Universidad de la República

Facultad de Ingeniería

Procesamiento Digital de Señales de Audio

Práctica 4

Author:
Julian O'Flaherty

Número de hojas: 14
15 de junio de 2022

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Índice

1. Ejercicio 1 - Modelo de producción de voz

1.1. Parte 1

1.1.1. Cepstrum complejo para sonidos sonoros

Los sonidos sonoros se modela con un tren de pulsos periódico $p[n]$ como excitación del tracto vocal:

$$p[n] = \beta^n \sum_{k=0}^{\infty} \delta[n - kP] \tag{1}$$

Se desea calcular el cepstrum complejo $\hat{p}[n]$. El cepstrum complejo se define como:

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log\left(|X(e^{jw})|\right) e^{jwn} dw = \mathcal{Z}^{-1}\left\{\log\left(|X(z)|\right)\right\}[n] \tag{2}$$

Comencemos calculado la transformada $\mathcal{Z}$ del tren de pulsos.

$$\mathcal{Z}\{p[n]\}(z) = P(z) = \mathcal{Z}\left\{\beta^n \sum_{k=0}^{\infty} \delta[n - kP]\right\} = \sum_{n=-\infty}^{\infty} \sum_{k=0}^{\infty} \delta[n - kP]\beta^n z^{-n} = \sum_{k=0}^{\infty} \beta^{kP} z^{-kP}$$

$$\mathcal{Z}\{p[n]\}(z) = \frac{1}{1 - \beta^P z^{-P}} \tag{3}$$

Para el cepstrum complejo tenemos que tomar el logaritmo complejo de la expresión 2:

$$\log\left(|P(z)|\right) = -\log(1 - \beta^P z^{-p}) = \sum_{n=1}^{\infty} \frac{\beta^{nP}}{n} z^{-nP}$$

Podemos reescribir esta expresión para que aparezca la transformada de un tren de impulsos.

$$\log\left(P(z)\right) = P \sum_{n=1}^{\infty} \frac{\beta^{nP}}{nP} z^{-nP} = \sum_{n=1}^{\infty} \left(P \sum_{k=1}^{\infty} \delta[n - kP] \frac{\beta^n}{n} \right) z^{-n} \tag{4}$$

Por lo tanto, tomando la anti-transformada $\mathcal{Z}$ de la ecuación 2. Obtenemos:

$$\hat{p}[n] = P \frac{\beta^n}{n} \sum_{k=1}^{\infty} \delta[n - kP] \tag{5}$$

En la figura 1 se muestra el cepstrum del peine.

1.1.2. Cepstrum complejo secuencia $h[n]$

Se tiene una secuencia $h[n]$ con transformada $\mathcal{Z}$ conocida:

$$H(z) = \frac{(1 - bz)(1 - b^* z)}{(1 - cz^{-1})(1 - c^* z^{-1})}, \quad \text{con } |b|, |c| < 1$$

Tomando el logaritmo, podemos separar la expresión en sumas.

$$\log\left(H(z)\right) = \log(1 - bz) + \log(1 - b^* z) - \log(1 - cz^{-1}) - \log(1 - c^* z^{-1})$$

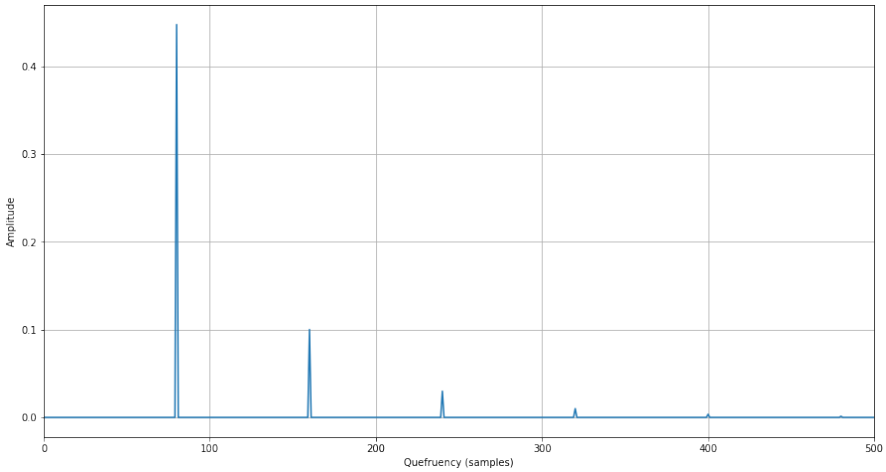


Figura 1: cepstrum del peine de deltas.

Y desarrollar los logaritmos en series de potencia

$$\log\left(H(z)\right) = \sum_{k=1}^{\infty}\left(b^n + (b^*)^n\right)\frac{z^n}{n} - \sum_{k=1}^{\infty}\left(c^n + (c^*)^n\right)\frac{z^{-n}}{n}$$

Identificando la forma de la transformada z, podemos tomar la anti-transformada del logaritmo y obtener el cepstrum complejo.

$$\begin{aligned}\hat{h}[n] &= \left(\frac{b^{-n}}{n} + \frac{(b^*)^{-n}}{n}\right)u[1-n] + \left(\frac{c^n}{n} + \frac{(c^*)^n}{n}\right)u[n-1] \\ \hat{h}[n] &= \frac{2}{n}|b|^{-n}cos(n\theta_b)u[1-n] + \frac{2}{n}|c|^n u[n-1]\end{aligned} \tag{6}$$

En la figura 2 se muestra el cepstrum del filtro $h[n]$.

