Universidad de Buenos Aires
Facultad de Ingeniería
Año 2025 - 2º cuatrimestre

Procesos Estocásticos

Trabajo Práctico integrador

ESTUDIANTES: Grupo 6

Del Rio, Francisco Agustín	110761
fadelrio@fi.uba.ar	
Monti, Martina	110574
mmonti@fi.uba.ar	
Chahuillco Pichihua, Aleksey Victor	110929
achahuillco@fi.uba.ar	

Índice

1. Resumen

En el presente trabajo se busca desarrollar un sistema de codificación y síntesis de voz mediante la técnica de codificación predictiva lineal (LPC), con el objetivo de modelar y reconstruir señales de habla. Se busca representar señales de habla mediante un conjunto reducido de parámetros, para ello, a partir del análisis y procesamiento de fonemas y audios completos, se estiman dichos parámetros y se aplican métodos para su reconstrucción, incluyendo una detección de pitch y uso de distintos tipos de excitación con respecto al fonema. El objetivo es comprender y aplicar los conceptos teóricos del modelo LPC en una implementación practica, analizando los resultados sobre su capacidad para sintetizar señales de voz.

2. Introducción

La voz es un fenómeno en que existe correlación entre sus muestras tanto a lo largo como a corto plazo. A corto plazo, esta correlación se debe a la lenta variación de la envolvente de su espectro en amplitud, lo que significa que las muestras cercanas en el tiempo son parecidas entre si. En el caso de los sonidos sonoros, las muestras se relacionan unas con otras de forma que puede predecirse aproximadamente el valor de una muestra a partir de las anteriores. Esta característica hace que la técnica de predicción lineal sea una herramienta adecuada par estimar dichas muestras. Por otro lado la correlación a largo plazo está asociada a la periodicidad de la señal, especialmente en fonemas sonoros que repiten un patrón con una frecuencia determinada (pitch o frecuencia fundamental). Detectar ese patrón permite comprimir la señal, eliminando información repetitiva sin afectar significativamente la comprensión del mensaje.

A lo largo del trabajo se estudia la técnica de codificación predictiva lineal (LPC) con el objetivo de representar cada segmento de audio como la salida de un modelo autorregresivo de orden P, excitado por una fuente periódica o ruido blanco. Este modelo aprovecha la correlación temporal en la voz para comprimir eficientemente la señal.

Con el fin de poder realizar un análisis adecuado de la voz, en orden a esta clasificación de sonidos sordos y sonoros, es necesario dividirla en segmentos en los que puede considerar que las características de la voz permanecen constantes como la de un proceso estocástico estacionario.

3. Modelo

A continuación, se definen los recursos teóricos utilizados en la implementación para lograr un correcto modelado.

Para el análisis LPC y extracción de parámetros, la señal de voz $X(n)$ se modela mediante la siguiente ecuación en diferencias:

$$X(n) = \sum_{k=1}^{P} a_k X(n-k) + G \cdot U(n) \tag{1}$$

Donde a_k son los coeficientes del modelo autor regresivo, G la ganancia del sistema y $U(n)$ la señal de excitación definida como un tren de impulsos o ruido blanco. Luego, para encontrar los coeficientes, se parte del calculo de la autocorrelacion $R_x(k)$ de la señal:

$$r = [R_X(0), R_X(1), \ldots, R_X(P)] \tag{2}$$

A partir de esto, se forma la matriz de Toeplitz R y se resuelve el sistema de ecuaciones de Yule-Walker:

$$\mathbf{a} = R^{-1} \cdot \mathbf{r} \tag{3}$$

Para el calculo del valor cuadrático de la ganancia G^2 se utiliza la siguiente ecuación:

$$G^2 = R_X(0) - \sum_{k=1}^{P} a_k R_X(k) \tag{4}$$

Esta misma representa la energía del residuo o error obtenido luego de la aplicación del filtro LPC, el mismo se asocia a la intensidad de la fuente de excitación que produce el sonido.

Una vez calculado la ganancia es posible desarrollar la densidad espectral de potencia (PSD) del modelo mediante la siguiente ecuación.

$$S_X(\omega) = \frac{G^2 \cdot S_U(\omega)}{\left|1 - \sum_{k=1}^{P} a_k e^{-j\omega k}\right|^2} \tag{5}$$

En donde $S_U(w)$ es la PSD de la señal de excitación que puede tomar una valor constante y unitario si respecta a un ruido blanco gaussiano o un tren de impulsos si respecta a un sonido sonoro.

Para los fonemas sonoros, se puede estimar la frecuencia fundamental (pitch) analizando la autocorrelación del residuo de predicción resolviendo:

$$e(n) = X(n) - \sum_{k=1}^{P} a_k X(n-k) \tag{6}$$

Este residuo representa la parte no predecible de la señal y se usa tanto para estimar la energía como para detectar el pitch, haciendo uso de un método de autocorrelacion del residuo.

Conociendo la frecuencia de muestreo del audio, se puede estimar la posición del segundo pico máximo de la autocorrelacion k, asumiendo que dicho segmento es sonoro y calculando:

$$f_p = \frac{f_s}{k} \tag{7}$$

Es posible reconstruir la señal generando una señal de excitación, ajustando la ganancia y pasándolo por un filtro $H(z)$ construido a partir de los coeficiente LPC.

$$H(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{8}$$

4. Estimación de parámetros del modelo

A continuación se describe el algoritmo desarrollado para el calculo de los coeficientes de predicción lineal y la ganancia de un segmento especifico.

- Estimación de autocorrelacion: Se calcula como la señal de voz se correlaciona consigo misma en diferentes retardos, donde $R_x(0)$ representa la energía total del segmento.
- Construcción de la matriz Toeplitz: Asumiendo al segmento como un proceso ESA, se arma una matriz simétrica denominada Toeplitz a partir de los valores de autocorrealacion estimados.
- Resolución del sistema de ecuaciones: A partir de los parámetros obtenidos se desarrolla (3) obteniendo los P coeficientes del LPC para modelar el filtro.
- Calculo de ganancia: se computa el valor de la ganancia cuadrática G^2 a partir de 4. Luego, para evitar los términos complejos se realiza la raíz cuadrada y se consideran unicamente los valores positivos $G > 0$.

Como herramienta para el análisis visual se utilizan archivos provistos por la cátedra de distintos segmentos individuales de fonemas sonoros y sordos, en donde se estiman los parámetros de los coeficiente y ganancia con el algoritmo desarrollado.

Luego se realiza el calculo de la señal original mediante la Transformada rápida de fourier (FFT) para obtener una estimación de la densidad espectral de potencia (PSD) tanto de la señal real como el del modelo LPC. Específicamente en el modelo se computa 5, en donde se utiliza una PSD para la excitación $S_U(w)$ constante y unitaria en los casos de voz sorda, y una simulación de tren de pulsos en caso de voz sonora.

Las estimaciones y gráficos fueron simulados con modelos de ordenes $P = (5, 10, 30)$ y en escala logarítmica (dB), permitiendo comparar de forma directa las envolvente espectrales.

