Procesamiento de Señal Hablada utilizando PRAAT

Joselyn Mayte Fernández Martínez

23 de marzo de 2023

Introducción

Para obtener muestras de señal hablada, se utilizaron un conjunto de palabras repetidas cinco veces a una tasa de muestreo de 8kHz, la grabación .wav completa así como de la segmentación por palabras se puede encontrar en el repositorio de GitHub en el link: `https://github.com/YosLab-dev/BiosignalProcessing/tree/main/PRAAT`
Para las palabras analizadas se hizo una descripción de los sonidos de cada letra para que a partir de los análisis mediante PRAAT, se identifiquen las características relevantes de cada una de acuerdo con el tipo de consonante y vocal que compone a cada palabra.

- Terremoto
  Descripción de sonido: Tiene vocales clasificadas como medias (según la apertura de los labios), si se clasifica la palabra por sílabas, la primera consonante es oclusiva, la segunda es vibrante, la tercera es nasal, la última es oclusiva.

- Erupción
  Descripción de sonido: Sus vocales forman parte del grupo de medias (para 'e' y 'o') y cerradas (para 'i' y 'u'). Por sílabas, la primera es vibrante, la segunda es fricativa y termina en nasal.

- Albergue
  Descripción de sonido: Se compone de vocales abiertas y medias. La primera sílaba tiene una consonante líquida, la segunda es oclusiva y vibrante y termina en oclusiva.

- Código
  Descripción de sonido: Las vocales de esta palabra están en la clasificación de medias y cerradas. Esta palabra en particular tiene todas sus consonantes oclusivas.

- Simulacro
  Descripción de sonido: Esta palabra tiene vocales de los tres tipos de clasificación, la primera sílaba tiene consonantes fricativas, la segunda es nasal, la tercera es líquida, y la última sílaba termina en oclusiva y vibrante.

- Rescate
  Descripción de sonido: Tiene vocales de clasificación abierta y media, su primera sílaba es vibrante y fricativa, la segunda y la tercera es oclusiva.

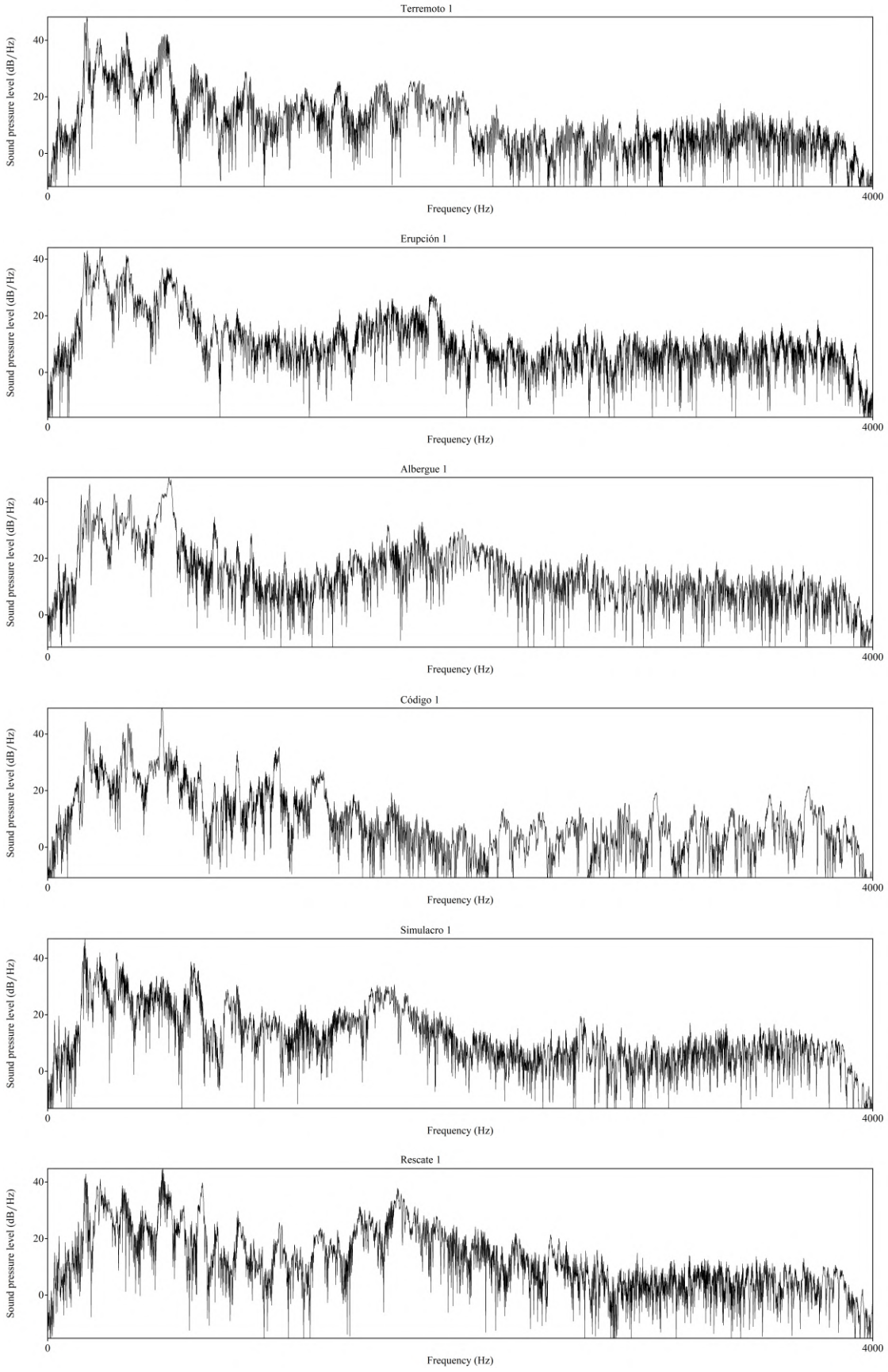
- Evacuación
  Descripción de sonido: Tiene vocales de los tres tipos de clasificación, la primer y segunda sílaba cuentan con consonantes oclusivas, la tercera tiene consonantes fricativas y nasales.

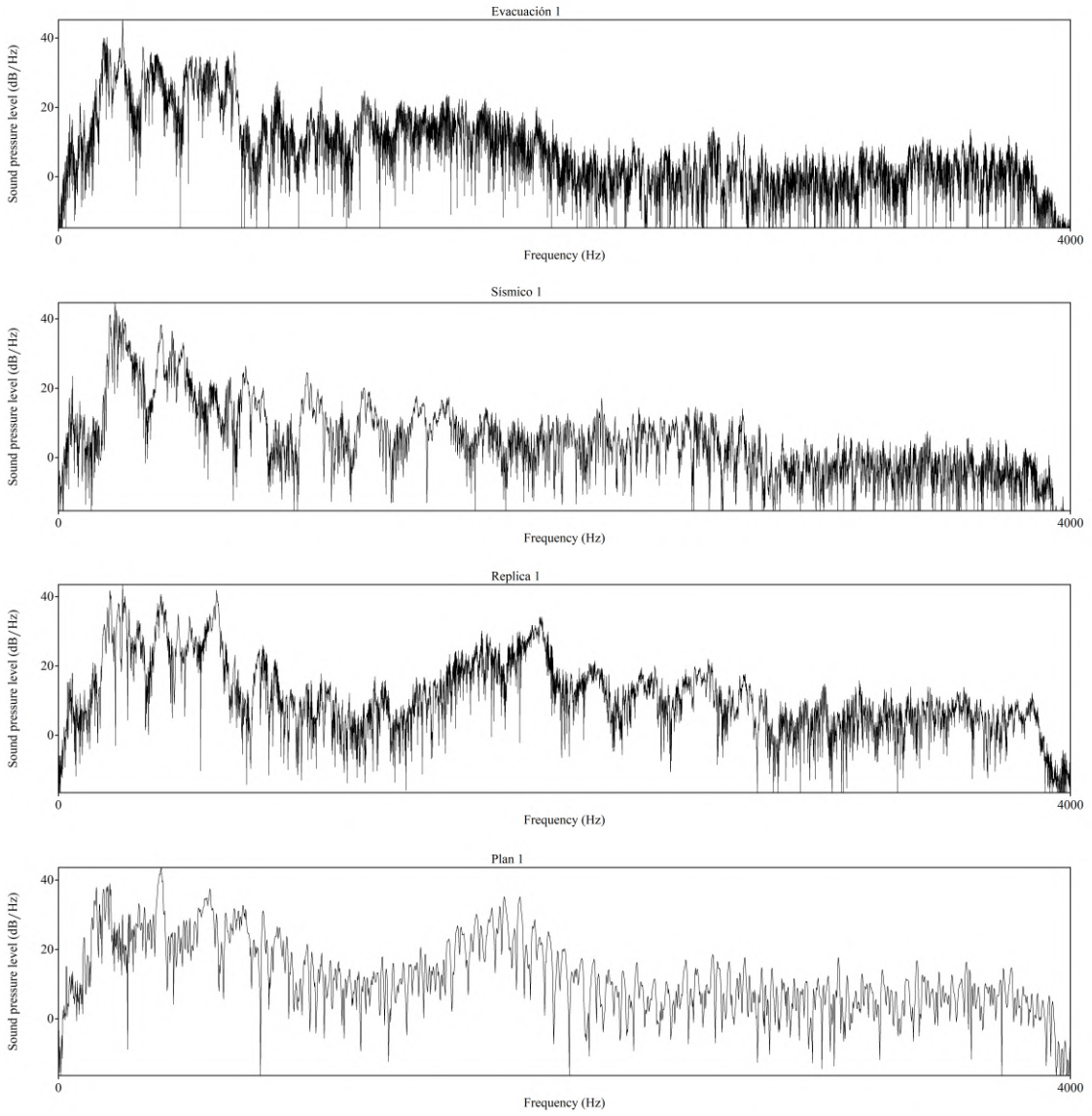
- Sísmico
  Descripción de sonido: Sus vocales son tipo media y cerrada. La primera y tercera sílaba tienen consonantes fricativas y la segunda es nasal.

