uc3m | Universidad Carlos III de Madrid

Grado Universitario en Ingeniería Electrónica Industrial y Automática
2016-2017

Trabajo Fin de Grado

"Sistema de análisis de audio"

Sergio Pérez García-Tenorio

Tutor/es

Ángel García Crespo

Leganés – Octubre 2017

Índice de contenidos

Índice de figuras

Índice de tablas

Índice de acrónimos

IPA *International Phonetic Alphabet*

RMS *Root Mean Square (amplitude)*

SI Sistema Internacional de Unidades

FAQ *Frequently Asked Questions*

WAV *Waveform Audio File Format*

AIFF *Audio Interchange File Format*

FLAC *Free Lossless Audio Codec*

MP3 *MPEG-1 Audio Layer III* o *MPEG-2 Audio Layer III*

MPEG *Moving Picture Experts Group*

HNR *Harmonic-to-noise Ratio*

SPL *Sound Pressure Level*

LTAS *Long Term Average Spectrum*

LPC *Linear Predictive Coding*

Resumen

Muchos lingüistas utilizan grabaciones de voz en su trabajo de investigación diario. En el trabajo descriptivo, las representaciones gráficas de las grabaciones eran generalmente anotadas manualmente. Con la irrupción de los ordenadores, se ha permitido sustituir un caro equipo de laboratorio por una sola máquina, y se han reducido el coste y el tiempo que lleva ejecutar estos análisis.

El programa Praat es una herramienta creada por el departamento de lingüística de la Universidad de Ámsterdam que permite multitud de operaciones relacionadas con análisis de voz, así como la creación de scripts que automatizan series de estas tareas.

Este trabajo se basa en la elaboración de un script de Praat que permita la detección de las vocales en una grabación de audio dada, así como el etiquetado de las mismas y un análisis preliminar, y en último lugar el catalogar la vocal entre los cinco sonidos vocálicos del castellano.

Abstract

Multiple linguists use voice recordings in their daily research job. In descriptive work, graphical representations of recordings were generally annotated manually. With the introduction of computers, expensive laboratory equipment has been substituted for only one machine, and the costs and time to do this analysis has been reduced.

Praat software is a tool created by the Linguistics department of the University of Amsterdam that allows multiple operations related to speech analysis, as well as the development of scripts that automate sets of this tasks.

This job is base in the development of a Praat script that will allow to detect vowels in a given audio recording, as well as the annotation of these vowels, and in the last term to catalog the vowel among the five existing vowel sounds in Spanish language.

1. Motivación y objetivos

Muchos lingüistas utilizan grabaciones de voz en su trabajo de investigación diario. En el trabajo descriptivo, las representaciones gráficas de las grabaciones son generalmente anotadas con símbolos IPA y otras etiquetas, y posteriormente usadas para ilustrar un fenómeno o defender una cierta postura acerca de la naturaleza de alguna propiedad fonética o fonológica del lenguaje en cuestión.

