20.10 dB) tienen los mayores errores debido a sus formantes estrechos y separados, mientras que las vocales abiertas /a/ (13.14 dB) y /e/ (13.88 dB) muestran mejor ajuste por sus formantes más anchos y superpuestos.

Entre las fricativas, /s/ tiene el mejor ajuste (6.78 dB), gracias a su espectro concentrado en altas frecuencias, mientras que /sh/ presenta el mayor error (7.16 dB), aunque dentro de un rango reducido.

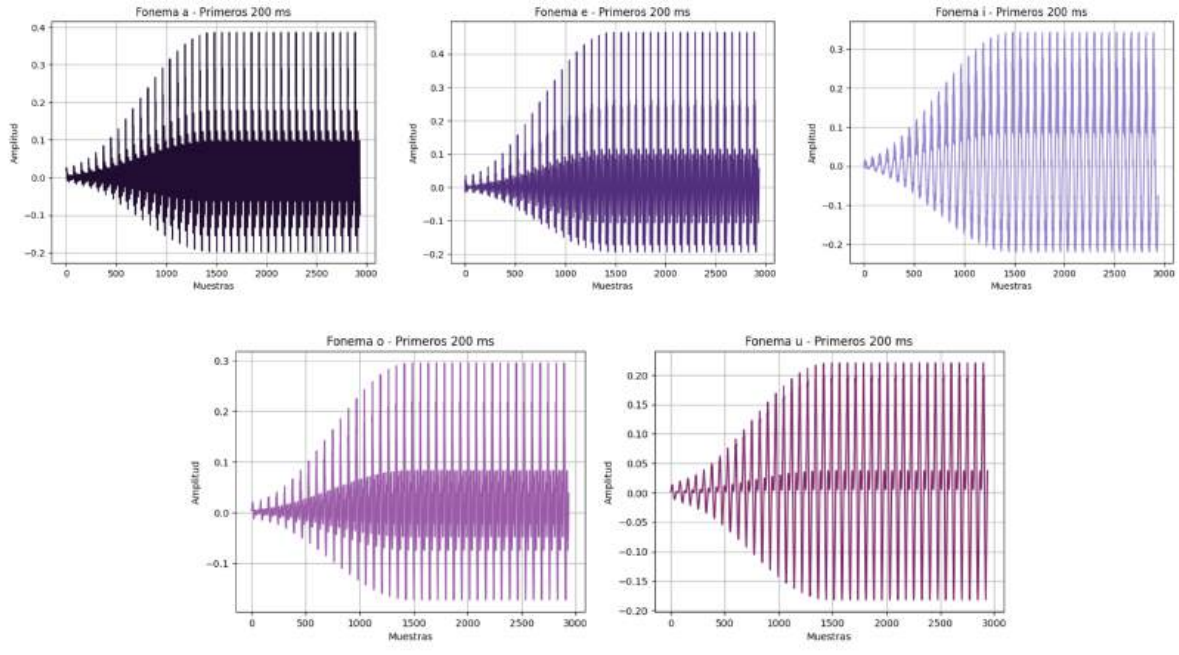

Figura 5: Amplitud de fonemas de vocales.