- Réplica
  Descripción de sonido: Cuenta con los tres tipos de vocales. Su primer sílaba tiene consonantes vibrantes, la segunda oclusivas y líquida y la tercera es oclusiva.

- Plan
  Descripción de sonido: Su única vocal es de tipo abierta, y sus consonantes son de tipo oclusivas, líquidas y nasales.

Espectro

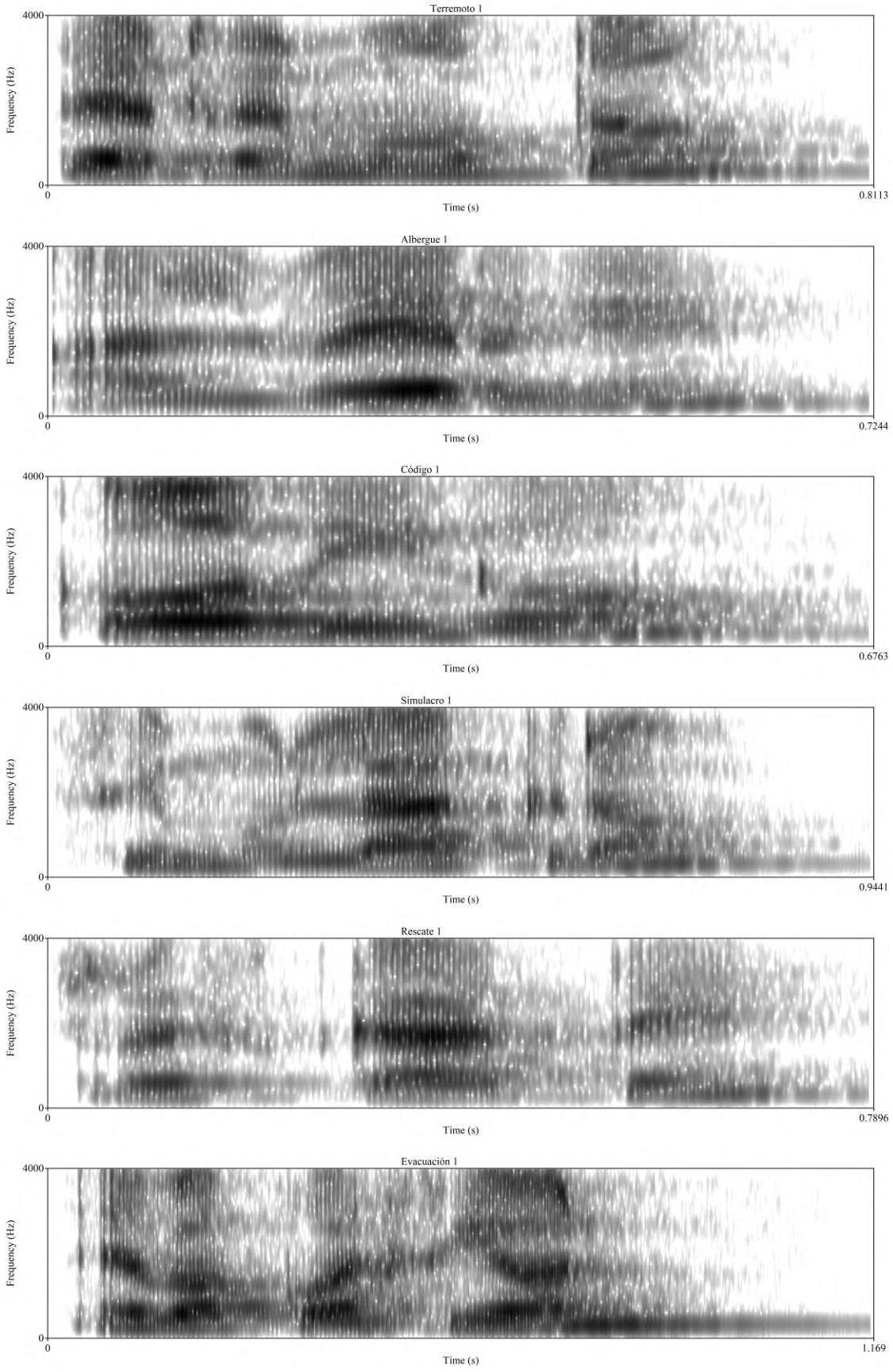

Terremoto 1
Sound pressure level (dB/Hz)
40
20
0
0
4000
Frequency (Hz)
Erupción 1
Sound pressure level (dB/Hz)
40
20
0
0
4000
Frequency (Hz)
Albergue 1
Sound pressure level (dB/Hz)
40
20
0
0
4000
Frequency (Hz)
Código 1
Sound pressure level (dB/Hz)
40
20
0
0
4000
Frequency (Hz)
Simulacro 1
Sound pressure level (dB/Hz)
40
20
0
0
4000
Frequency (Hz)
Rescate 1
Sound pressure level (dB/Hz)
40
20
0
0
4000
Frequency (Hz)

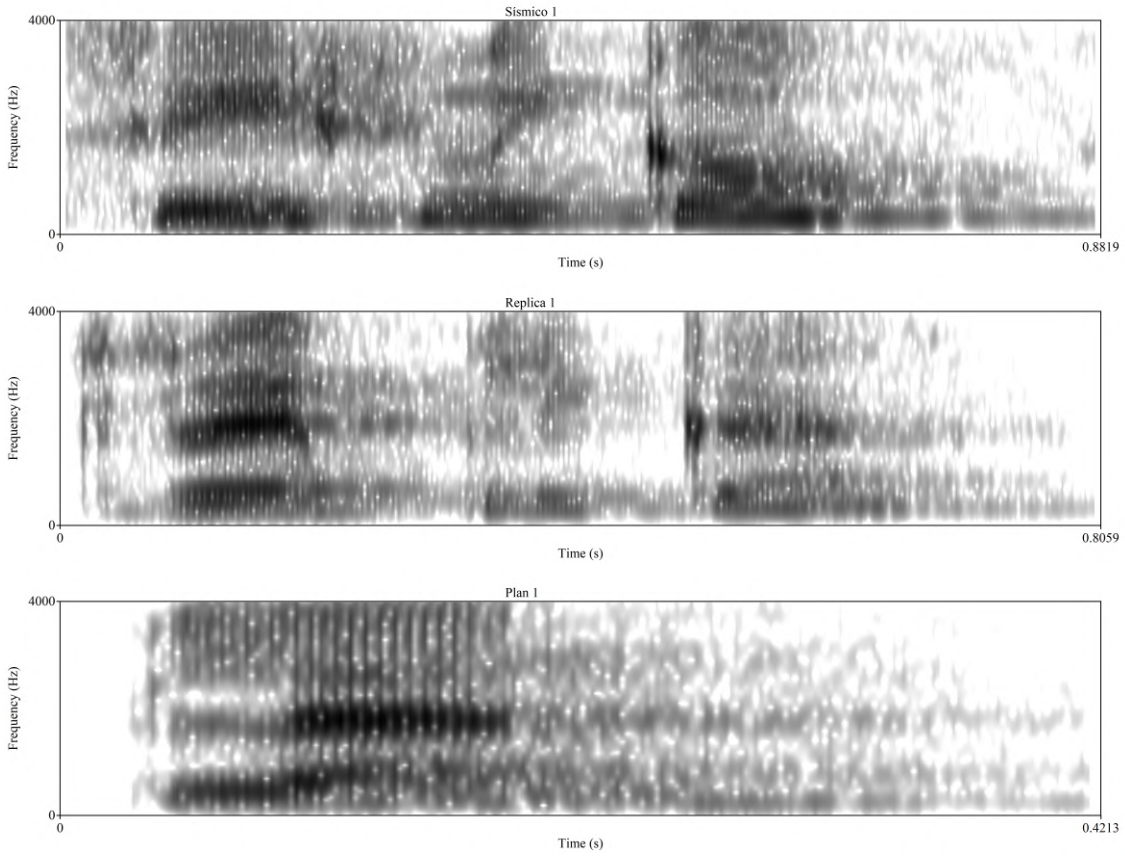


Observaciones

El espectro de cada audio nos permite observar las magnitudes de las componentes frecuenciales de la señal envolvente del fonema analizado en cada caso. Se puede notar que las palabras Albergue y Plan tienen picos similares en la región de baja frecuencia, asumiendo que lo que tienen en común son las letras 'a' y 'l', un aspecto importante a resaltar es que este tipo de comparaciones no son tan exactas porque no se tienen letras de referencia para comparar y definir a qué componentes están asociadas las letras, siendo que la señal no ha sido filtrada y en las gráficas también pueden haber componentes de ruido que se han filtrado durante la grabación de la voz.