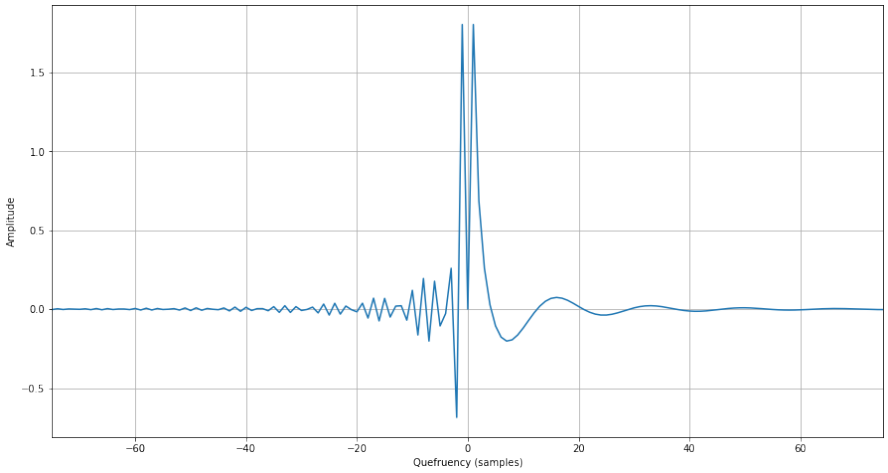


Figura 2: cepstrum del filtro

1.1.3. Cepstrum complejo $h * p[n]$

Como en las partes anteriores comenzamos tomando la transformada $\mathcal{Z}$ de la señal $s[n] = h * p[n]$:

$$\mathcal{Z}\{s[n]\}(z)\mathcal{Z}\{h * p[n]\}(z) = \mathcal{Z}\{h[n]\}(z)\mathcal{Z}\{p[n]\}(z) = H(z)P(z)$$

Tomando el logaritmo de la transformada:

$$\log\left(\mathcal{Z}\{s[n]\}\right) = \log(H(z)) + \log(P(z))$$

Tomando la anti-transformada:

$$\hat{s}[n] = \hat{h}[n] + \hat{p}[n] = \left(\frac{b^{-n}}{n} + \frac{(b^*)^{-n}}{n}\right)u[1-n] + \left(\frac{c^n}{n} + \frac{(c^*)^n}{n}\right)u[n-1] + P\frac{\beta^n}{n}\sum_{k=1}^{\infty}\delta[n-kP]$$

En la figura 3 se muestra el cepstrum de la convolución $h * p[n]$.

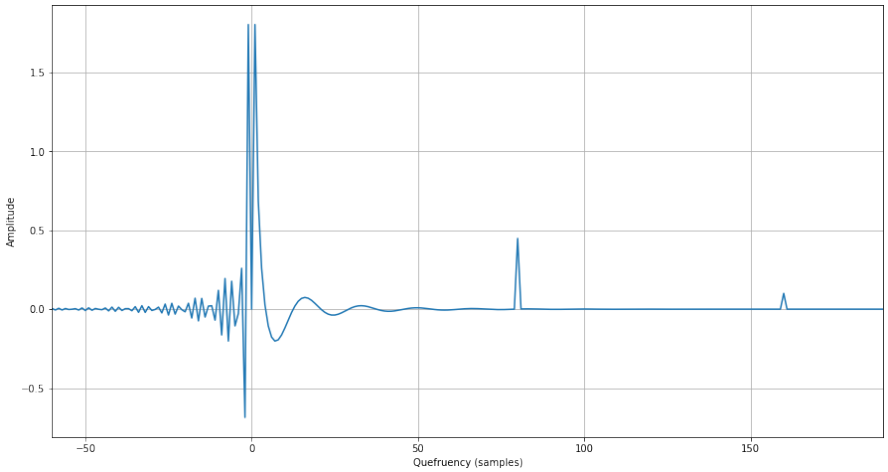


Figura 3: cepstrum de la convolución

1.1.4. Extracción respuesta del filtro

Se quiere extraer la respuesta al impulso del filtro a partir de la señal convolucionada. Este problema surge naturalmente en análisis de voz, dado que puede ser de interés extraer la respuesta de las cavidades resonantes de la voz. El cepstrum es muy poderoso en este sentido, ya que nos separa la convolución en una suma, y si el filtro y la señal original no se superponen (o lo hacen en cuefrencias que no contienen mucha información), basta con un simple liftrado para extraer el cepstrum del filtro y aplicarle el cepstrum inverso para obtener $\hat{h}[n]$.