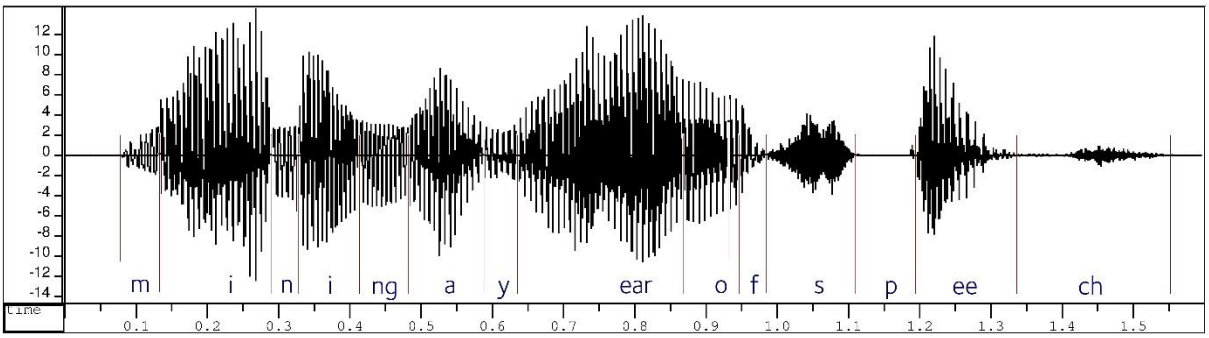


Figura 1: Representación de la forma de una onda de voz

La introducción de los ordenadores ha supuesto una revolución virtual de las ciencias lingüísticas con respecto al uso de grabaciones de voz. Un laboratorio lleno de maquinaria ha sido sustituido por un solo PC, donde cualquiera puede grabar, anotar y modificar audio con unos cuantos comandos o *clicks*. Incluso el cálculo de algunos parámetros del habla, que antes se antojaban muy complejos, pero sin embargo eran necesarios para el análisis, están ahora a sólo unos *clicks* de distancia.

El análisis de vocales es una parte muy importante en los estudios lingüísticos, ya que permite una mejor comprensión de la manera en la que los seres humanos producen los sonidos que dan lugar al lenguaje. Un detenido estudio de las mismas permite establecer diferencias en el habla por origen, edad o sexo, entre otros; así como entender y ayudar a mejorar el habla en las personas que presentan disfuncionalidades en la misma.

En este marco, el objetivo de este Trabajo Fin de Grado es el de desarrollar un *script* que detecte automáticamente las vocales en una grabación de voz y calcule los parámetros característicos de las mismas para su posterior procesado y análisis. Este *script* debe ser accesible tanto para personas que tienen conocimientos de informática como para aquellas que no, de forma que cualquiera pueda usarlo de manera fácil e intuitiva; y asimismo ser fácil de modificar en caso de necesitar un análisis más específico y detallado.

2. Estado del arte

2.1. Fonética

La fonética es la rama de la lingüística que estudia los sonidos del habla humana, o los aspectos equivalentes en caso del lenguaje de signos[1]. Se ocupa por tanto de las propiedades físicas del habla, los fonemas: la forma fisiológica en que se producen, sus propiedades acústicas, percepción auditiva y estado neurofisiológico. Por otro lado, la fonología se ocupa de la parte abstracta del lenguaje.

Esta rama se estudia probablemente desde el siglo IV a.C., incluso es posible que anteriormente, en torno al VI a.C. La fonética moderna se inicia con intentos de notación precisa de los sonidos del habla entre finales del siglo XVIII y el siglo XIV[2]. Su estudio experimentó un rápido crecimiento en el siglo XIV con la aparición del fonógrafo, lo que permitió la grabación de voz, facilitando que los fonetistas pudieran reproducir la señal varias veces y aplicarle filtros. Gracias a ello, pudieron deducir mejor la naturaleza acústica de las señales de voz.

Usando un fonógrafo de Edison, Ludimar Hermann investigó las propiedades espectrales de las vocales y consonantes, dando lugar a la aparición por primera vez del término formante[3], del que se hablará en profundidad más adelante. Hermann también reprodujo grabaciones de vocales hechas con el fonógrafo de Edison a diferentes velocidades con el objetivo de probar las teorías de Robert Willis y Charles Wheatstone de producción de vocales.

La teoría de Willis establecía una correspondencia entre la producción de vocales y la producción de notas musicales usando un órgano: los pulmones actuaban como fuelles, las cuerdas vocales como válvula y la cavidad oral como los tubos del órgano[4]. Por lo tanto, diferentes vocales correspondían a diferentes cavidades orales (tubos de órgano) de diferentes longitudes, que eran independientes de la vibración de las cuerdas vocales (válvula). Ésta teoría se contrasta con la teoría armónica de Wheatstone de la producción de vocales.