4.1. Sonido “a”

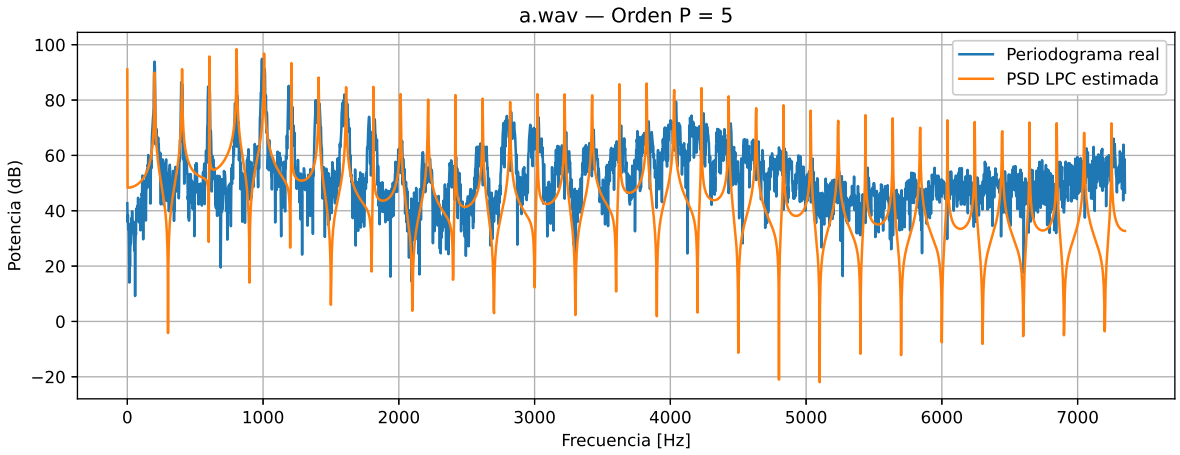

Figura 1: PSD real superpuesta a la obtenida con el modelo de orden 5

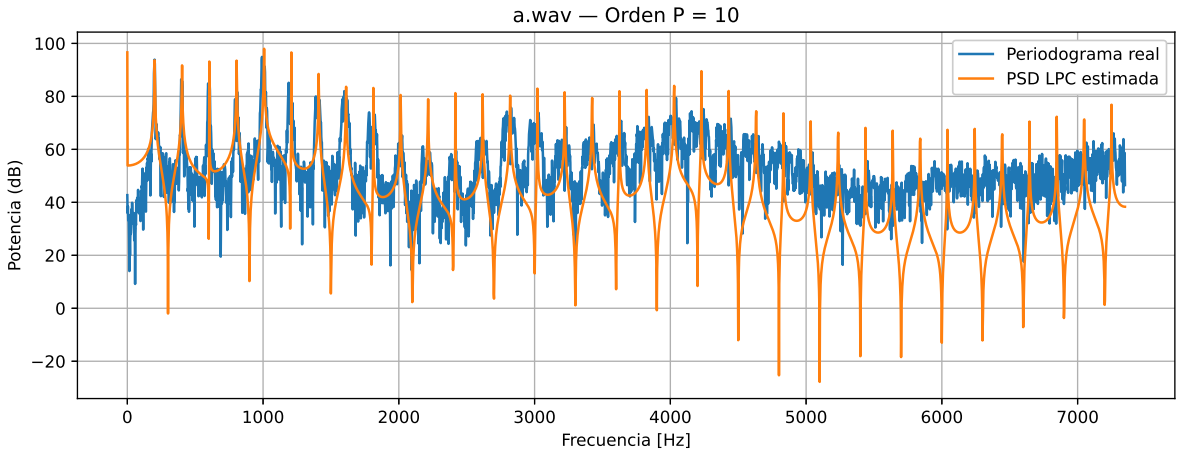

Figura 2: PSD real superpuesta a la obtenida con el modelo de orden 10

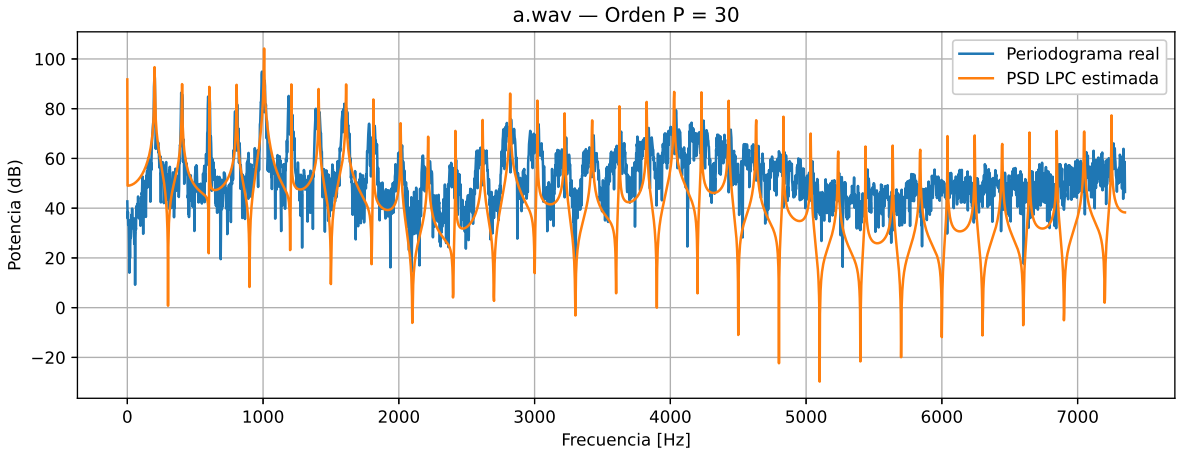

Figura 3: PSD real superpuesta a la obtenida con el modelo de orden 30

4.2. Sonido “e”

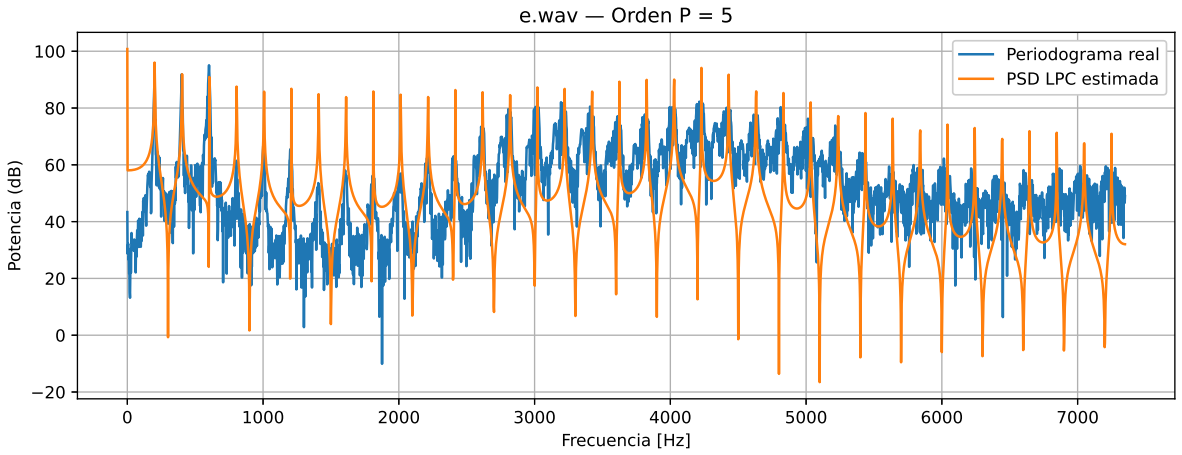

Figura 4: PSD real superpuesta a la obtenida con el modelo de orden 5

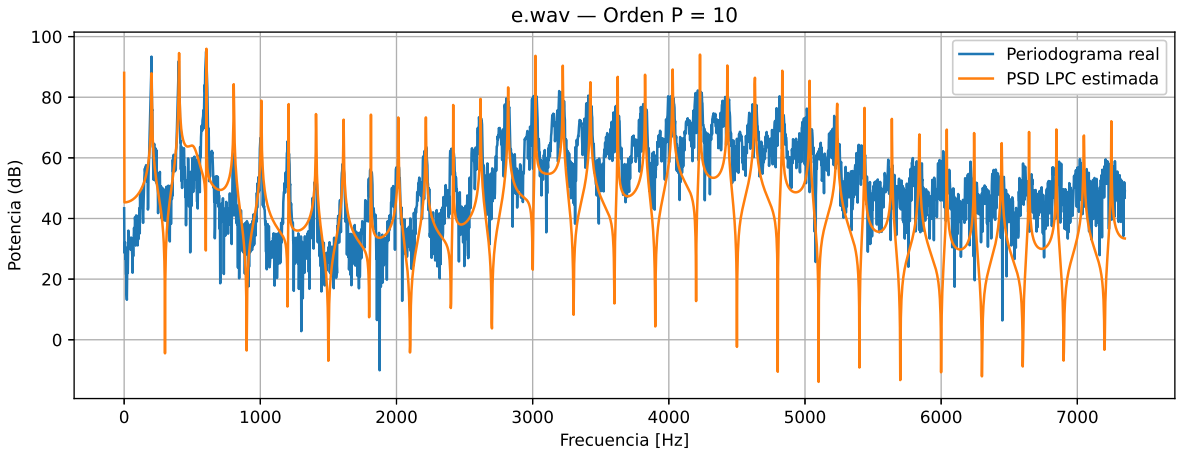

Figura 5: PSD real superpuesta a la obtenida con el modelo de orden 10

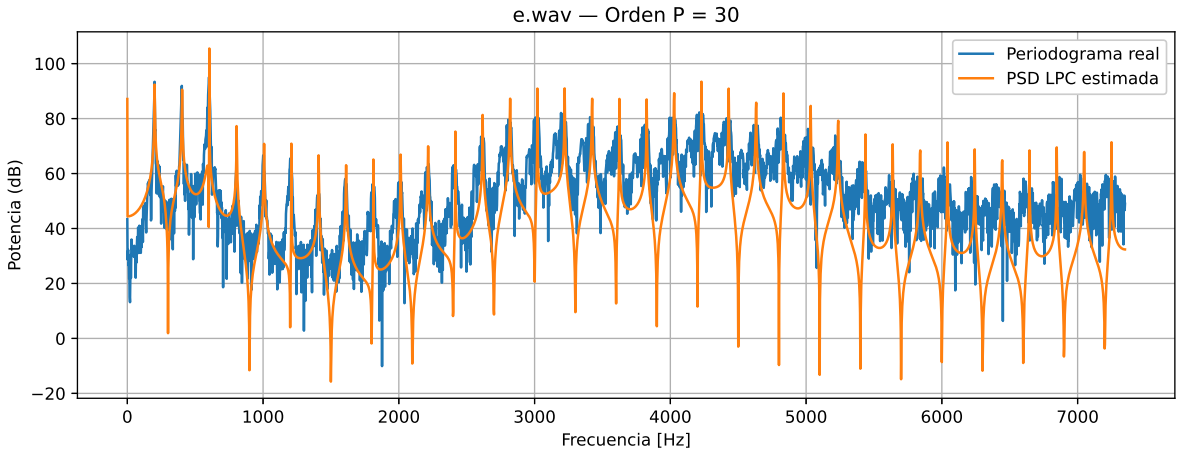

Figura 6: PSD real superpuesta a la obtenida con el modelo de orden 30

4.3. Sonido “s”

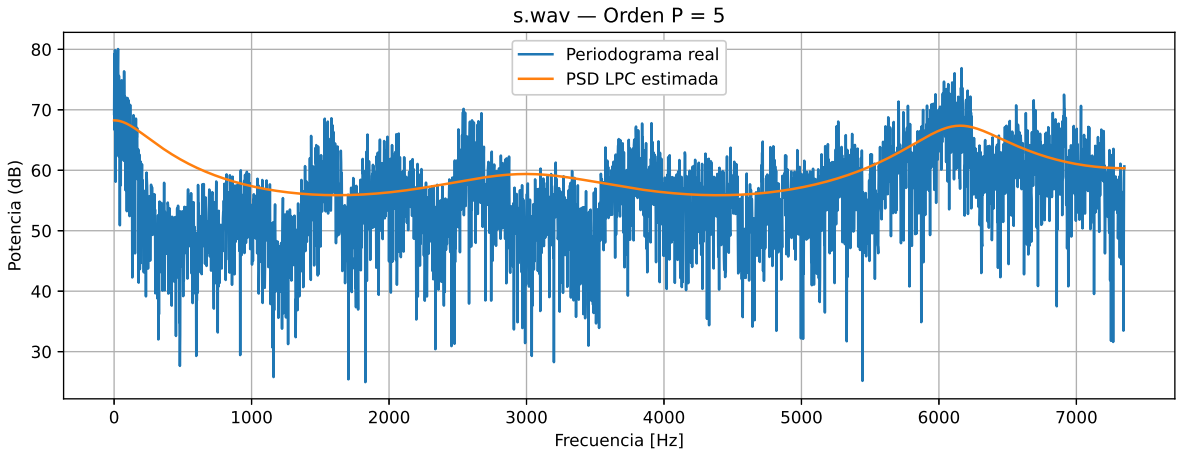

Figura 7: PSD real superpuesta a la obtenida con el modelo de orden 5

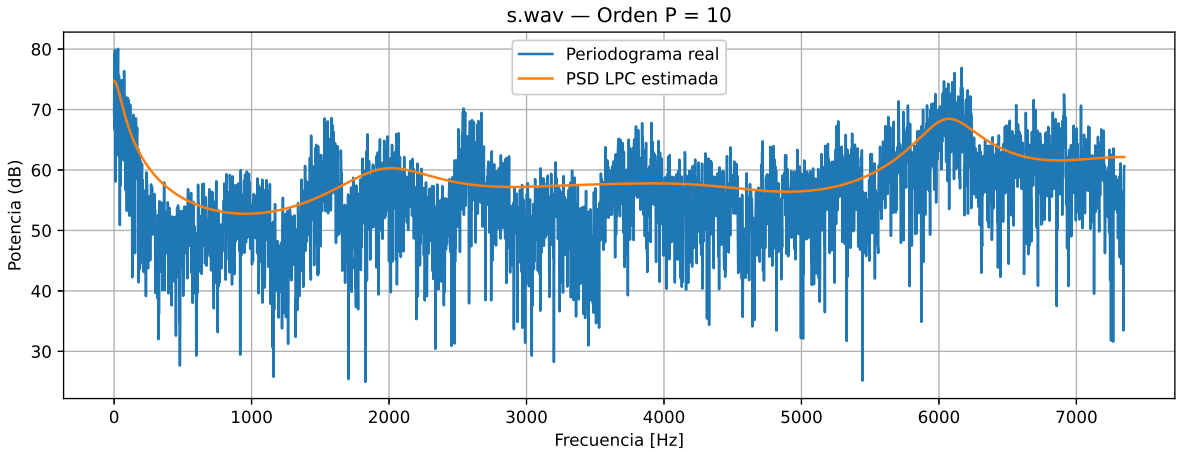

Figura 8: PSD real superpuesta a la obtenida con el modelo de orden 10

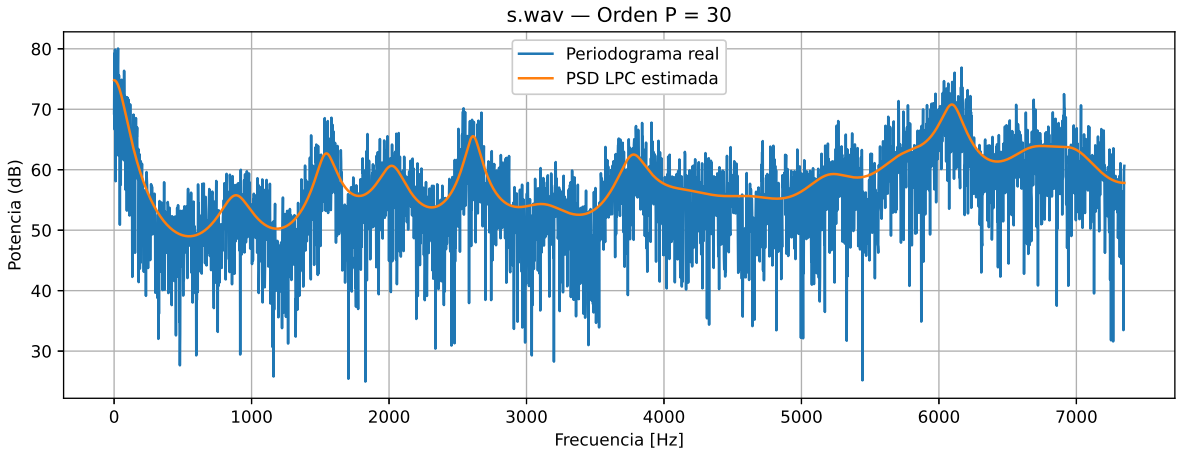

Figura 9: PSD real superpuesta a la obtenida con el modelo de orden 30

4.4. Sonido “sh”

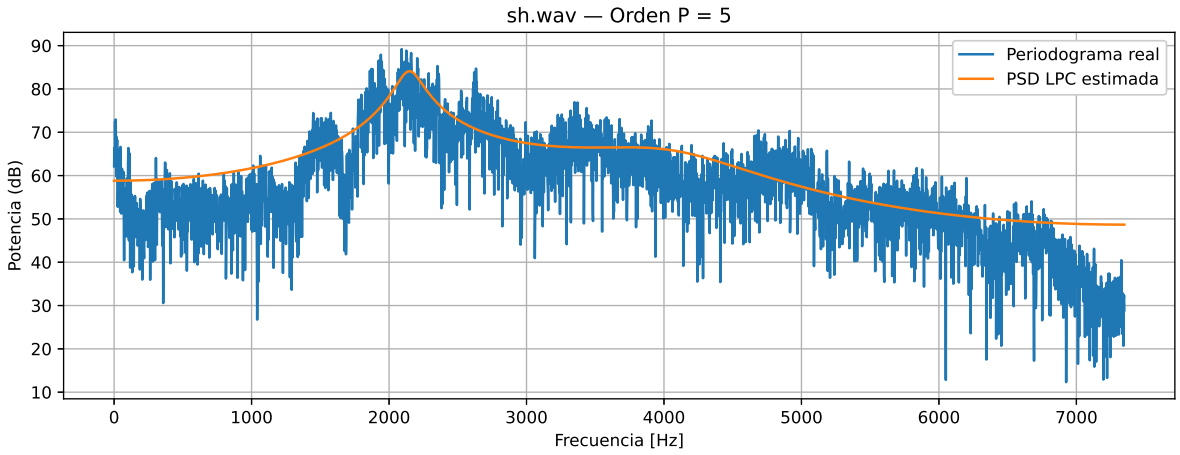

Figura 10: PSD real superpuesta a la obtenida con el modelo de orden 5

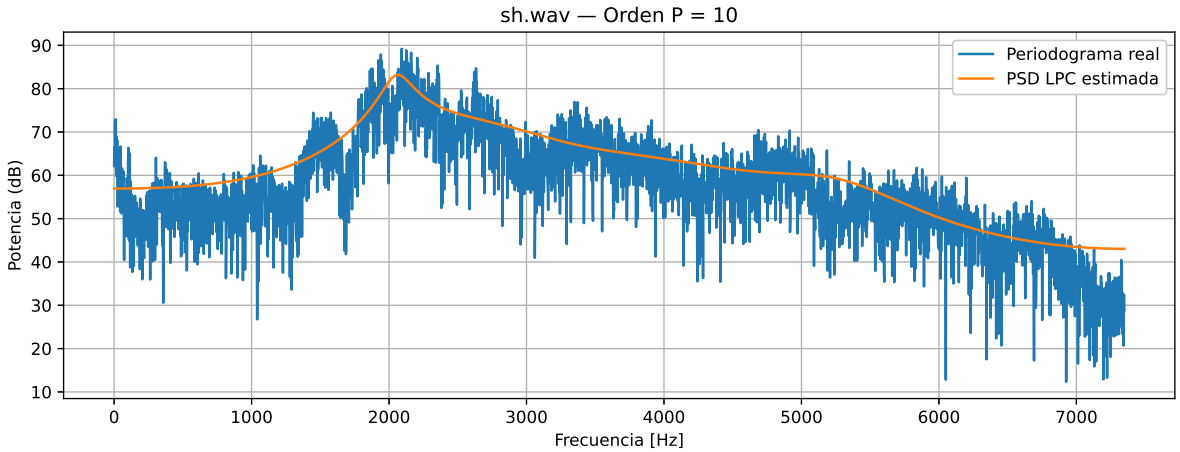