Espectrograma



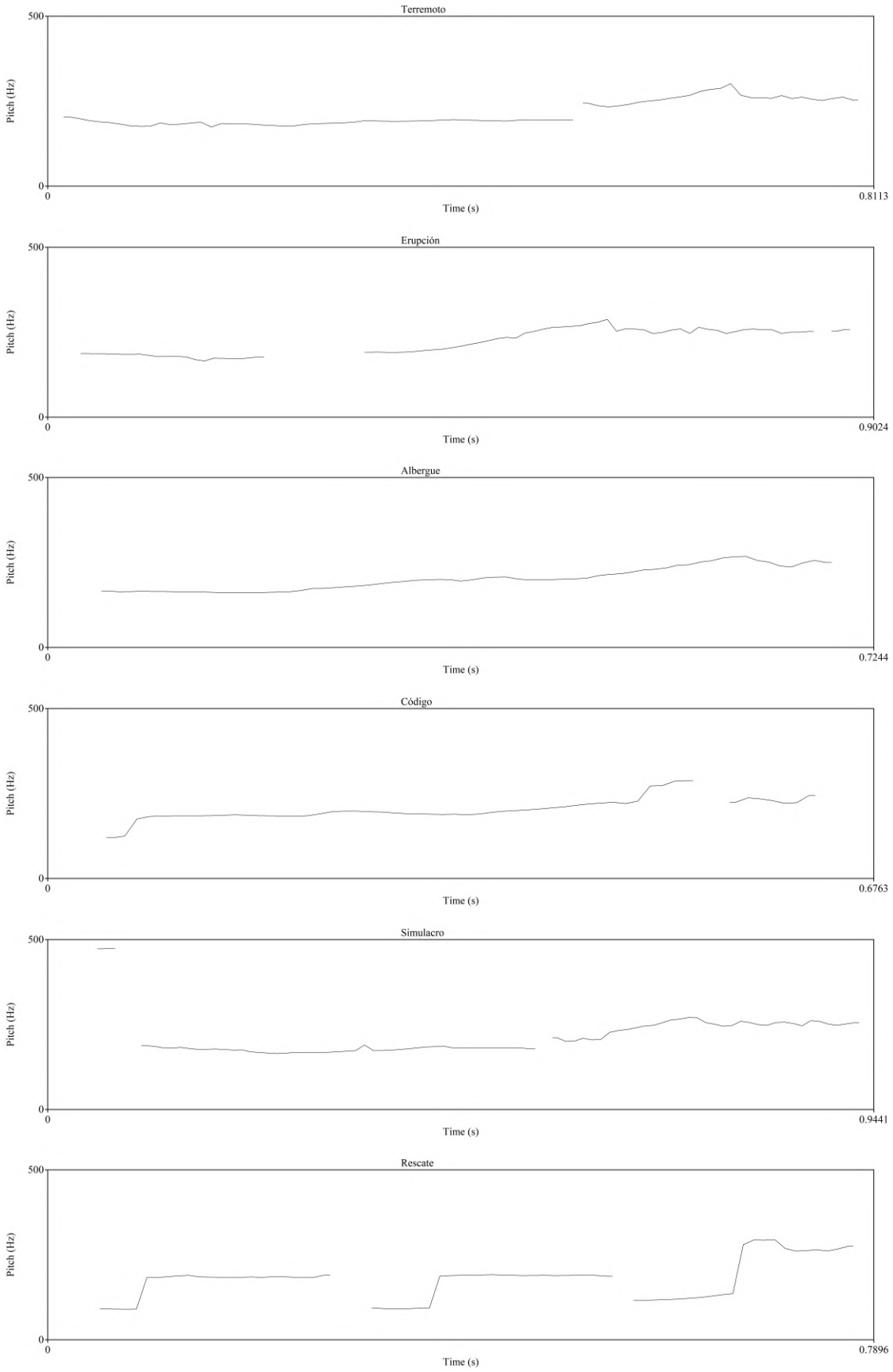


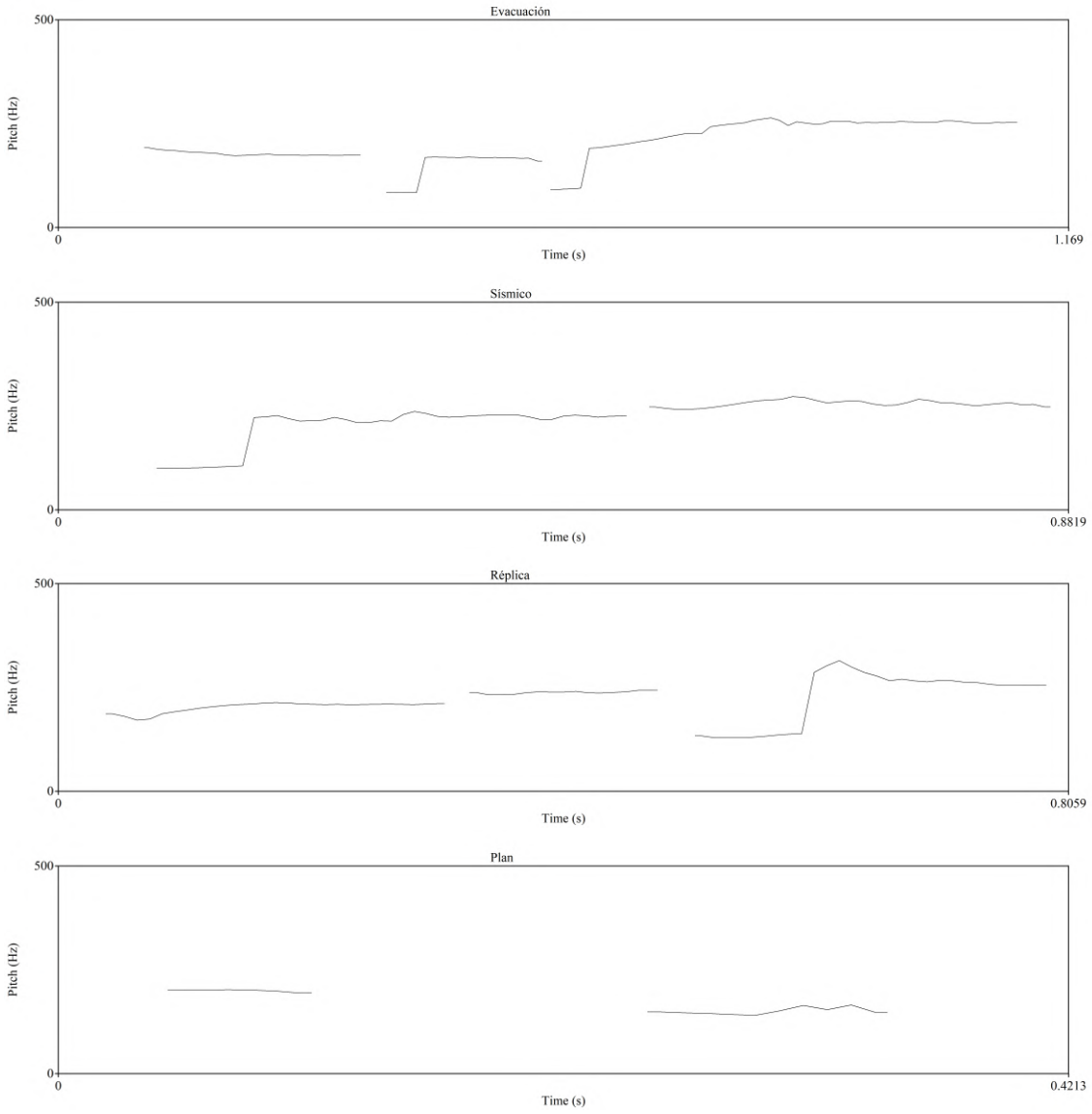
Observaciones

Para la palabra **Terremoto** la sílaba que tiene mayor magnitud en sus componentes frecuenciales, es la 'te' compuesta de consonante oclusiva y vocal media, la que tiene menores componentes frecuenciales es la compuesta por consonantes nasales y vocal media. Por otra parte, la palabra **Albergue** cuenta con una sílaba más predominante 'ber' compuesta por consonantes oclusivas y vibrantes, en cambio la que tiene menores componentes frecuenciales es la que está formulada por consonantes oclusivas y vocales cerradas.

En la palabra **Simulacro**, la sílaba con mayores componentes es 'la', compuesta por consonantes líquidas y vocales abiertas; similarmente con la palabra **Rescate**, la sílaba 'ca' es la que tiene mayores componentes frecuenciales debido a su consonante oclusiva pero también debido a su vocal abierta. Para la palabra **Evacuación**, la sílaba con mayores componentes frecuenciales es la última 'ción', cuyas consonantes son fricativas y debido al acento. Para la palabra **Plan**, se puede notar la variación de la frecuencia para la consonante oclusiva y la consonante líquida, y se puede notar que la componente nasal tiene menores componentes frecuenciales.