En la figura 3 se observa que el cepstrum de la convolución cumple lo anteriormente mencionado, por lo que eligiendo una cuefrencia de corte menor que P podemos extraer el filtro. En la figura 4 se muestra el cepstrum de la convolución con las cuefrencias de corte elegidas. En este caso se uso $0{,}9P$ como cuefrencia de corte.

Se aplico un liftrado pasa-bajos, es decir, se llevo a 0 todas las cuefrencias mayores en valor absoluto a la quefrencia de corte. Si le aplicamos la transformada inversa al cepstrum extraído se obtiene la respuesta al impulso $\hat{h}[n]$ mostrada en la figura 5 se muestra el resultado de la extracción.

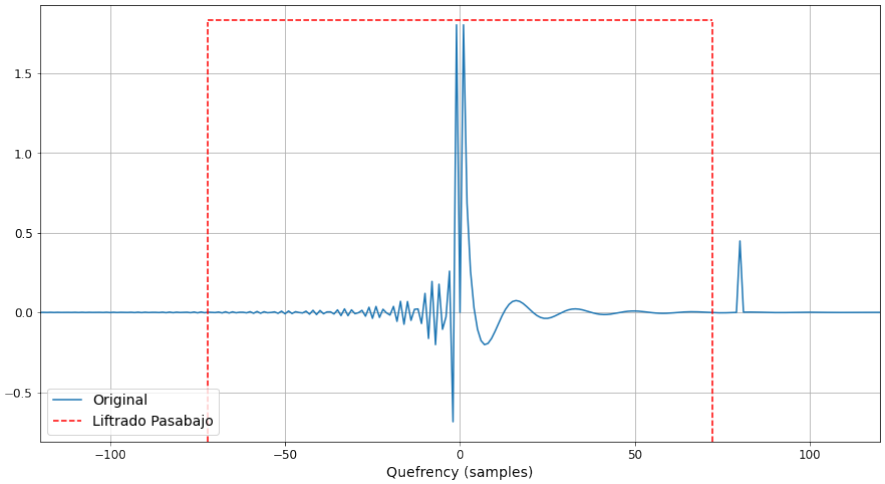


Figura 4: cepstrum de la señal s con liftrado pasa-bajos superpuest

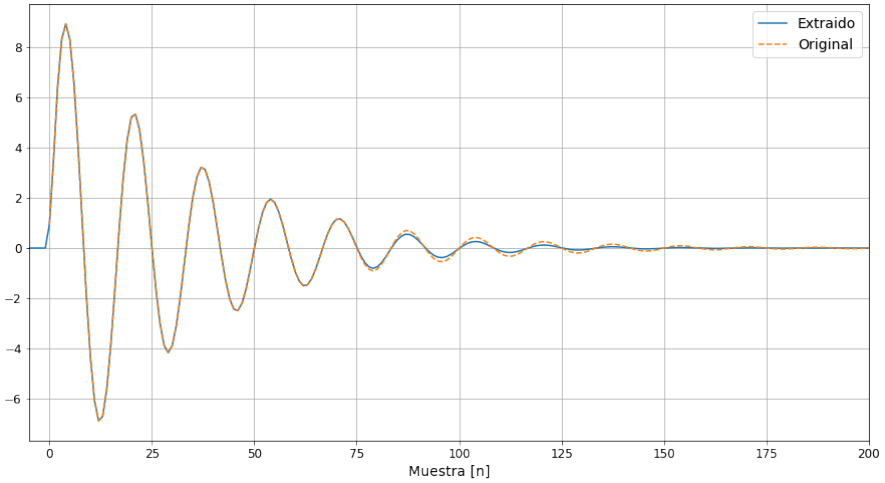


Figura 5: comparación de la respuesta del filtro original con la extraída.

Se observa que la estimación es muy buena para las primeras muestras, notándose diferencias recién a partir de la muestra 75. De allí en adelante la estimación n es buena, teniendo un desfase notable en los con la respuesta original. Sin embargo, como la amplitud de la respuesta decayó mucho cuando se comienzan a ver estos imperfectos, el impacto que tienen es pequeño.

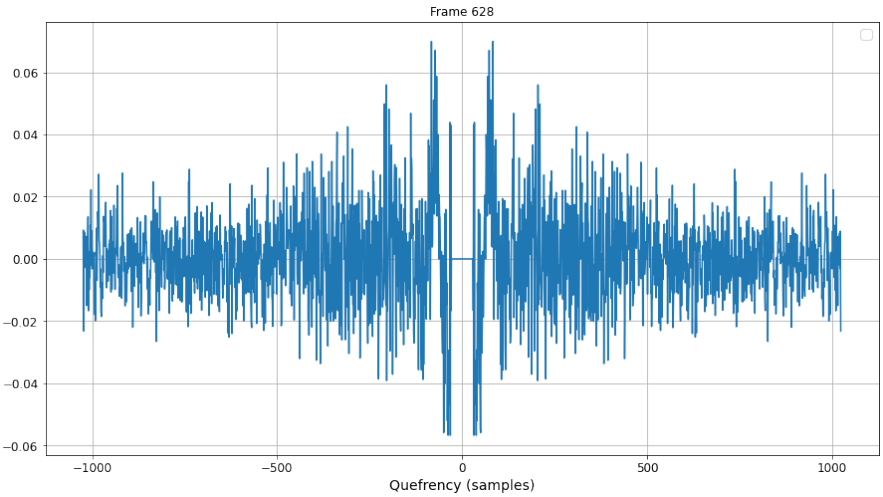
1.2. Estimación de frecuencia fundamental

Otra aplicación del cepstrum es la estimación de la frecuencia fundamental de la voz. Así como se realizo un litrado pasabajos en la parte anterior para obtener la respuesta del filtro, podemos aplicar un liftrado pasa-altos para extraer la excitación. Detectando la frecuencia de los pulsos, podemos obtener la frecuencia fundamental de la voz que está hablando.