Figura 11: PSD real superpuesta a la obtenida con el modelo de orden 10

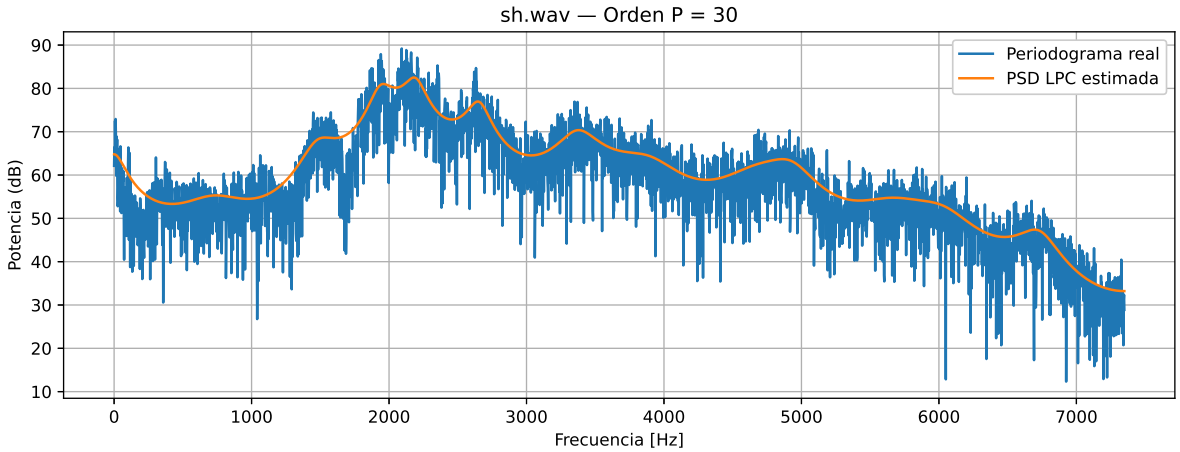

Figura 12: PSD real superpuesta a la obtenida con el modelo de orden 30

4.5. Análisis de resultados

En el caso de los fonemas sonoros el ajuste es cercano para todos los ordenes, sin embargo se denota una clara mejora al aumentar la cantidad de coeficientes. El ajuste de la ganancia es considerablemente más preciso y el comportamiento fuera de los picos es menos brusco.

Sin embargo, el efecto de variar el orden tiene mayor peso en los fonemas sordos. En estos se ve claramente como la forma estimada con el orden 5 sigue la forma general de la onda pero no la de sus picos. Al aumentar el orden el seguimiento mejora notablemente, en el orden 30 se ve como la curva se ajusta a los picos principales de la señal ademas de seguir de forma mas fiel la señal general.

5. Detección de pitch

Como se mencionó anteriormente, es necesario para la sinterización de los fonemas sonoros, no solo estimar los coeficientes del LPC, sino que también detectar la frecuencia fundamental con el objetivo de reconstruir el mismo sin perder la entonación del audio original.

Para ello es posible modelar el residuo LPC o la autocorrelacion del error de predicción $R_e(k)$. Se modela este parámetro ya que la señal original de voz incluye una excitación periódica y una resonancia del tracto vocal (los formantes), que crean picos adiciones en el espectro, esto hace que la autocorrelacion de la señal completa tenga varios picos grandes y sea difícil de saber cual corresponde al pitch real, utilizando el residuo LPC es posible eliminar esta resonancia y conservar la periodicidad del sonido.

A continuación se resume los pasos desarrollados para la implementación del detector:

- Calculo de la señal de error: Se calcula el error de predicción computando 6

- Autocorrelación de la señal de error: Se calcula la autocorrelación del error y se normaliza con su valor $R_e(0)$ para independizarlo de su amplitud.

- Detección de picos: Se implementa una función que busca el segundo pico significativo en la autocorrelacion dentro de un rango. utilizando un umbral α y calculando la frecuencia en la que se encuentra mediante 7

Al momento de realizar la detección de la frecuencia del pitch se investigo el rango de frecuencias de la voz humana. Se limito la búsqueda a ese rango para disminuir la complejidad computacional del problema y evitar picos espurios. El rango de frecuencias considerado fue $[80Hz, 400Hz]$.