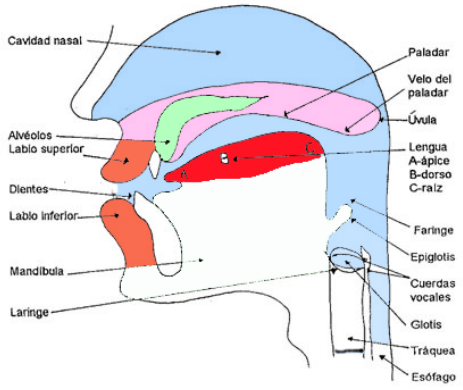


Figura 2: Partes del aparato fonador

Figura 3: Partes de un órgano

La fonética a su vez está dividida en varios campos, que en caso del lenguaje oral son tres principalmente:

- Fonética articulatoria: estudio de la producción de sonidos del habla desde el punto de vista fisiológico, es decir, de los órganos que intervienen en el habla[5].
- Fonética acústica: estudio de la transmisión física de los sonidos del emisor al receptor, en la que nos fijaremos más en profundidad.
- Fonética auditiva: estudio de la recepción y percepción de los sonidos del habla por el receptor.

En el siguiente capítulo se hablará más en profundidad de la fonética acústica, ya que se trata del área de más interés en el desarrollo de este proyecto.

2.1.1. Fonética acústica

La fonética acústica, como se dijo en el anterior capítulo, es el campo de la fonética que trata con los aspectos acústicos de la voz. Este campo investiga tanto las características en el dominio del tiempo como la amplitud de una forma de onda, la frecuencia, la duración o la frecuencia fundamental; como en el dominio de la frecuencia. Adicionalmente, estudia aspectos espectrotemporales y la relación de estas propiedades con las otras ramas de la fonética, o con conceptos lingüísticos abstractos como fonemas, frases o enunciados.

En adición a los avances en fonética citados en el apartado anterior, los avances en fonética acústica se hicieron posibles con el desarrollo de la industria telefónica. Durante la II Guerra Mundial, el trabajo de los *Bell Telephone Laboratories* facilitó el estudio sistemático de las propiedades espectrales de sonidos periódicos y aperiódicos, resonancias de tracto vocal y formantes de vocales, calidad de la voz, prosodia, etcétera.

A nivel teórico, la acústica de la voz puede ser modelada de forma análoga a los circuitos eléctricos. Lord Rayleigh estuvo entre los pioneros en reconocer que la nueva teoría eléctrica podría ser usada en acústica, pero no fue hasta 1941 cuando el modelo de circuitos fue usado de manera efectiva. Sin embargo, fue en 1960 cuando se publicó la obra más importante y que supone los cimientos de la acústica fonética tanto en investigación como en industria: *"Acoustic Theory of Speech Production"* de Gunnar Fant.

2.2. La voz humana

La voz humana es el sonido emitido por un ser humano utilizando sus cuerdas vocales para hablar, cantar, reír, llorar, gritar, etcétera. La frecuencia de la voz humana es específicamente una parte de la producción de sonido por parte de los humanos donde las cuerdas vocales son la fuente de sonido primaria.