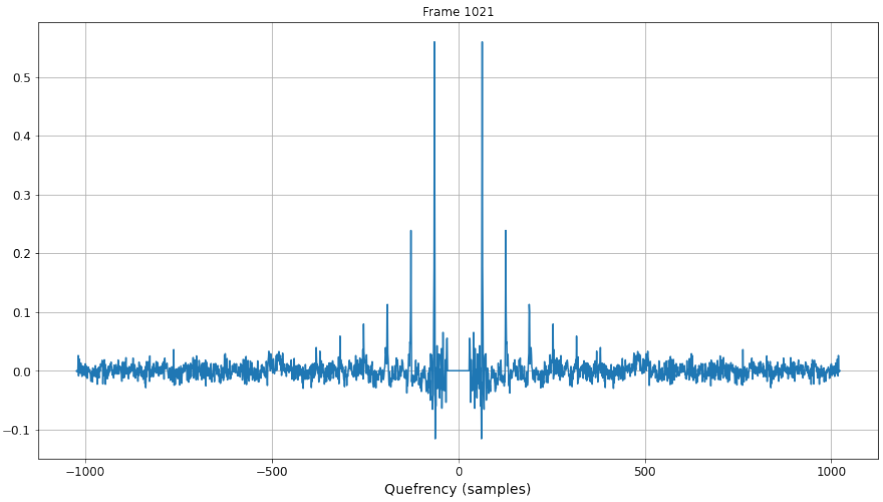
1.2.1. Cepstrum de tiempo corto

Para analizar una señal de audio e ir estimando la frecuencia fundamental debemos realizar un análisis de tiempo corto. El proceso es igual que el caso de la STFT. Se define un largo L de ventana y un salto de R muestras entre ventanas. Se va tomando tramas de la señal y enventanando con una ventana suavizante, y se aplica el cepstrum real sobre la trama enventanada.

Enventanar la señal produce replicas de la respuesta del filtro en los múltiplos de la frecuencia del peine, por lo que obtendremos una versión ruidosa del peine al liftrar. Sin embargo, como lo que buscamos es la quefrencia a la que se da el primer pico, este no es un problema que afecte la aplicación.



(a) trama de la señal con sonido sordo



(b) trama de la señal con sonido sonoro

Figura 6: cepstrum real de dos tramas de la señal

En la figura 6 se muestra el resultado del cepstrum de tiempo corto en 2 tramas. Para estas tramas se utilizó un largo de ventana $L = 2048$, salto $R = 256$ y quefrencia de corte $qc = 30$. En

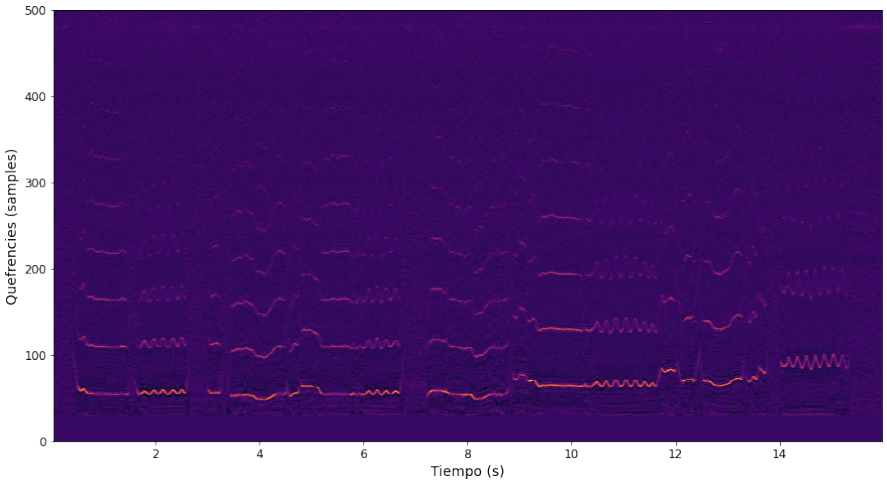


Figura 7: cepstograma del audio *LP-mem-6*

la figura 6b se logra observar los picos del tren de impulsos que genera funciona de excitación del tracto vocal, por lo que podemos tomar la ubicación del primer pico para estimar la frecuencia fundamental. En la figura 6a se muestra un frame de silencio, o en presencia de una consonante fricativa, por lo que no se tiene un tren de pulsos como exitación. Es importante notar que la amplitud de los picos en este tipo de señales es mucho menor que el de los picos, y que los picos son muchos más distinguibles. Esto nos permitirá discernir entre el tipo de sonido, que será útil para estimar la frecuencia fundamental.