El valor de α fue elegido observando el gráfico de la auto-correlación del error de las vocales. En ambos se puede observar que el valor del segundo pico es siempre mayor a 0.4, por lo tanto se elige un valor apenas menor para tener un margen de error y una seguridad para el resto de las vocales. Teniendo en cuenta todas estas consideraciones se eligió usar $\alpha = 0{,}4$.

El orden elegido para el LPC fue P = 20. Se considero que sea suficiente para poder diferenciar de manera clara los picos y donde ya no se perciba una diferencia notable al aumentarlo, ya que esto elevaría la complejidad del problema sin presentar un beneficio perceptible.

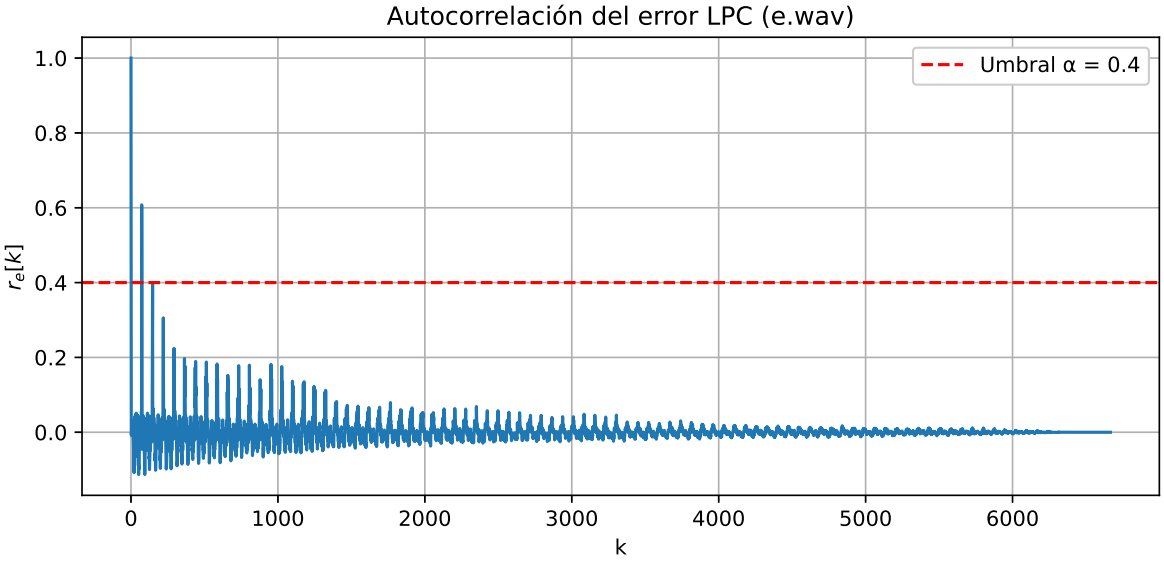

Figura 14: Detección de pitch con autocorrelacion de fonema “e”

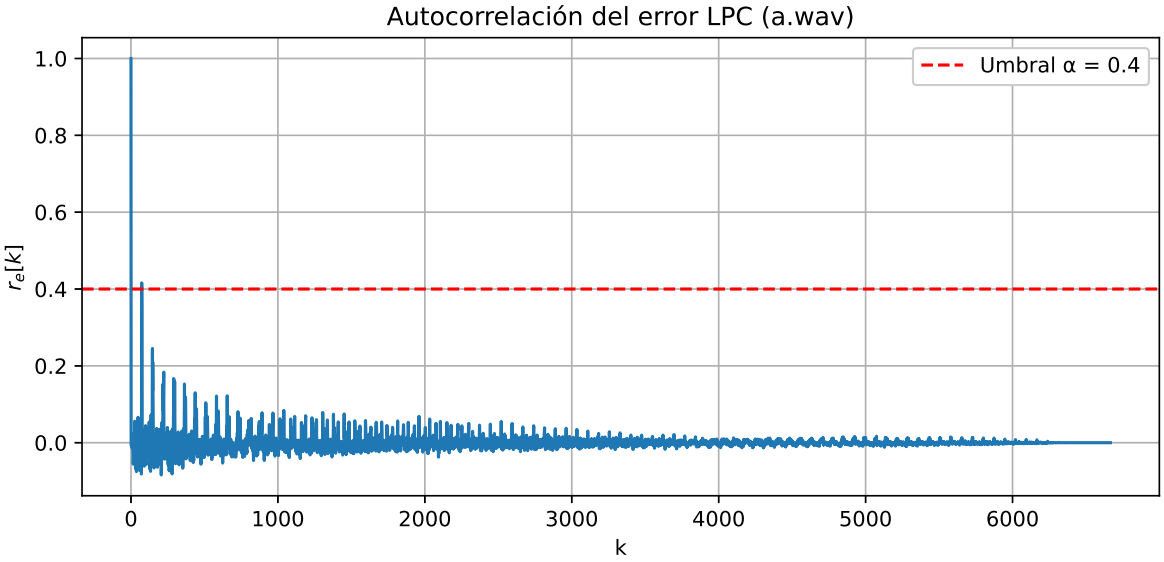

Figura 13: Detección de pitch con autocorrelacion de fonema “a”

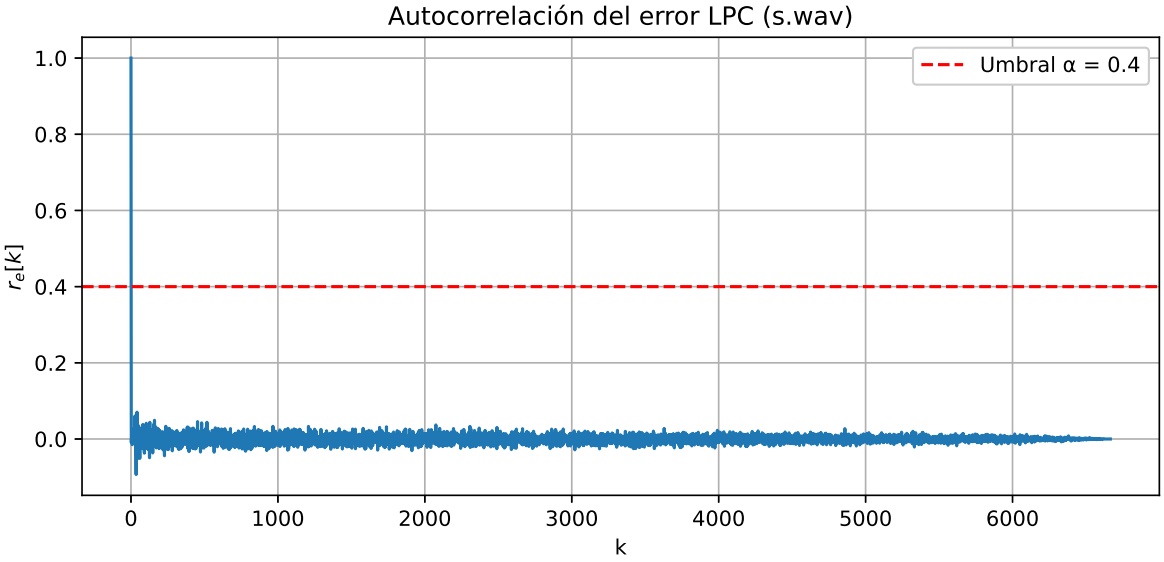

Figura 15: Detección de pitch con autocorrelacion de fonema “s”