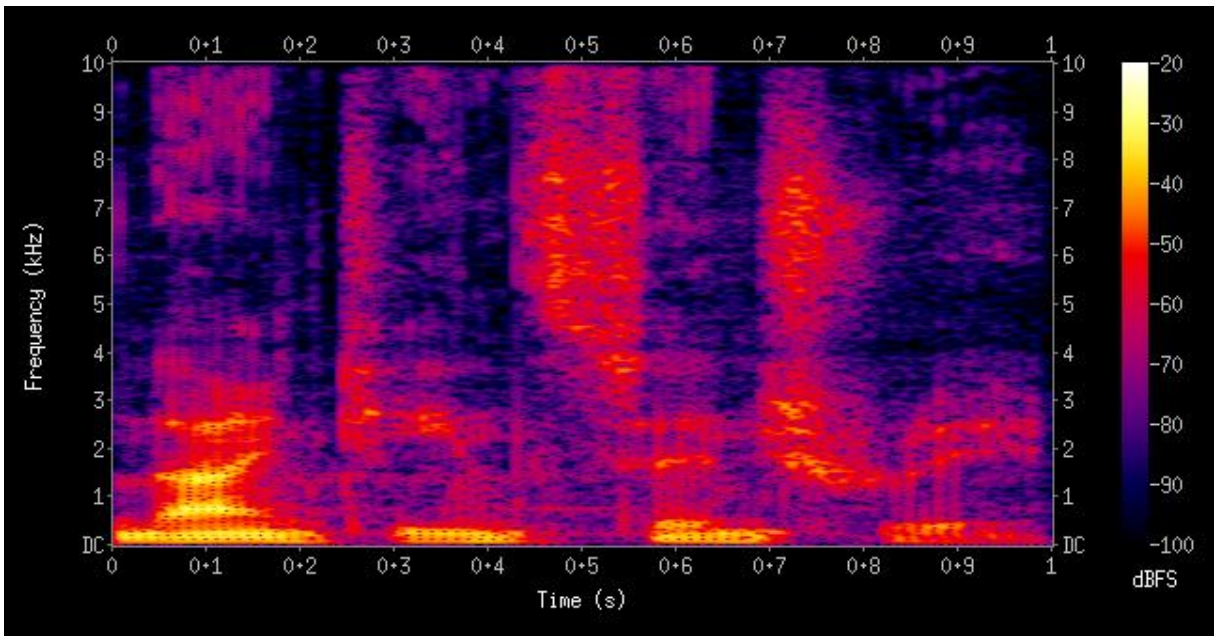


Figura 4: Espectrograma de la voz humana

En general, el mecanismo para la generación de la voz humana puede dividirse en tres partes: los pulmones, las cuerdas vocales en la laringe, y los articuladores[6].

- El pulmón debe producir el flujo de aire y la presión adecuados para hacer vibrar las cuerdas vocales.
- Las cuerdas vocales son una válvula vibrante que transforma el flujo de aire de los pulmones en impulsos audibles que son la fuente del sonido de la laringe. Los músculos de la laringe ajustan la longitud y la tensión de las cuerdas vocales para producir el tono.
- Los articuladores son las partes del tracto oral sobre la laringe, es decir: la lengua, el paladar, la parte interna de las mejillas, los labios, etcétera. Su función es filtrar el

sonido que emana de la laringe y en cierto modo interactuar con el flujo de aire de la laringe para atenuarlo o aumentarlo como fuente de sonido.

Las cuerdas vocales en combinación con los articuladores son capaces de producir sonidos tremendamente complejos. El tono de voz puede modularse para sugerir emociones como la ira, la sorpresa o la felicidad. Los cantantes pueden utilizar su voz como un instrumento para crear música.

2.2.1. Tipos de voces

Los hombres y mujeres adultos generalmente tienen diferentes tamaños de cuerdas vocales, como reflejo de las diferencias en el tamaño de la laringe. Las voces de los hombres adultos generalmente tienen un tono más bajo debido a las cuerdas vocales más largas que las de las mujeres. En general, el tamaño medio de las cuerdas vocales (en adultos) de una mujer está entre 12.5 y 17.5 mm, mientras que en los hombres el tamaño varía entre 17 y 25 mm[7].

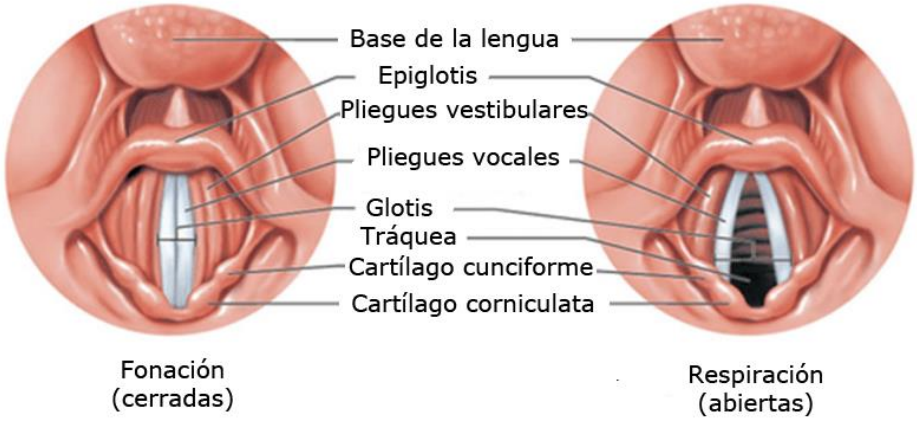


Figura 5: Las cuerdas vocales

Las cuerdas se encuentran en la laringe. Están unidas por atrás, en la parte más cercana a la médula espinal, a los cartílagos aritenoides; y por delante, en la parte bajo la barbilla, al cartílago de la tiroides. No tienen borde externo ya que se unen en los lados con la tráquea, mientras que en la parte interna están libres para vibrar. Tienen una construcción en tres capas consistente en el epitelio, el ligamento vocal y el músculo tiroartenoide, que se encarga de acortar y alargar las cuerdas vocales.