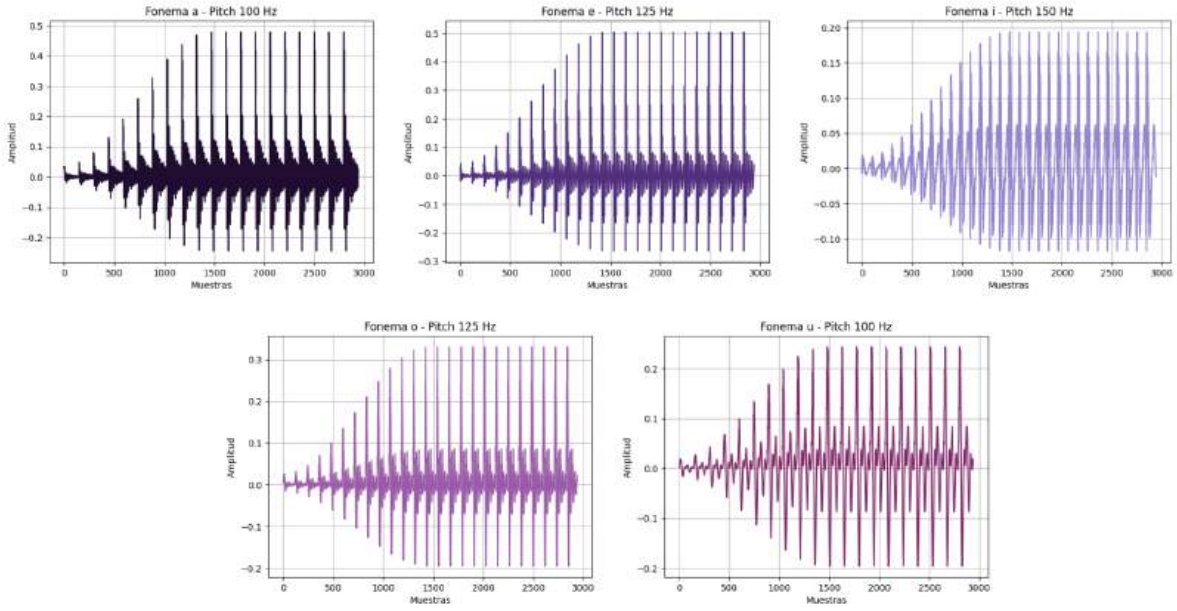

Figura 6: Amplitud de fonemas de vocales con distintos pitch.

3.2.2. Síntesis de vocales con diferentes pitch (frecuencias de la onda sonora)

Con el objetivo de explorar la influencia de la frecuencia fundamental en la percepción de los fonemas vocálicos, se implementó una variante del modelo en la que cada vocal fue sintetizada utilizando un pitch distinto, asignado según valores típicos de frecuencia fundamental para el habla humana:

`'a': 100 Hz, 'e': 125 Hz, 'i': 150 Hz, 'o': 125 Hz, 'u': 100 Hz`

Se buscó evaluar si el modelo AR es capaz de preservar la identidad de cada vocal incluso cuando la frecuencia de excitación cambia, lo que resulta fundamental en contextos reales donde el pitch varía constantemente (por ejemplo, entre hablantes o según la entonación).

En la Figura (6) se muestran las formas de onda de las vocales sintetizadas con sus respectivos pitches. Las diferencias en la frecuencia fundamental se manifiestan visualmente en la densidad de los ciclos oscilatorios y auditivamente en la percepción de la altura tonal. Se puede observar como hay mayor densidad de ciclos en los pitches más altos, como en la /i/, la cual se escucha más aguda que la /a/ o la /u/, donde la densidad de ciclos es menor (un pitch más grave).

Desde el punto de vista perceptual, las señales sintetizadas mantienen su carácter fonémico distintivo, lo que confirma que el modelo propuesto es robusto ante variaciones del pitch.

Este resultado valida la hipótesis de que el contenido armónico del habla, gobernado por la fuente de excitación, modula la percepción de la altura sonora (pitch), mientras que el contenido formántico, definido por el tracto vocal (el filtro),

determina el timbre o identidad fonética.

Además, se comprobó que los espectros de las vocales mantienen su estructura formántica general, con desplazamientos proporcionales de los armónicos según el pitch, sin que ello afecte la localización de los formantes principales. Esto refuerza el modelo fuente-filtro tradicional de la producción del habla, donde la fuente determina la frecuencia fundamental y el filtro da forma a la envolvente espectral.

4. Conclusión

En este trabajo evaluamos la efectividad del modelo autoregresivo (AR) basado en la técnica de Linear Predictive Coding (LPC) para la síntesis de fonemas del español, específicamente vocales (/a/, /e/, /i/, /o/, /u/) y consonantes fricativas (/j/, /f/, /s/, /sh/). Los resultados muestran que el modelo AR captura adecuadamente las características espectrales generales de ambos tipos de fonemas, como la ubicación de los formantes en vocales y las tendencias de espectros suaves en fricativas. Sin embargo, el análisis cuantitativo mediante el RMSE revela una mayor precisión en las consonantes fricativas (RMSE promedio de 6.95 dB) frente a las vocales (RMSE promedio de 16.19 dB), lo que se atribuye a la simplicidad de modelar señales con excitación de ruido blanco frente a estructuras armónicas complejas. Esta diferencia resalta la capacidad del modelo para representar fonemas no periódicos con mayor fidelidad, mientras que las resonancias marcadas y los valles entre formantes de las vocales representan un desafío significativo.

A pesar de sus limitaciones, el enfoque AR ofrece una herramienta valiosa en el procesamiento de voz debido a su simplicidad computacional y su capacidad para modelar las resonancias del tracto vocal de manera compacta. No obstante, la subestimación de la energía en regiones específicas del espectro, como los valles entre formantes en vocales y las fluctuaciones rápidas en fricativas, sugiere la necesidad de ajustes o extensiones del modelo, como el uso de órdenes superiores o la incorporación de técnicas no lineales, para mejorar la precisión en fonemas sonoros y capturar detalles finos en señales aperiódicas. Estos hallazgos subrayan la importancia de adaptar el modelo al tipo de fonema analizado, abriendo la puerta a futuras investigaciones que optimicen su desempeño en aplicaciones de síntesis y reconocimiento del habla.