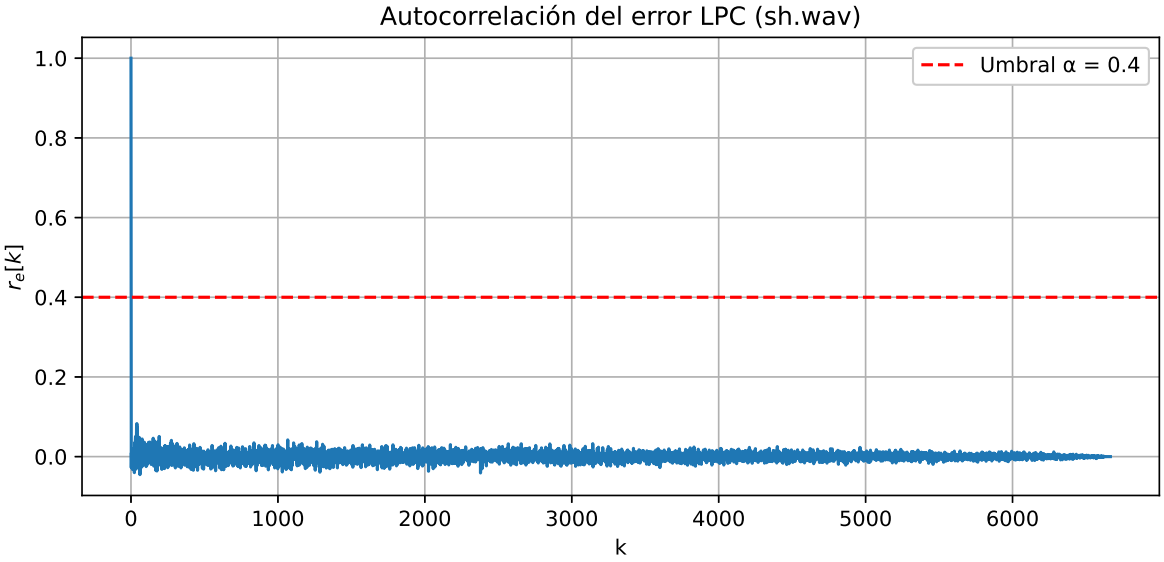

Figura 16: Detección de pitch con autocorrelacion de fonema “sh”

En las Figuras 13, 14, 15 y 16 se puede apreciar como en las vocales el segundo pico supera el umbral elegido y en los fonemas sordos se observa un ruido blanco. Además se observa el valor de los segundos picos en las vocales que fue usado como criterio para elegir donde se ubicó el umbral.

5.1. Estimación de PSD con pitch detectado

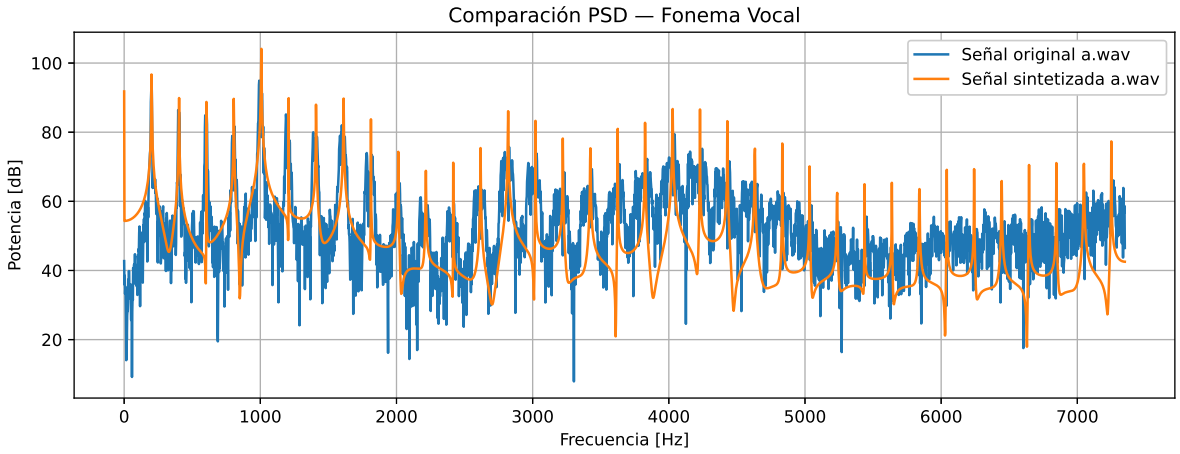

Figura 17: PSD obtenida con el pitch detectado

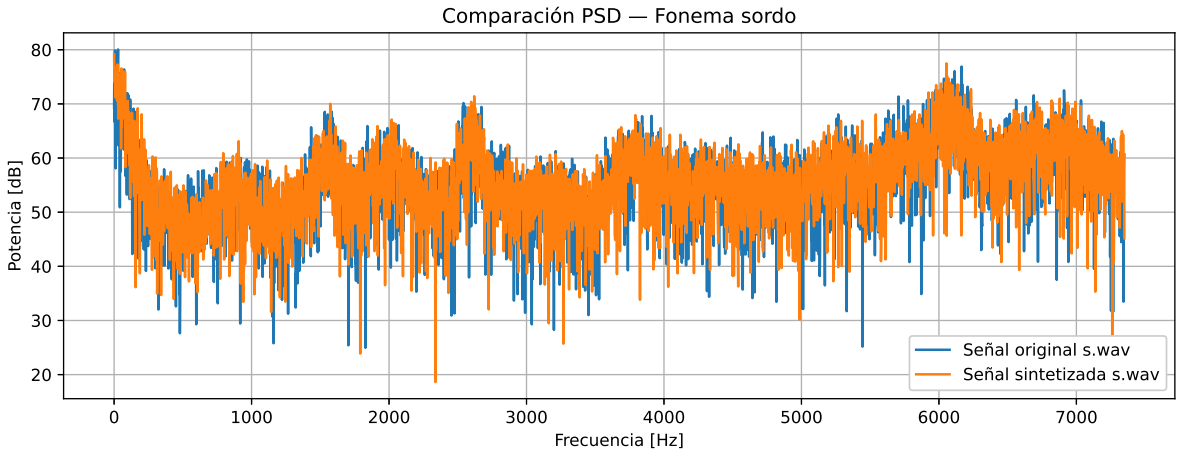

Figura 18: PSD de fonema no vocal

La frecuencia obtenida con la función *pitch_lpc* fue de 201.37Hz para ambas vocales, y como se esperaba 0Hz para los fonemas sordos. La frecuencia obtenida es cercana a la usada anteriormente, por lo que se puede ver que la estimación ajusta también de buena forma a la señal real.

6. Codificación y decodificación

A partir de las pruebas realizadas anteriormente para el correcto funcionamiento la implementación del detector de pitch y estimador de modelo LPC, se procede a realizar una implementación de una codificación y decodificación completa a partir de audios mas extensos provistos por la cátedra.

Al momento de codificar el audio se debió considerar que el mismo no es un proceso estacionario. Sin embargo, para poder trabajarlo como si lo fuera se segmento la señal en pequeños fragmentos en donde puedan ser considerados estacionarios y de esa manera usar las simplificaciones de un proceso estacionario.

Teniendo esto en cuenta se resume la implementación de la codificación:

- Segmentación: La señal de voz se divide en segmentos cortos solapados en una duración de unos pocos milisegundos. La duración de cada segmento debe considerar el compromiso de mantenerse corto para la valides de la cuasi-estacionariedad, y lo suficientemente largo para capturar varios ciclos del tono fundamental respectivo.

- Ventaneo: Se opta por multiplicar cada segmento por una ventana de Hamming por sugerencia de la cátedra.

- Extracción de parámetros LPC: A cada segmento ventaneado se le extraen los coeficientes y ganancia para el modelo LPC y el pitch respecto en caso de ser detectado como sonoro

Un receptor que quiera reconstruir la señal puede utilizar el siguiente algoritmo implementado para la decodificación:

- Síntesis iterativa de segmentos: La función de síntesis procesa cada conjunto de parámetros obtenidos de la codificación de forma secuencial.

- Selección de variante de tono: A modo de experimentación se opta por sintetizar con 4 tonos distintos, el estimado por el detector en cada segmento a partir α, un tono fijo e invariante en todos los segmentos, sin tono para forzar la excitación a un ruido blanco y un tono sintético que varia en cada segmento adaptándose a la frecuencia fundamental respectiva.