Como es un análisis de tiempo corto, tiene sentido hacer un gráfico que muestre la evolución temporal de las quefrencias. En la figura 7 se muestra el resultado del análisis cepstral de tiempo corto, resultando en un cepstograma. Notar que puede verse el trazo de la frecuencia fundamental invertido, dado que las quefrencias son proporcionales al tiempo, y los momentos de silencio.

1.2.2. Estimar f_0

Como último paso nos queda estimar la frecuencia fundamental a partir del análisis de tiempo corto. Viendo el cepstograma de la figura 7 se puede observar claramente la traza de la frecuencia fundamental, por lo que tomando el máximo (y ajustando la quefrencia de liftrado) es suficiente para las zonas sonoras. Queda definir un criterio para poder detectar las zonas de silencio e imponer $f_0 = 0$, es decir, ausencia de frecuencia fundamental.

Se empezó probando con un umbral fijo, pero se noto que había zonas de ruido donde del valor del pico era mayor que él umbral, y subirlo hacía que momentos sonoros con bajo volumen fueran llevados a 0. Se terminó optando por un umbral adaptivo, donde se selecciona el umbral en cada trama basándose en el promedio y la desviación estándar. Como el pico es un dato muy atípico, podemos definir el criterio de una forma estadística. La trama n-ésima es sonora si su máximo tiene un valor mayor que $\langle c_n \rangle + tol \cdot \sigma_{c_n}$, donde $\langle x_n \rangle$ es el promedio del cepstrum la trama, σ_{c_n} la desviación estándar y tol un parámetro que regula a cuán atípico tiene que ser el pico para ser detectado como tal.

Aún con este criterio es necesario agregar un filtro de media móvil de orden 3, dado que aún sigue habiendo picos aislados detectados en zonas sordas.

En la figura 8 se muestra el resultado de la estimación de f_0. Se aprecian ciertos imperfectos, en particular un pequeño retardo sobre el final de la señal. Este retardo es muy pequeño, de

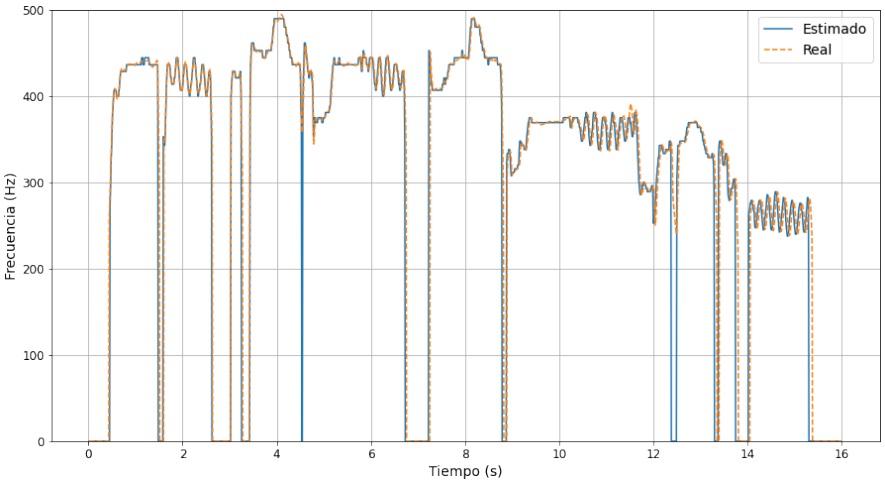


Figura 8: estimación de la frecuencia fundamental

apenas unos milisegundos, por lo que no afecta de forma destructiva la aplicación.

2. Ejercicio 2 - Predicción lineal

2.1. Modelado del problema

Se quiere definir un modelo auto-regresivo de orden p para la señal, donde la n-ésima muestra depende de la p muestras anteriores (ver ec 7).

$$s[n] = \sum_{k=0}^{p} \alpha_k s[n-k] \tag{7}$$

Obtendremos los coeficientes α_k minimizando el error cuadrático medio. Definimos el error cuadrático medio como:

$$E_n = \sum_m e_n^2 = \sum_m (s_n[m] - \tilde{s}_n[m])^2 = \sum_m \left(s_n[m] - \sum_{k=1}^{p} \alpha_k s_n[m-k] \right)^2 \tag{8}$$

Para minimizarlo, tomamos las derivadas respecto a $\alpha_i \quad \forall i = 1, \ldots, p$ e igualamos a 0:

$$\frac{\partial}{\partial \alpha_i} E_n = \sum_m 2 s_n[m-i] \left(s_n[m] - \sum_{k=1}^{p} \alpha_k s_n[m-k] \right) = 0$$

Por lo tanto:

$$\sum_{k=1}^{p} \alpha_k \sum_m s_n[m-k] s_n[m-i] = \sum_m s_n[m] s_n[m-i] \tag{9}$$

Obtenemos entonces una ecuación de la forma de la ecuación 9 para cada $i = 1, \ldots, p$. A estas ecuaciones se las denominan ecuaciones normales.