Pitch

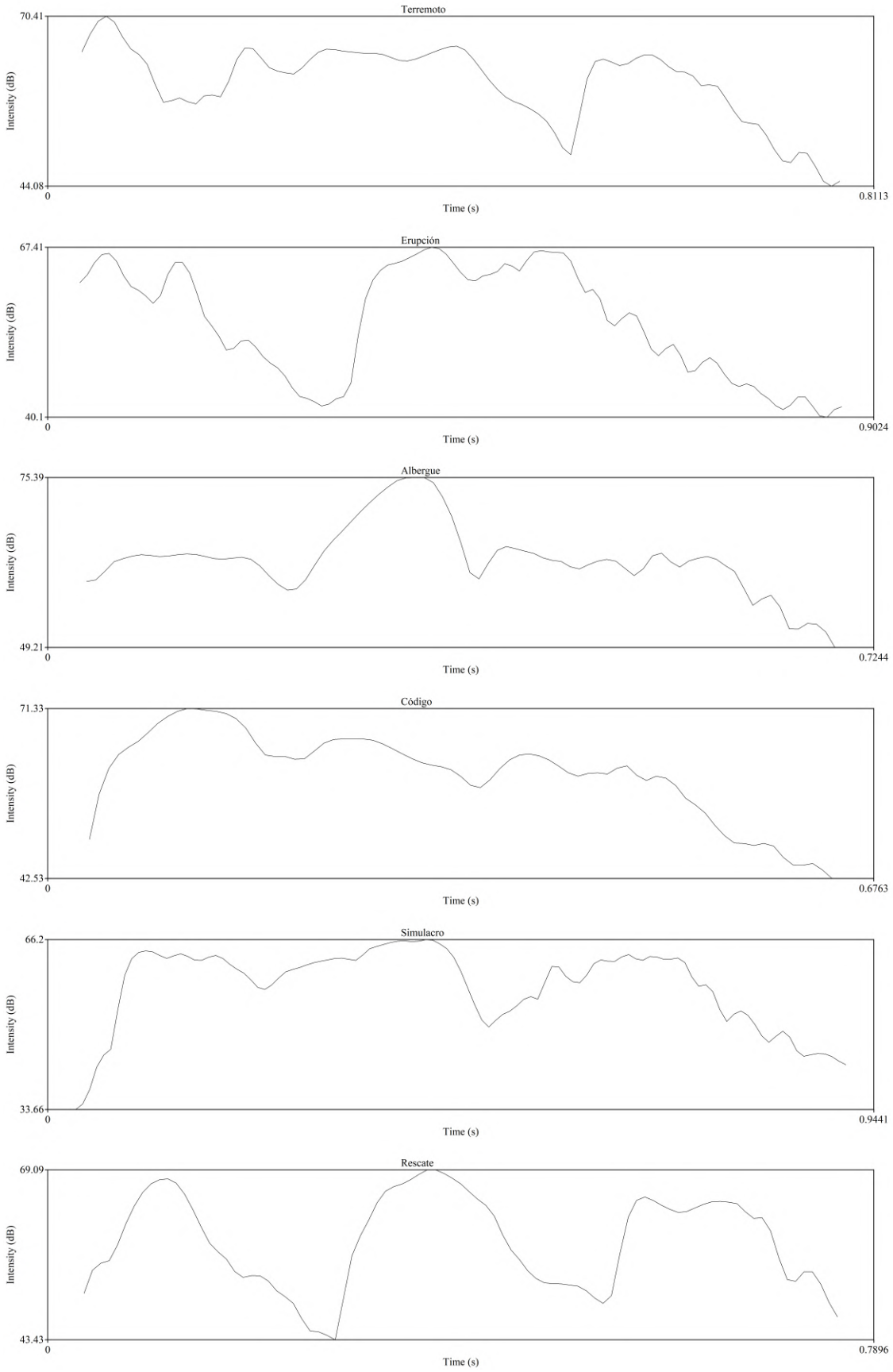

Terremoto
500
Pitch (Hz)
0
0
0.8113
Time (s)
Erupción
500
Pitch (Hz)
0
0
0.9024
Time (s)
Albergue
500
Pitch (Hz)
0
0
0.7244
Time (s)
Código
500
Pitch (Hz)
0
0
0.6763
Time (s)
Simulacro
500
Pitch (Hz)
0
0
0.9441
Time (s)
Rescate
500
Pitch (Hz)
0
0
0.7896
Time (s)

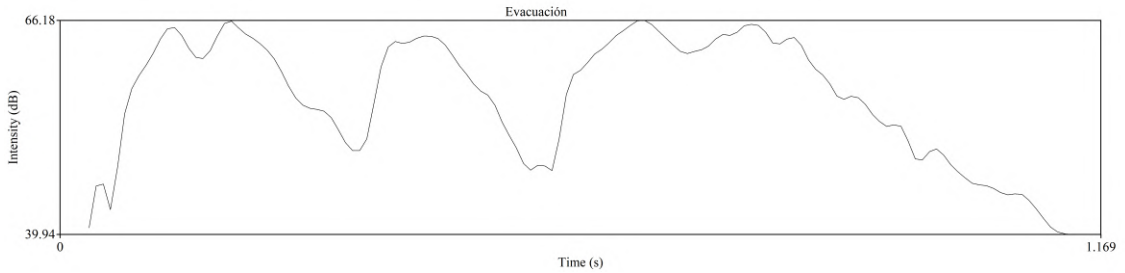


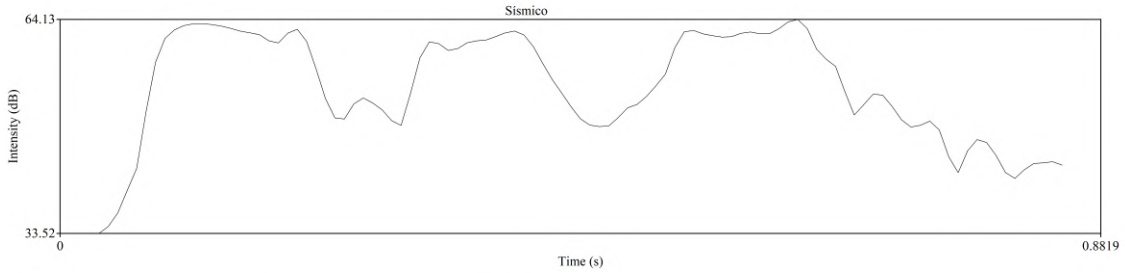
Observaciones

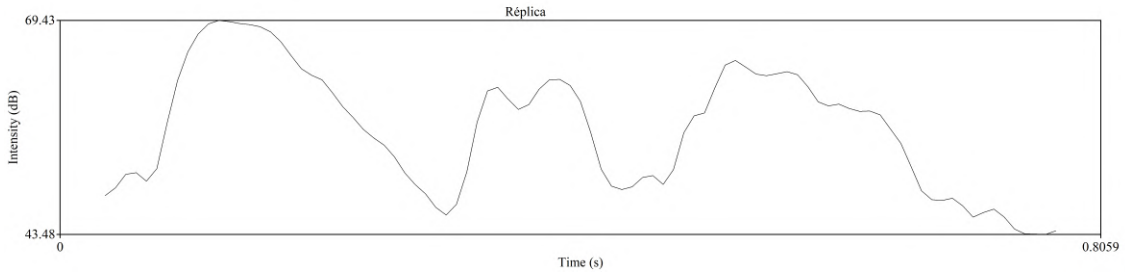
Un aspecto que se logra identificar en todas las gráficas es que la forma de onda en todas las palabras no es continua, esto debido a que el Pitch se asocia al movimiento que realiza la glotis en la generación del sonido por esto, la oscilación glotal es una función cuasi-periódica, y esto se ve más marcado en las palabras **Rescate**, **Erupción**, **Réplica**, **Plan**, un aspecto que tienen en común todas estas palabras es que están compuestas por consonantes oclusivas y vibrantes.

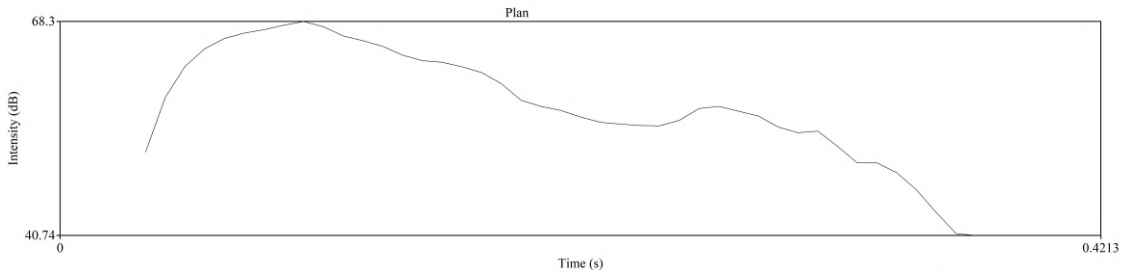
Intensidad


Terremoto
70.41
Intensity (dB)
44.08
0
0.8113
Time (s)
Erupción
67.41
Intensity (dB)
40.1
0
0.9024
Time (s)
Albergue
75.39
Intensity (dB)
49.21
0
0.7244
Time (s)
Código
71.33
Intensity (dB)
42.53
0
0.6763
Time (s)
Simulacro
66.2
Intensity (dB)
33.66
0
0.9441
Time (s)
Rescate
69.09
Intensity (dB)
43.43
0
0.7896
Time (s)

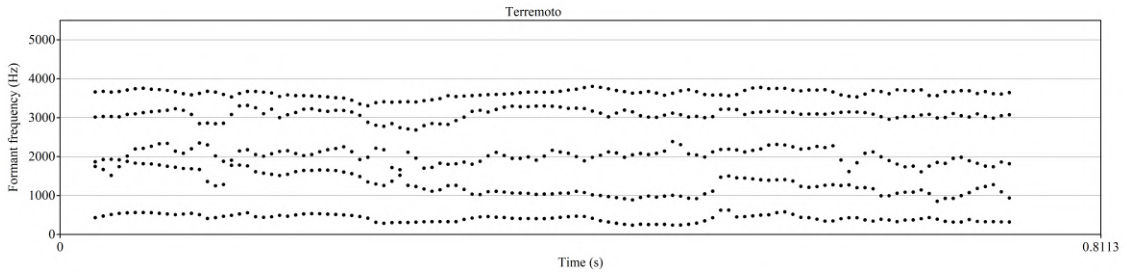

Evacuación
66.18
39.94
Intensity (dB)
0
1.169
Time (s)

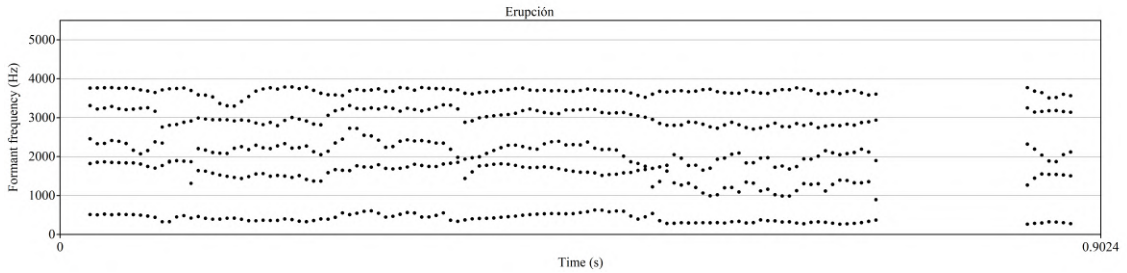

Sísmico
64.13
33.52
Intensity (dB)
0
0.8819
Time (s)