- Filtrado IIR: Se selecciona la señal de excitación correspondiente al tipo de fonema detectado y con los programaros extraídos de la síntesis se arma el filtro con 8 para multiplicarlo con la excitación y obtener el segmento reconstruido

- Superposición: Se multiplica el segmento reconstruido por una ventana de Hamming con los mismos parámetros con los que se codificó.

Se decidió usar una ventana de Hamming con un largo de 30 ms. El largo de la ventana fue elegido experimentalmente. Inicialmente se uso uno pequeño de 5 ms pero la reconstrucción del audio no resultaba satisfactoria, al aumentar el tamaño de la ventana el resultado mejoro considerablemente. Además para suavizar la transición entre cada segmento se utilizo un solapamiento del 50 %

6.1. Reconstrucción con pitch estimado

Para este caso, al reconstruir la señal se utilizó el pitch estimado para cada fragmento.

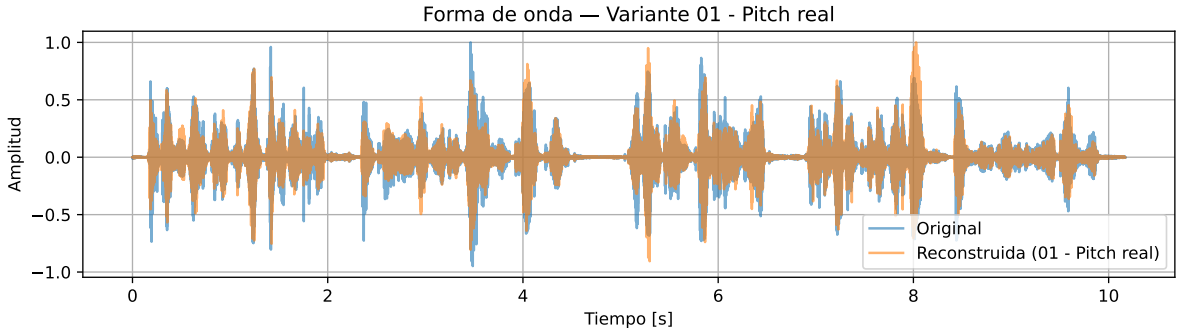

Figura 19: Superposición de señal real y reconstruida en el tiempo

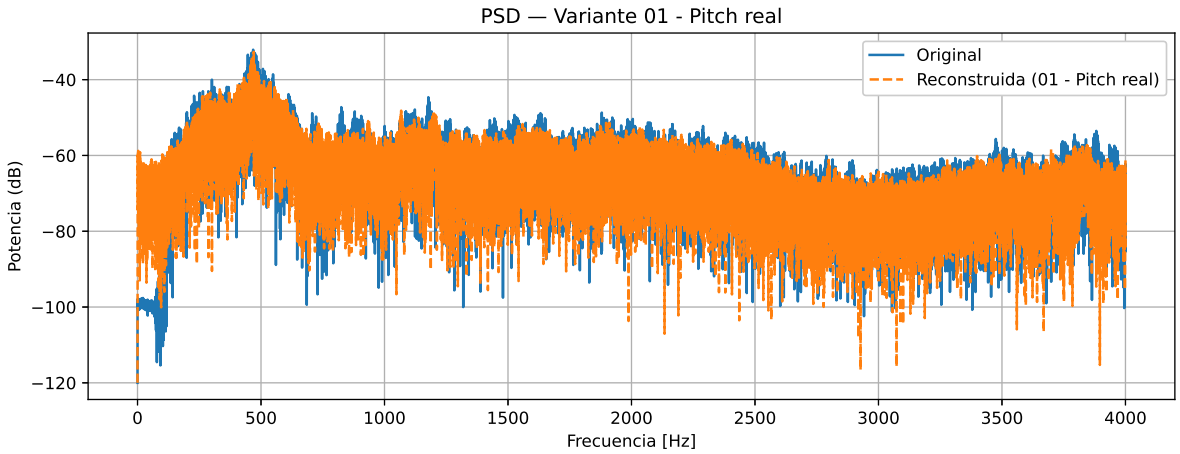

Figura 20: Superposición de PSD real y reconstruida

En este caso se puede ver que, si bien ni la PSD ni la señal en el tiempo son coincidentes en todos los puntos, en el caso de la señal en el tiempo los máximos de amplitud coinciden, y en la PSD la forma general, ignorando el ruido, es similar.

Escuchando el audio de la señal reconstruida, es posible entender el relato de la persona, si bien el sonido no es idéntico.

6.2. Reconstrucción con pitch fijo

En este caso se utilizó un pitch fijo de 200 Hz para todos los fragmentos.

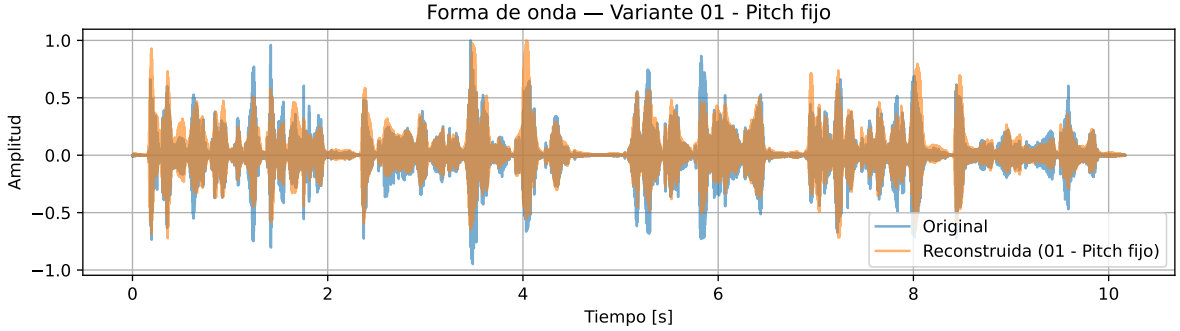

Figura 21: Superposición de señal real y reconstruida en el tiempo

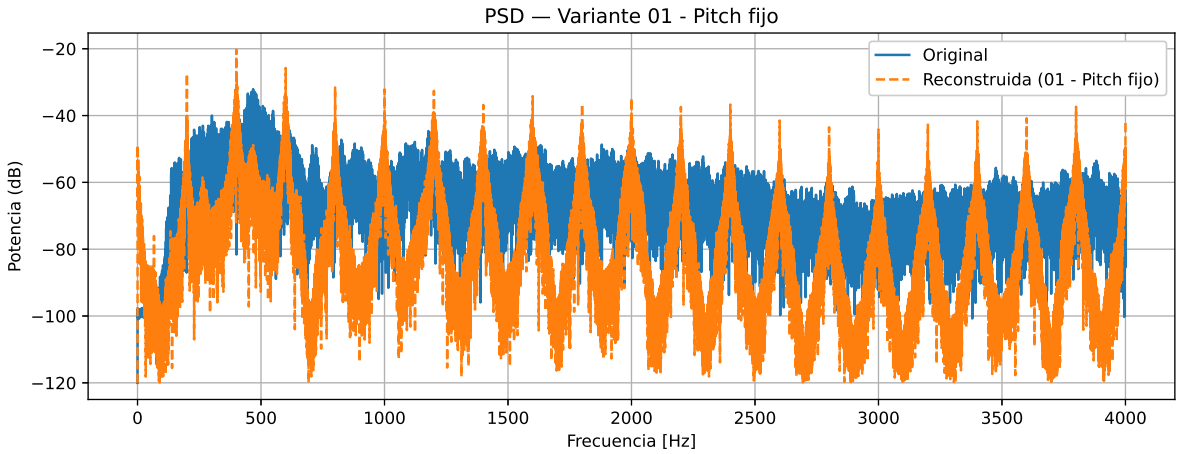

Figura 22: Superposición de PSD real y reconstruida

En la Figura 22 se puede ver claramente el efecto de usar un pitch fijo, en este caso los gráficos no coinciden en lo absoluto, si bien la forma de onda temporal pueda ser similar.