Es de interés calcular el error mínimo al que llegamos con está solución. Para esto, comenzamos desarrollando la ecuación 2.1:

$$E_n = \sum_m s_n^2[m] - \sum_m 2 s_n[m] \sum_{k=1}^{p} \alpha_k s_n[m-k] + \sum_m \left(\sum_{k=1}^{p} \alpha_k s_n[m-k] \right)^2$$

$$E_n = \sum_m s_n^2[m] - 2 \sum_m \sum_m 2 s_n[m] \sum_{k=1}^{p} \alpha_k s_n[m-k] + \sum_m \sum_{k=1}^{p} \sum_{i=1}^{p} \alpha_k \alpha_i s_n[m-k] s_n[m-i]$$

Multiplicando la expresión 9 por α_i y sumando en i obtenemos:

$$\sum_{i=0}^{p} \sum_{k=1}^{p} \hat{\alpha}_i \hat{\alpha}_k \sum_m s_n[m-k] s_n[m-i] = \sum_{i=0}^{p} \hat{\alpha}_i \sum_m s_n[m] s_n[m-i]$$

Por lo tanto, siendo $\hat{\alpha}_k$ la de las ecuaciones normales (coeficientes a los que se obtiene el error mínimo):

$$E_n = \sum_m s_n^2[m] - 2 \sum_m s_n[m] \sum_{k=1}^{p} \hat{\alpha}_k s_n[m-k] + \sum_{k=0}^{p} \alpha_k \sum_m s_n[m] s_n[m-k]$$

Por lo que el error mínimo resulta:

$$E_n = \sum_m s_n^2[m] - \sum_{k=1}^{p} \alpha_k \sum_m s_n[m] s_n[m-k] \tag{10}$$

2.2. Clasificador de vocales

Como se mencionó anteriormente, el LPC es un método para describir la secuencia como un modelo auto-regresivo (solo polos). Esto nos permite fácilmente aproximar la envolvente de una transferencia, siempre que esta no tenga cero, o los ceros no sean necesarios de modelar para el problema.

En el caso de clasificación de vocales, nos interesa obtener las dos primeras formantes del audio. Estas dos formantes son suficiente para, con bastante precisión, discernir entre dos vocales diferentes [Listerri, 2022]. En la figura 9 se muestra la envolvente espectral y el diagrama cero polo devuelto por el LPC. En el diagrama cero-polo se marcaron en rojo los polos correspondientes a las formantes.

El desafío en la clasificación es la detección de las formantes, dado que la clasificación se realiza por proximidad a la formantes estipuladas en el documento de Joaquín Listerri[Listerri, 2022]. Para la detección se sigue un procedimiento simple:

1. **Enventanado de la señal:** se enventana la señal con una ventana suavizante de largo L. El enventanado se realiza respecto a la muestra central de la señal.

2. **LPC:** se realiza un LPC de orden p. Es aquí que se obtienen los polos de la señal. La regla del pulgar es $p \approx \frac{f_s}{1000}$

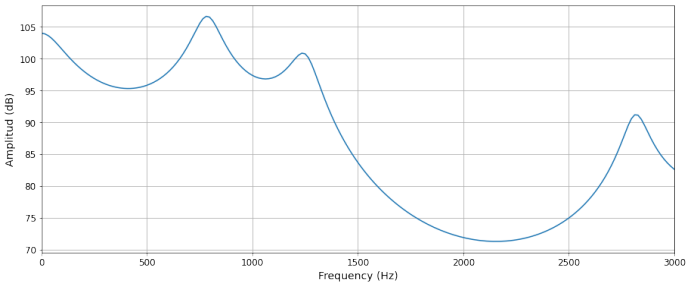
3. **Descartar polos reales y mayores a π:** se descartan polos reales, correspondientes a componentes de continua, y polos con ángulo mayor que pi, puesto que son complejos conjugados y estaríamos repitiendo la información.

4. **Descartar polos con ancho de banda grande:** se impone un umbral BWthres para el ancho de banda de un polo a partir del cual se lo descarta. El ancho de banda de un polo se estima como $BW_i = -\frac{f_s}{\pi} \log(|p_i|)$, donde p_i es el i-ésimo polo.

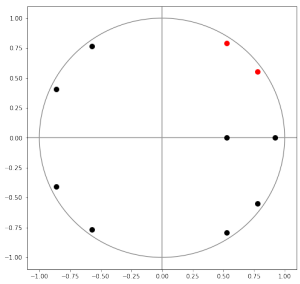
Obtenidos los polos de las formantes, se los convierte a frecuencia y se toma las dos menores como las formantes de la señal. Como se mencionó antes, la clasificación se realiza por proximidad a las formantes conocidas. El algoritmo implementado compara las formantes extraídas de la señal con las formantes de las vocales para un hombre y para una mujer, dando como predicción aquella vocal que tenga la menor distancia euclidiana. La decisión de comparar con ambas formantes viene del hecho que no se tiene información a priori si la voz es femenina o masculina, además de que dentro de los sexos hay una desviación, por lo que se cubren más casos realizando la comparación con ambas.