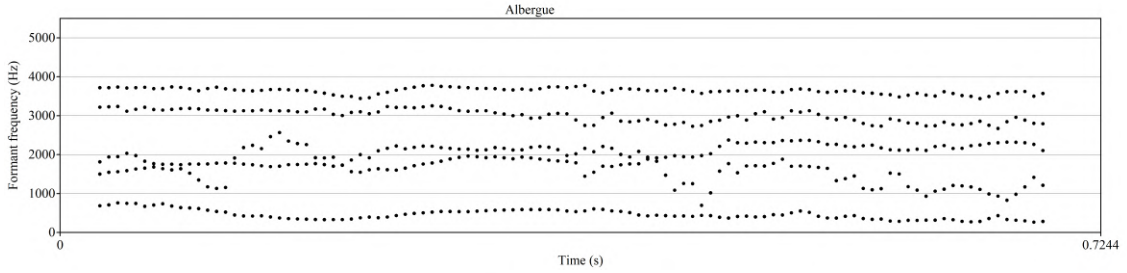

Réplica
69.43
43.48
Intensity (dB)
0
0.8059
Time (s)

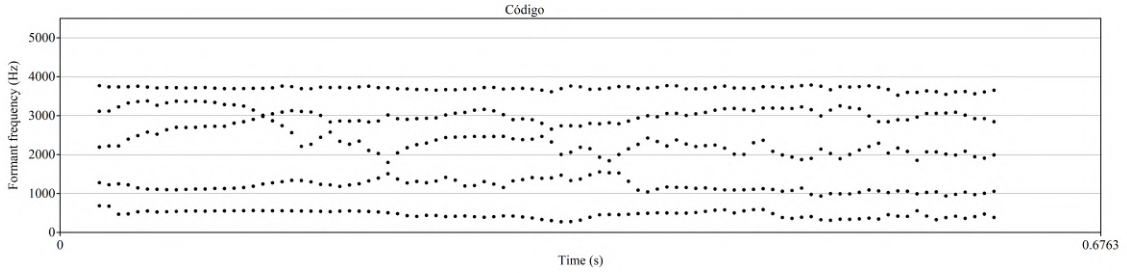

Plan
68.3
40.74
Intensity (dB)
0
0.4213
Time (s)

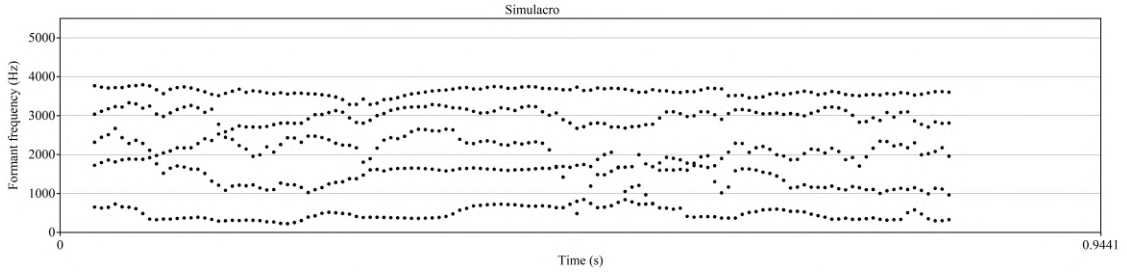
Formantes

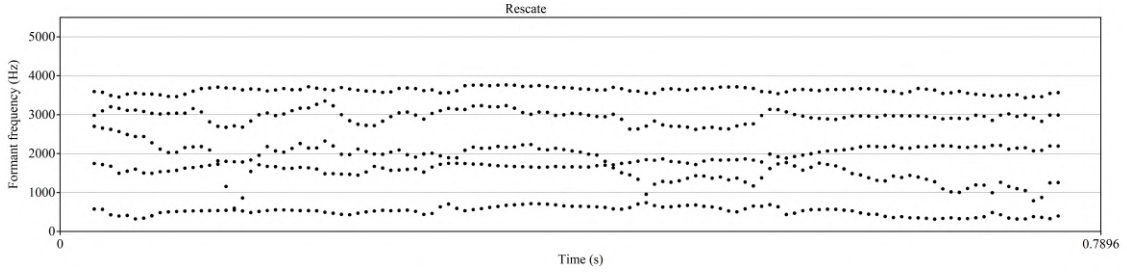

Terremoto
Formant frequency (Hz)
5000
4000
3000
2000
1000
0
0
0.8113
Time (s)

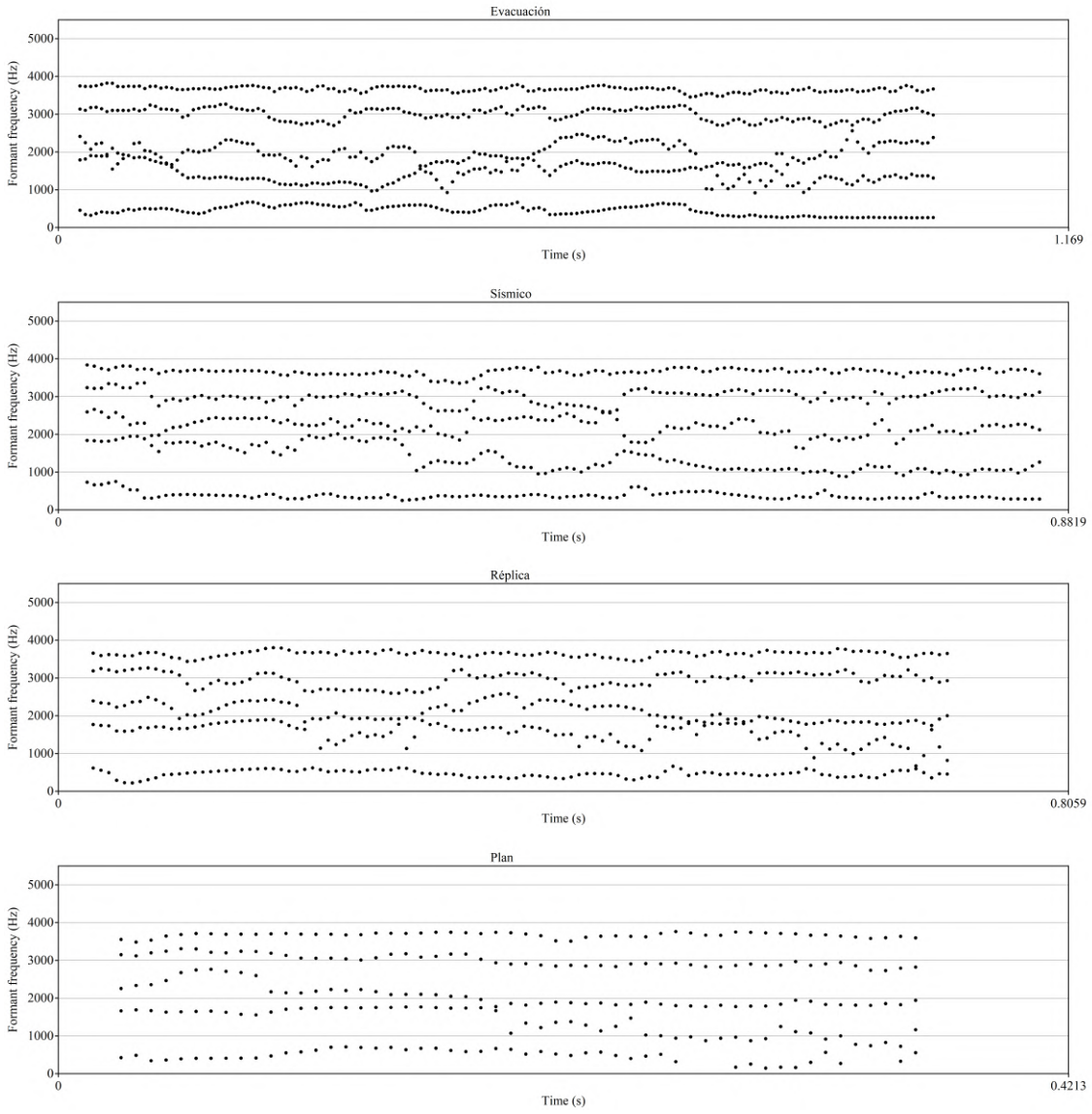

Erupción
Formant frequency (Hz)
5000
4000
3000
2000
1000
0
0
0.9024
Time (s)


Albergue
Formant frequency (Hz)
5000
4000
3000
2000
1000
0
0
0.7244
Time (s)


Código
Formant frequency (Hz)
5000
4000
3000
2000
1000
0
0
0.6763
Time (s)


Simulacro
Formant frequency (Hz)
5000
4000
3000
2000
1000
0
0
0.9441
Time (s)


Rescate
Formant frequency (Hz)
5000
4000
3000
2000
1000
0
0
0.7896
Time (s)