Esto resulta en que el audio obtenido mantenga el sonido de la voz diferenciable, pudiendo entender el relato, pero con un sonido *robótico*.

6.3. Reconstrucción sin pitch

Ahora en vez de utilizar un pitch para la reconstrucción de los fragmentos, se utiliza ruido blanco en todos los casos.

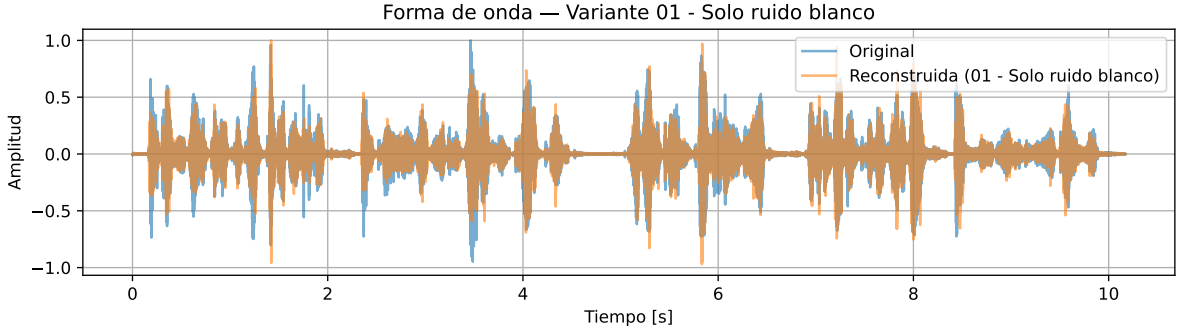

Figura 23: Superposición de señal real y reconstruida en el tiempo

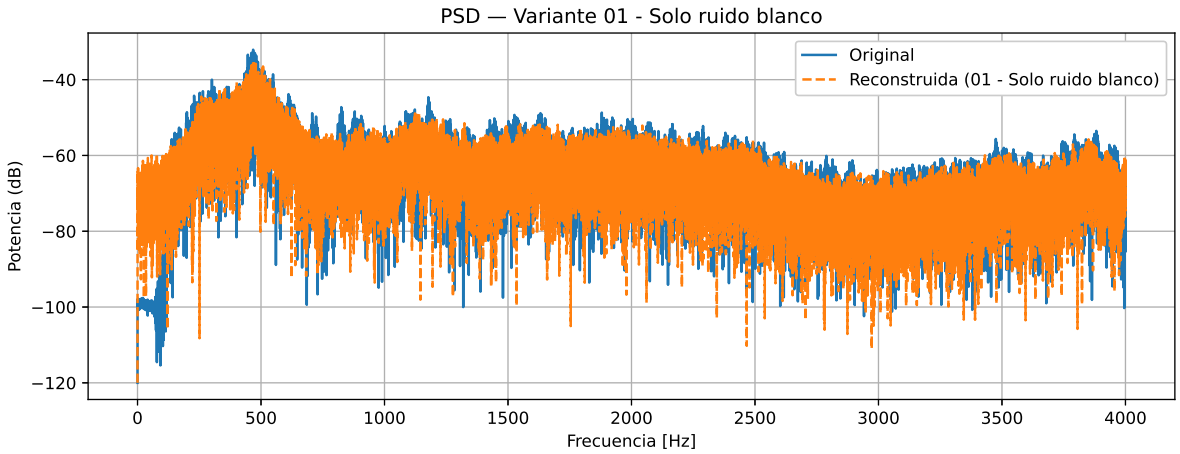

Figura 24: Superposición de PSD real y reconstruida

En las figuras de este caso se puede ver que no se presenta demasiada diferencia entre el reconstruido y el real. Sin embargo, al escuchar el audio reconstruido se nota una clara diferencia, el no usar pitch y reconstruir la señal solamente con ruido blanco resulta en una voz similar a la que utiliza al susurrar.

6.4. Reconstrucción con pitch sintético

Para este ultimo caso se utilizó un pitch artificial generado por el fragmento de código proporcionado por los docentes.

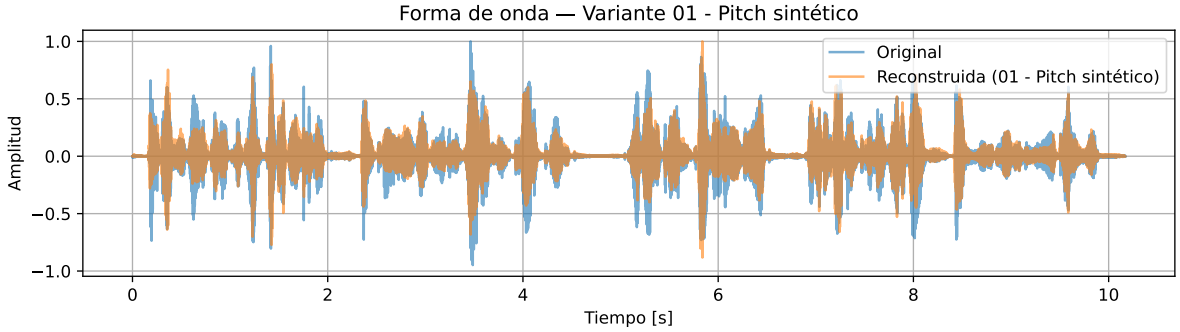

Figura 25: Superposición de señal real y reconstruida en el tiempo

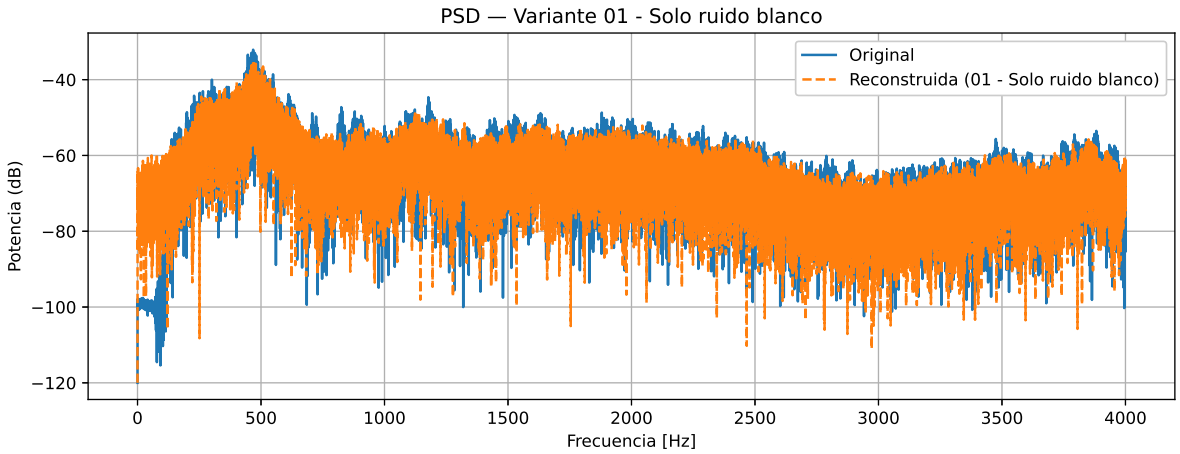

Figura 26: Superposición de PSD real y reconstruida

Nuevamente en este caso no se presenta amplia diferencia en los gráficos, pero al escuchar el audio reconstruido es posible apreciar el efecto del pitch variable.

7. Conclusiones

A pesar de que los audios sintetizados no sean exactamente iguales a los originales, se puede identificar el relato de manera clara en todos los casos. Se observó el efecto de utilizar pitchs obtenidos desde distintas fuentes en los audios.

Se comprobó que el rango de ventaneo en este tipo de estimación no puede ser tan chico porque es posible que no hayan suficientes periodos muestreados, disminuyendo la resolución espectral.

Comenzando por la obtención de los coeficientes del modelo LPC y estimando el pitch de los fragmentos se pudo codificar audios a través de un ventaneo. Luego con los datos obtenidos se pudieron reconstruir los fragmentos, con suficiente exactitud para entender los relatos en estos.