Hablante	**A**	**E**	**I**	**O**	**U**	**Total**
Martín	90 %	100 %	100 %	60 %	100 %	90 %
Cecilia	80 %	90 %	80 %	90 %	90 %	86 %
Total	85 %	95 %	90 %	75 %	95 %	**88 %**

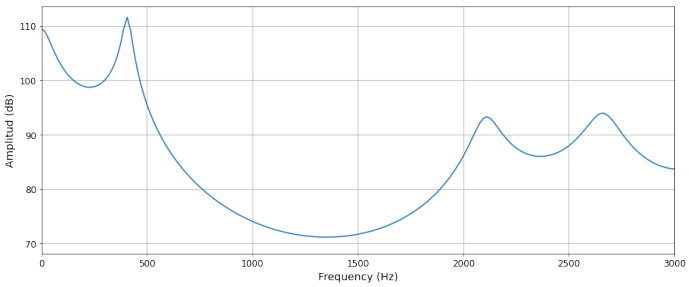
Cuadro 1: resultados clasificación



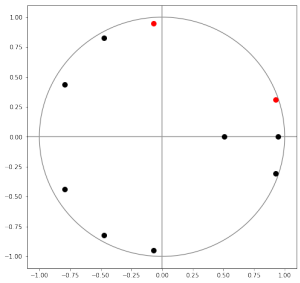
(a) envolvente A



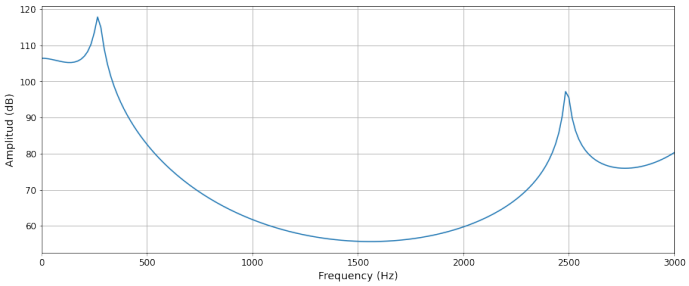
(b) Polos A



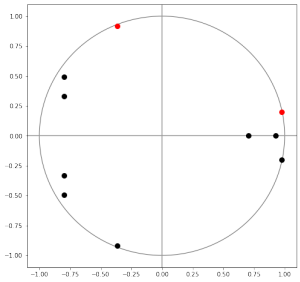
(c) envolvente E



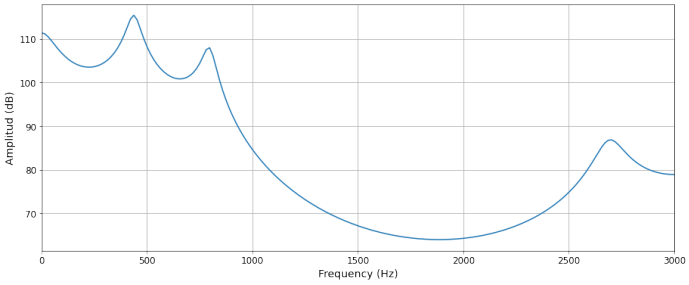
(d) Polos E



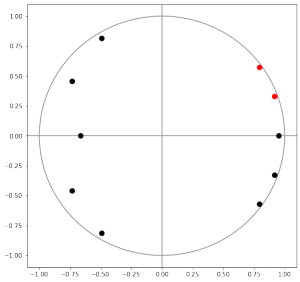
(e) envolvente I



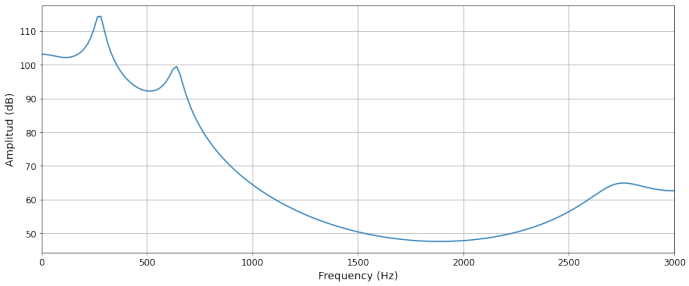
(f) Polos I



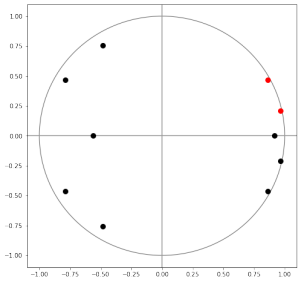
(g) envolvente O



(h) Polos O



(i) envolvente U



(j) Polos U

Figura 9: envolvente espectral y diagrama cero polo para cada vocal

En la tabla 1 se muestran los resultados obtenidos con el clasificador diseñado ($p = 9$, $BW_{trhes} = 200$, $L = 300$). Las figuras 10 y 11 se muestra los mapas de vocales para Martín y Cecilia respectivamente. El color es la etiqueta real y la letra anotada al lado la etiqueta asignada por el clasificador. Las estrellas corresponden a las formantes de referencia del hablante masculino y los pentágono del hablante femenino.