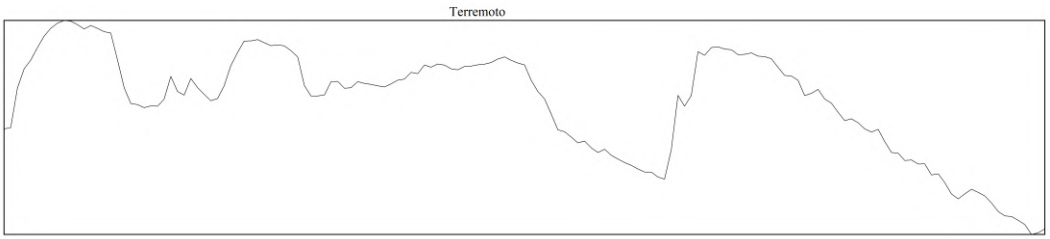


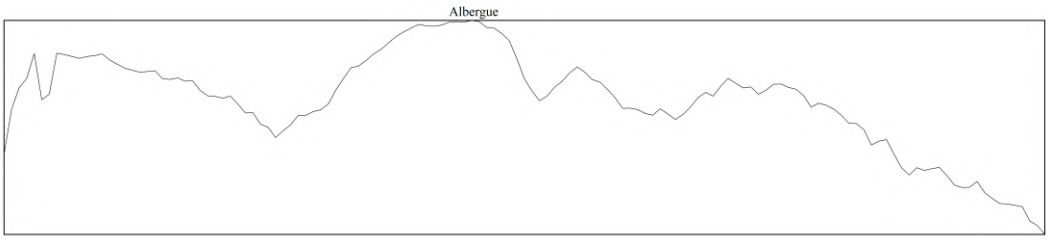
El proceso de articulación determina las frecuencias de los formantes vocales. La anatomía vocal que él asocia con las frecuencias de los formantes. La apertura de la mandíbula, que constriñe el tracto vocal hacia el final de la glotis y lo expande en el extremo del labio, es el factor decisivo para el primer formante. Esta frecuencia de formantes aumenta a medida que la mandíbula se abre más. El segundo formante es más sensible a la forma del cuerpo de la lengua, y el tercero formante es más sensible a la punta de la lengua. En este sentido, Benade sugiere los siguientes rangos de frecuencias de los formantes de una voz

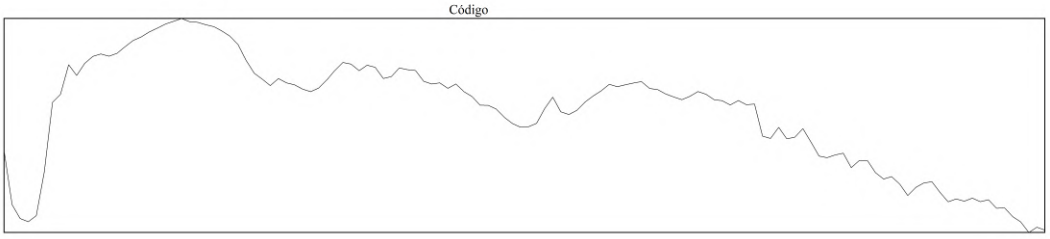
- 1° Formante: 150-850 Hz
- 2° Formante: 500-2500 Hz
- 3° Formante: 1500-3500 Hz
- 4° Formante: 2500-4800 Hz

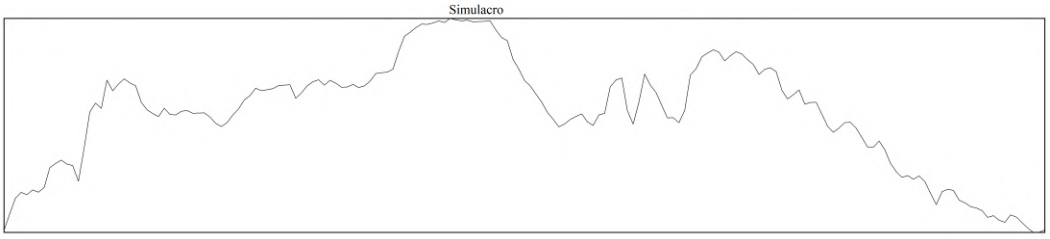
De esta manera se puede notar que para las palabras que contienen una misma vocal como es el caso de **Plan**, es posible identificar los formantes debido a 'a', por ejemplo para la palabra **Código** existe una variación en los formantes 3 y 4 que van de 3kHz a 2 kHz.