La diferencia en las cuerdas vocales entre hombres y mujeres es la causa de que tengan tonos de voz diferentes. Incluso dentro del mismo sexo, la genética causa variaciones en la voz, clasificándose las voces al cantar en varios tipos (de timbre más bajo a más alto):

- En mujeres: Soprano, Mezzo-Soprano y Contralto.
- En hombres: Tenor, Barítono y Bajo[8].

La voz humana está en un constante estado de cambio y desarrollo, al igual que el propio cuerpo, la voz cambiará al mismo tiempo que la persona envejece. Entre los niños, el rango vocal y las diferencias en el timbre son menos variadas que en adultos, de hecho, en la etapa pre-pubertad tanto los niños como las niñas tienen un rango vocal y timbre equivalentes, ya que el tamaño y la altura de sus laringes es similar, al igual que lo son sus las estructuras de sus laringes. Con la llegada de la pubertad, las voces tanto de mujeres como de hombres se alteran al definirse los ligamentos vocales y endurecerse los cartílagos de la laringe, siendo estos cambios más acentuados en el caso del hombre[9].

2.2.2. Modulación de la voz, fisiología y timbre

El lenguaje hablado hace uso de la habilidad que casi toda la gente en una sociedad concreta para modular dinámicamente ciertos parámetros de la fuente de voz en la laringe de manera consistente. Los parámetros comunicativos (o fonéticos) más importantes son el tono de voz (determinado por la frecuencia de vibración de las cuerdas vocales) y el grado de separación de las cuerdas vocales, llamado aducción de las cuerdas vocales (cuando se juntan) o abducción (cuando se separan)[10].

Si un movimiento de aducción o abducción es lo suficientemente fuerte, la vibración de las cuerdas vocales se detendrá o no llegará a iniciarse. Si el gesto de abducción es parte de un sonido de voz, este sonido se denominará sordo.

El sonido de la voz de cada individuo es único no solamente debido al tamaño y forma de sus cuerdas vocales, sino del tamaño de su cuerpo en conjunto, especialmente el tracto oral y la manera en que los sonidos son formados y articulados. La forma del pecho y el cuello, la posición de la lengua y la tensión de algunos músculos en otros casos no relacionados con la producción de voz pueden ser alterados, resultando en un cambio de tono, volumen o timbre del sonido producido. El sonido también resuena en diferentes partes del cuerpo, de forma que el tamaño y la estructura ósea afectan al sonido producido por un individuo.

Los registros vocales son una serie particular de tonos producidos por el mismo patrón de vibración de las cuerdas vocales. Cualquiera de estos patrones de vibración que aparece en un registro vocal produce ciertos sonidos característicos. En lingüística, el registro del lenguaje combina el tono y la fonación de vocales en un mismo sistema fonológico.

La resonancia vocal es el proceso por el cual la producción básica de la voz es realzada en timbre o intensidad por las cavidades a través de las que pasa en su camino al exterior del aparato fonador. Algunos términos relacionados con el proceso de resonancia son amplificación, enriquecimiento, alargamiento, mejora, intensificación o prolongación. El punto principal es que el resultado final de la resonancia debe ser emitir un mejor sonido[11].

Existen multitud de desórdenes del habla: desde impedimentos hasta lesiones de las cuerdas vocales. Estos desórdenes pueden ser comprendidos de mejor manera mediante el análisis de la voz de la persona que los padece.

2.3. Conceptos técnicos

2.3.1. Ondas de sonido

El sonido puede propagarse por un medio como el aire, agua o sólidos como ondas longitudinales o transversales. Las ondas de sonido se generan por la vibración de una fuente de sonido, como puede ser la membrana de un altavoz o las cuerdas vocales, y esta vibración crea vibraciones en el medio que rodea a la fuente. Al continuar esta vibrando, las vibraciones se propagan alejándose de la fuente según la velocidad del sonido en el medio, creando la onda de sonido.