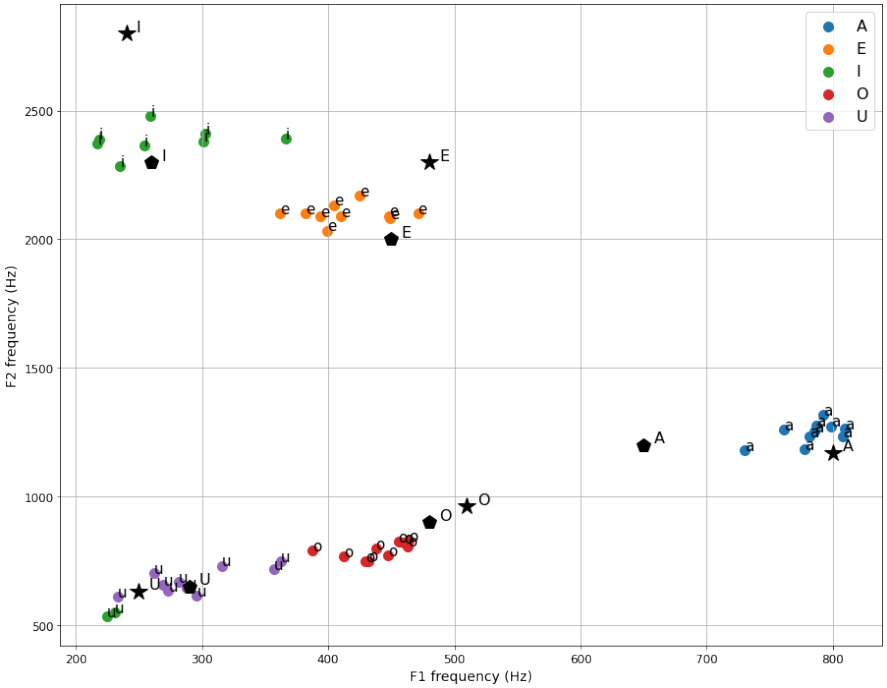


Figura 10: Mapa de vocales Martín

El peor rendimiento se obtiene con la vocal O para martín. En el mapa de vocales de la figura 10 es puede observar que las O y las U están muy cercano, por lo que varias O con frecuencia de la primer formante baja se las esta clasificando como U.

Para el caso de Cecilia, la mayor parte de los errores se dieron por una mala estimación de formantes. La voces agudas presentan un problema cuando se quiere estimar las formantes, y surge desde la generación del sonido. La voz comienza como un tren de impulsos que pasa por las cavidades resonadoras para darle "forma" al sonido que se quiere producir. Al tener una frecuencia fundamental más alta, el espaciamiento entre los impulsos es mayor, por lo que el espectro de las cavidades resonadoras tiene un muestreo menos fino, dificultando la detección de características del sonido que se produjo.

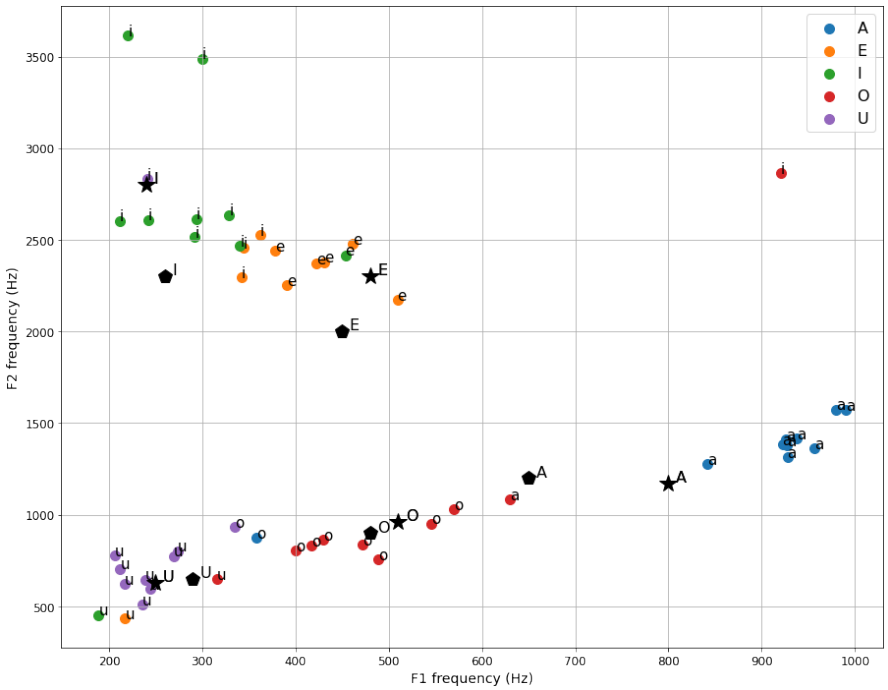


Figura 11: Mapa de vocales Cecilia

Referencias

[Listerri, 2022] Listerri, J. (2022). Las características acústicas de los elementos segmentales.