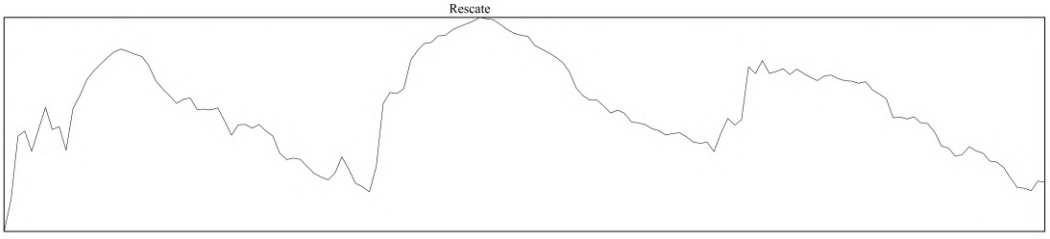
MFCC

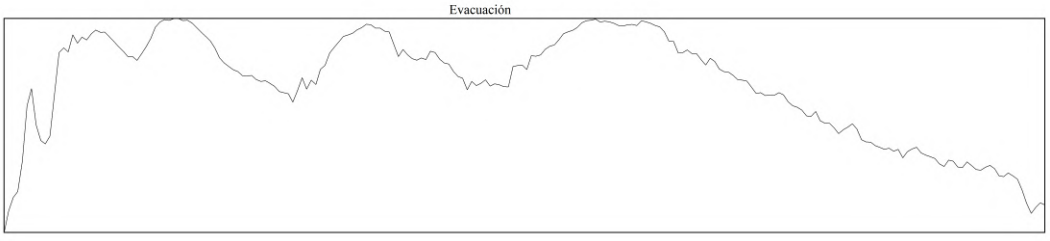


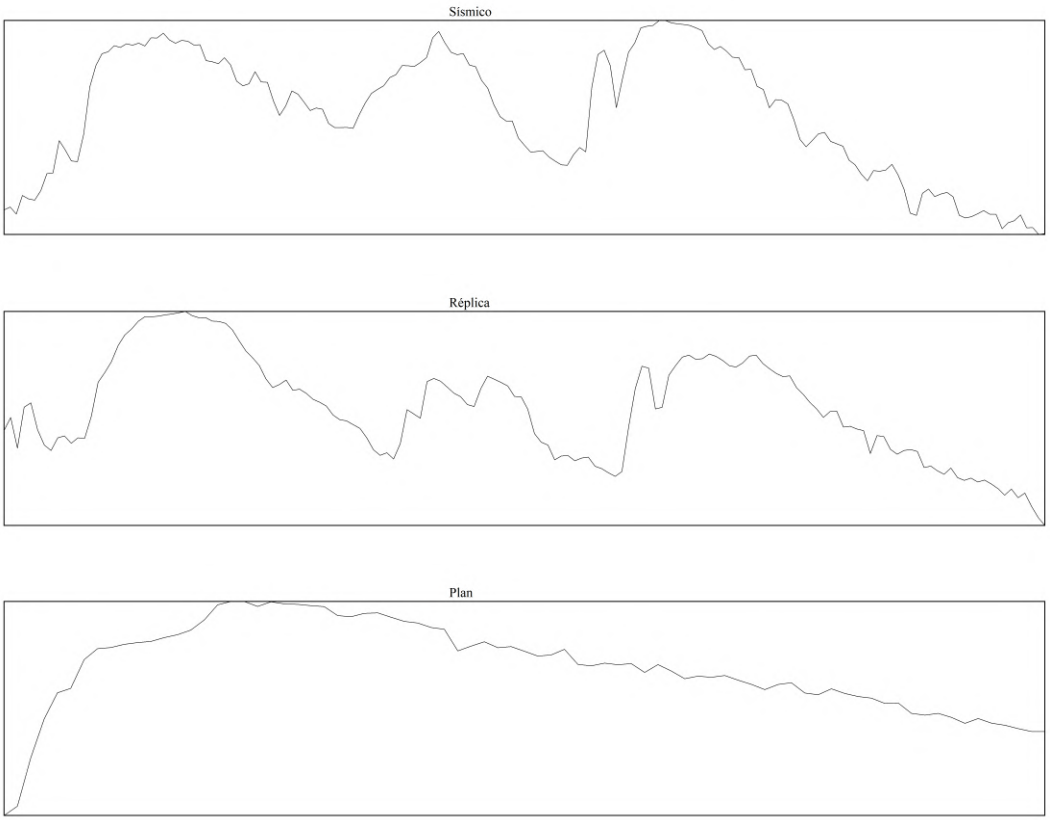






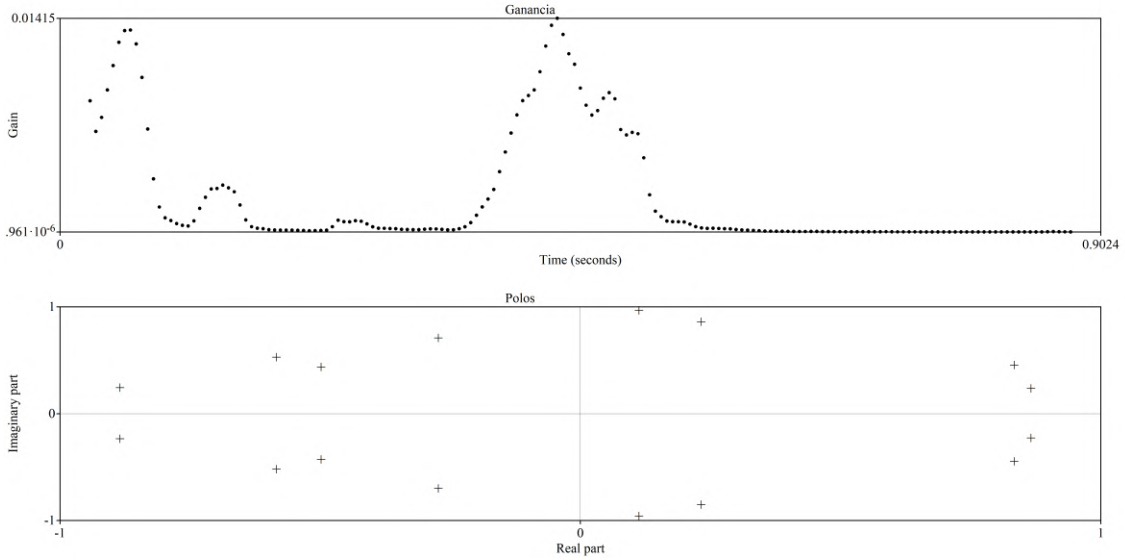




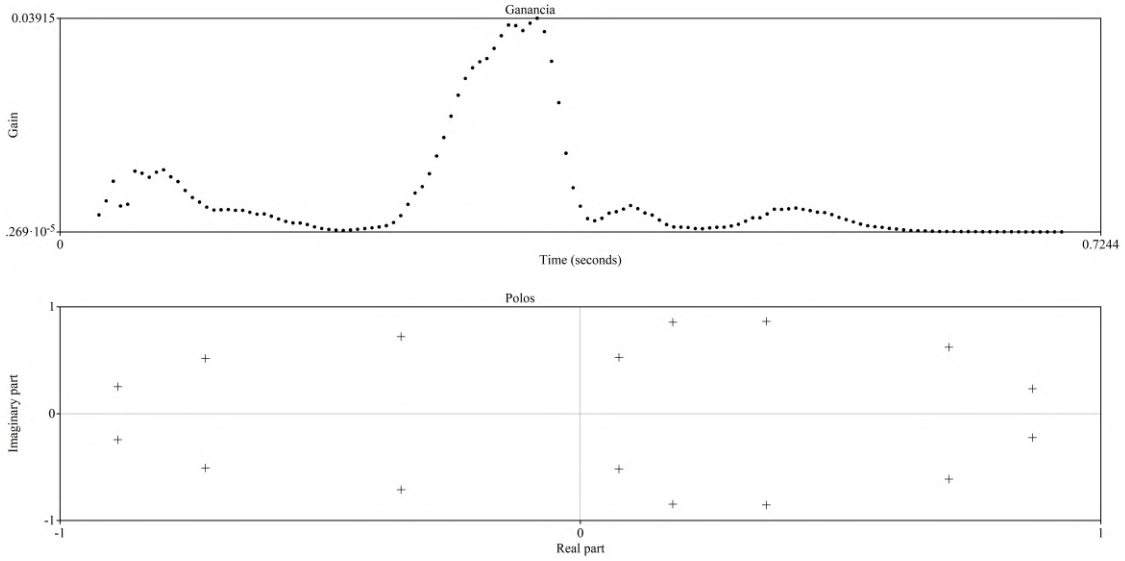


Los Coeficientes Cepstrales en la Escala de Mel (MFCC) nos permiten representar la amplitud del espectro del habla de manera compacta, aplicando un filtro de pre-énfasis a la señal y posteriormente dividiendo la misma en tramas y se le aplica una función de ventaneo, en este caso una ventana de Hamming. De esta manera se puede notar que los resultados están muy relacionados con las gráficas obtenidas en la sección Pitch.

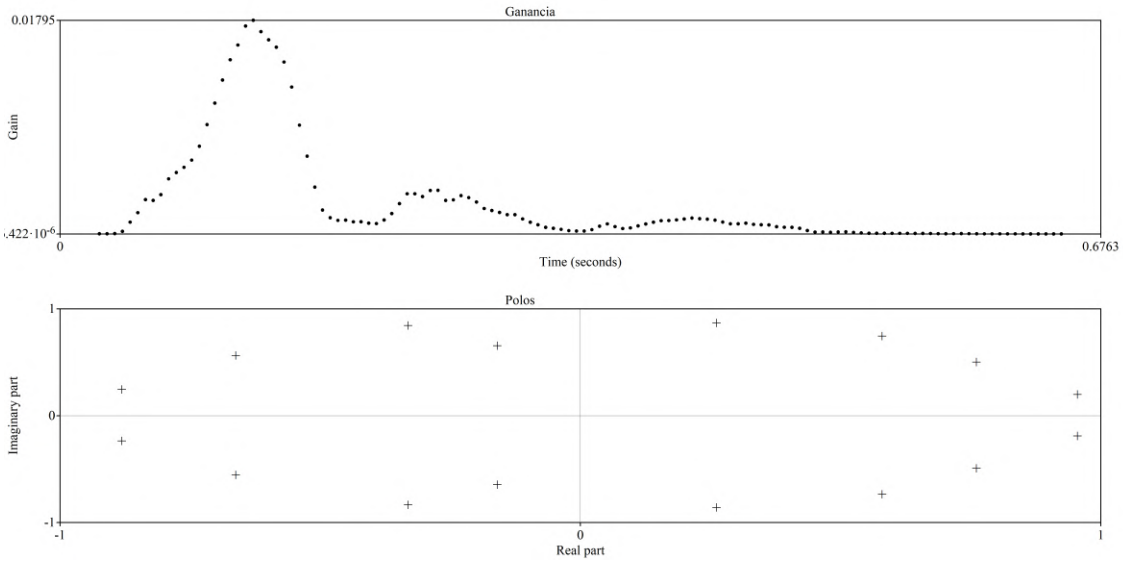
LPC



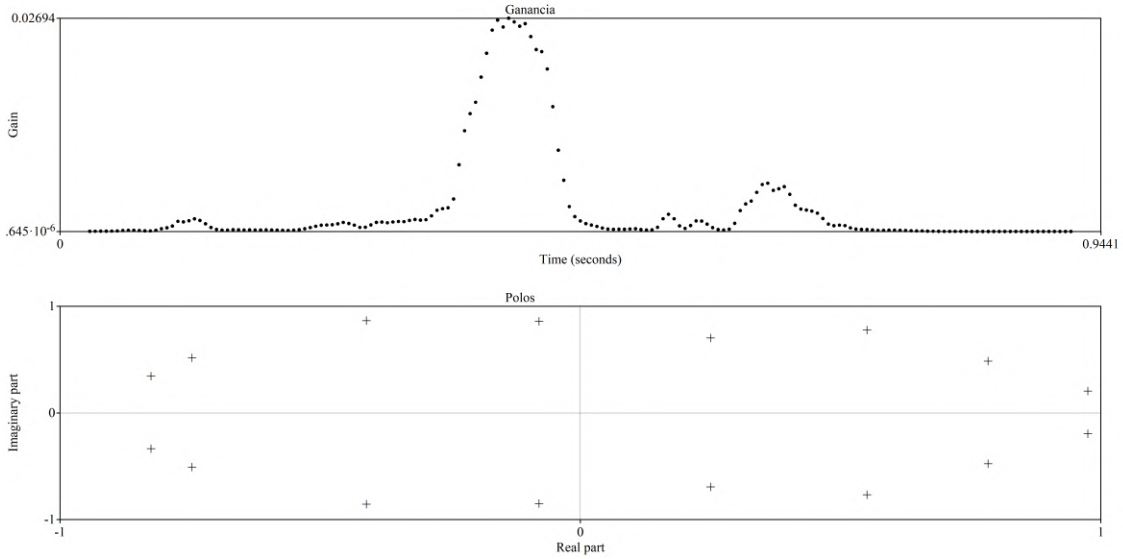
Albergue



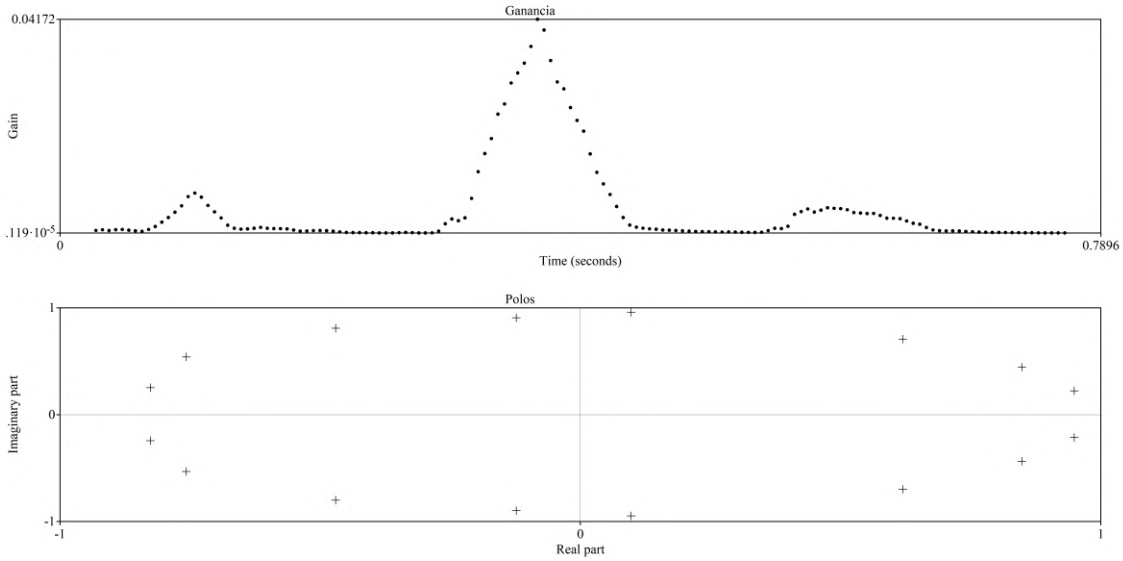
Código



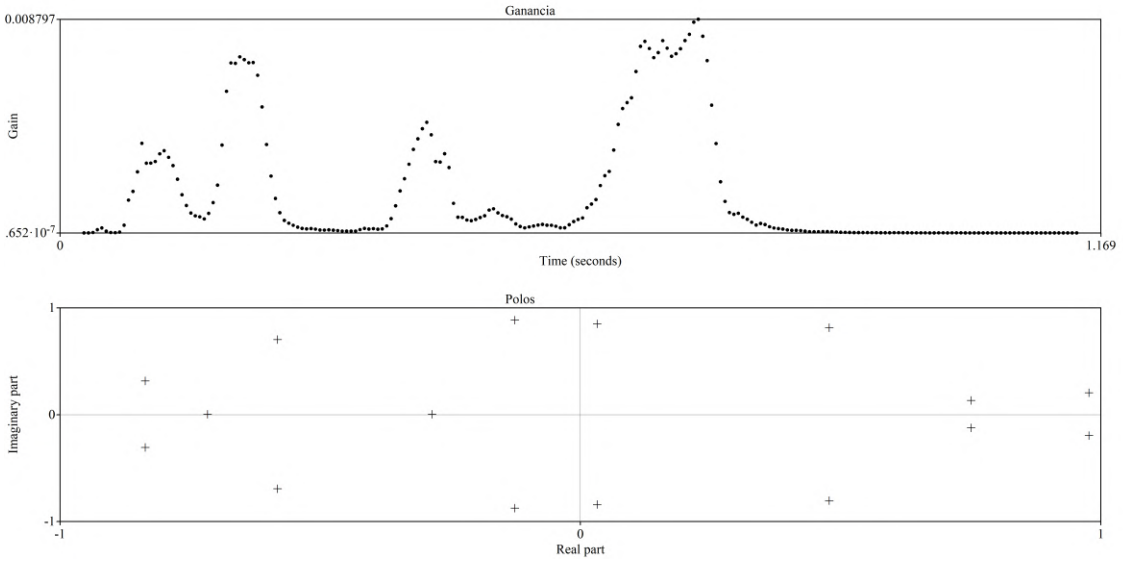
Simulacro



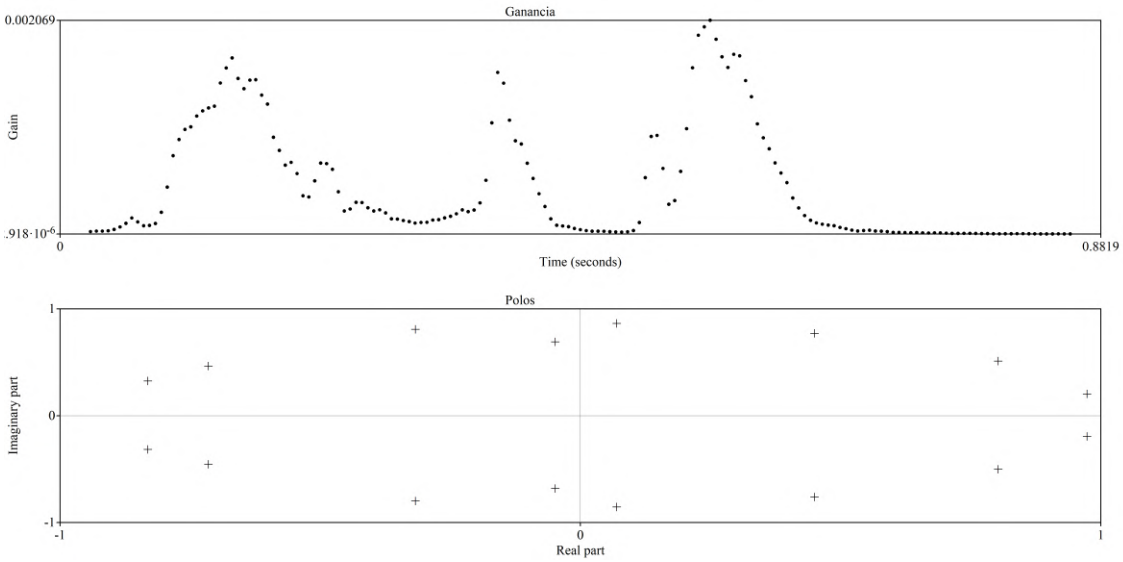
Rescate



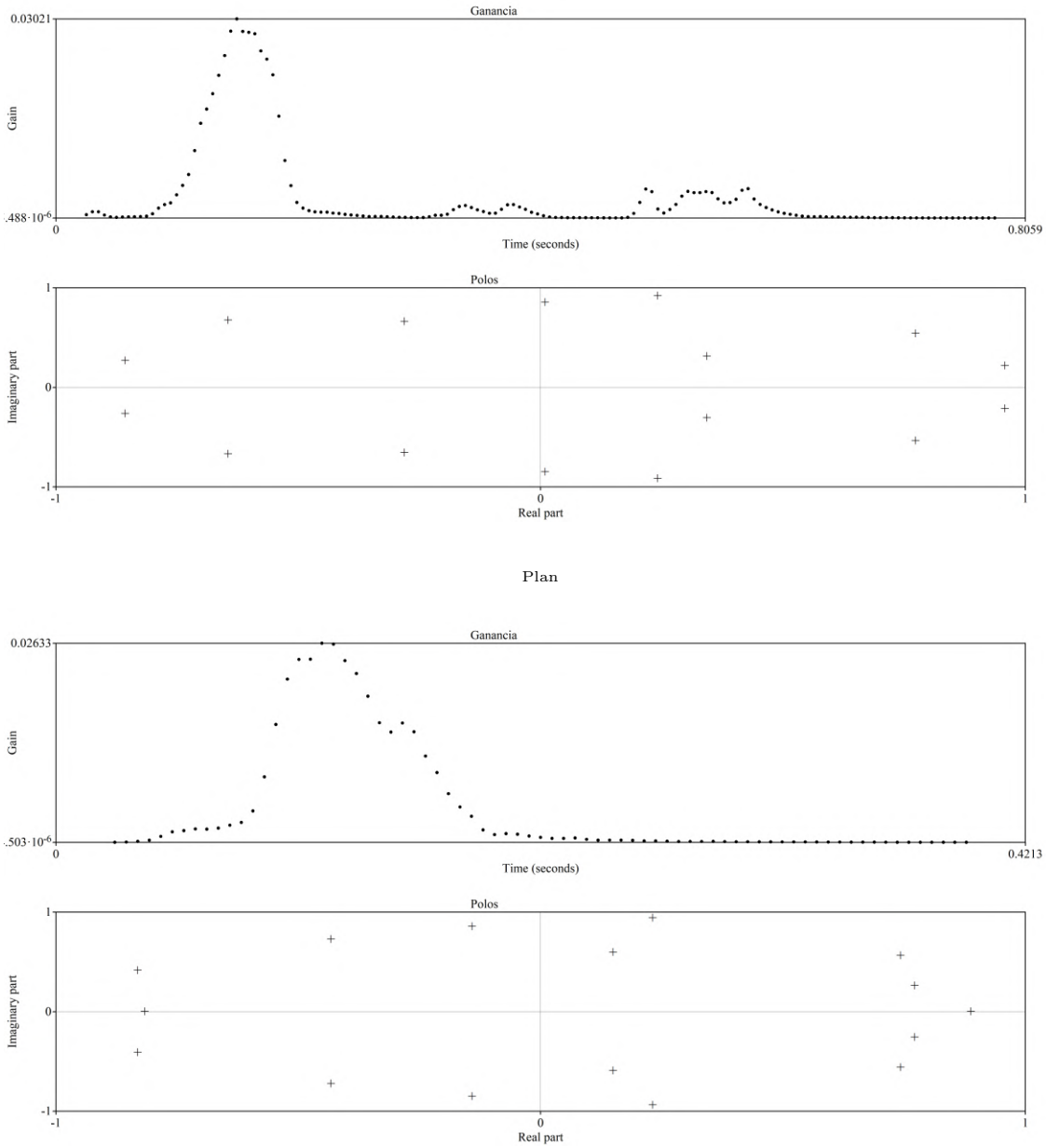
Evacuación



Sísmico



Réplica



Plan

Para este análisis, los coeficientes del filtro mostrados como función del tiempo, nos permiten observar la ganancia para las sílabas más relevantes, para la palabra **Erupción** la última sílaba compuesta por las vocales medias y cerradas con acento, muestra el mayor pico de ganancia, el número total de polos de la señal es de 16.
La palabra **Albergue**, tiene una ganancia mayor en la sílaba 'ber', y cuenta con un número total de 16 polos, por otra parte la palabra **Código** tiene un pico de ganancia precisamente en la sílaba asociada a la vocal con acento 'Có' y cuenta con 16 polos. Al analizar la palabra **Simulacro**, tiene un pico de ganancia en la primer vocal 'Si', debido a su consonante fricativa y al igual que el caso anterior cuenta con 16 polos, en la palabra **Evacuación**, su pico más grande no es debido a la vocal con acento como el caso de código, es en la unión de las dos vocales, este pico es más predominante que el asociado a la sílaba 'ción'.
En la respuesta del LPC para la palabra **Sísmico**, la mayor ganancia se obtuvo en la sílaba 'Sís', y al igual que en los casos anteriores se tienen 16 polos totales, finalmente la palabra **Plan** tiene solamente un pico debido a su única vocal.

Conclusiones

Cuando pronunciamos una vocal, las cuerdas vocales vibran, produciendo una vibración en el aire en forma de tren de pulsos, que a su vez, este tren de pulsos pasa a través del tracto vocal, el cual, dependiendo del fonema que se ha pronunciado actuará como un filtro y modificará las componentes frecuenciales del tren de pulsos.
Los métodos aquí analizados, nos permiten identificar las características y variaciones para cada palabra pronunciada, lo anterior es aplica en algoritmos reconocimiento del habla porque estas características del tracto vocal, corresponden a la forma de la envolvente del sonido visto desde el punto de vista frecuencial (como es el caso del MFCC). Otros parámetros como LPC nos permiten conocer la respuesta a un filtro.
En un principio se analizaron las 5 repeticiones para cada palabra, sin embargo añadir todos los resultados obtenidos en un solo archivo implicaba mucha información, por lo tanto, para la realización de este reporte se colocó el resultado del análisis para la primer palabra analizada con la diferencia de que con anterioridad se comparó entre ellas e identificó señales particulares que fueron mencionadas a lo largo del reporte. Todas las imagenes obtenidas pueden ser consultadas en el repositorio de GitHub, así como el audio completo utilizado y la segmentación por palabras del mismo.