La propagación del sonido en un medio dado se ve afectada principalmente por tres factores:

- La relación entre la densidad y la presión del medio, afectada por la temperatura. Esta relación afecta a la velocidad del sonido en el medio.
- El movimiento del medio en sí. Si se está moviendo, aumentará o disminuirá la velocidad de propagación de la onda de sonido en función del movimiento. Un ejemplo de esto es la acción del viento en ondas de sonido propagadas por el aire.
- La viscosidad del medio, que determina la atenuación a la que se somete la onda de sonido[12].

Velocidad del sonido en distintos medios (a 20º C)		
Sustancia	Densidad (kg · m-3)	Velocidad (m · s-1)
Aire	1,20	344
Etanol	790	1.200
Benceno	870	1.300
Agua	1.000	1.498
Aluminio	2.700	5.000
Cobre	8.910	3.750
Vidrio	2.300	5.170
Hierro	7.900	5.120

Tabla 1: Velocidad del sonido en distintos medios a 20º C

La transmisión de las ondas de sonido es compleja y presenta multitud de particularidades, en el punto de recepción puede reducirse a dos elementos simples: presión y tiempo. Estos pueden usarse para definir de manera absoluta cualquier sonido. Aun así, para entender el sonido más en profundidad, se separa en sus componentes, que son:

- Frecuencia, o su inversa, amplitud de onda. La frecuencia es el número de veces que se repite un evento por unidad de tiempo, medida en Hertzios[13]. La longitud de onda es la distancia en la onda donde su forma se repite[14].

$$f = \frac{v}{\lambda}$$

- Amplitud, presión de sonido o intensidad. La amplitud es la medida de la variación de una variable periódica en un único período. Puede medirse amplitud de pico a pico, de pico, semi-amplitud o amplitud RMS[15].

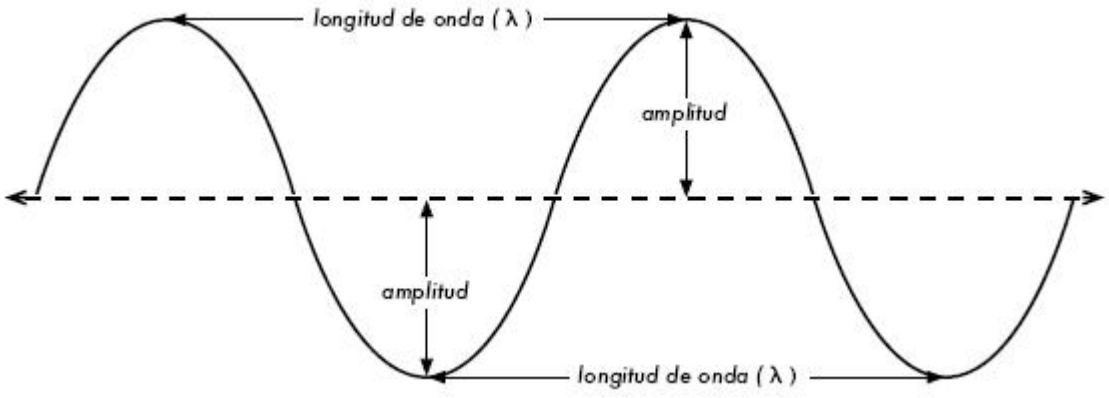


Figura 6: Magnitudes de onda

- Velocidad del sonido.

$$c = \sqrt{\frac{K}{\rho}}$$

Donde K es el módulo de compresibilidad (resistencia a la compresión uniforme)[16] y ρ es la densidad.

- Dirección.

El sonido es perceptible para los seres humano en frecuencias que van desde los 20 Hz hasta los 20 kHz. En el aire, eso corresponde a longitudes de onda en el rango de 17 m a 17 mm.

El espectro de un sonido muestra las diferentes frecuencias que componen el mismo. La mayoría de los sonidos están compuestos de una compleja mezcla de vibraciones, y el espectro de sonido representa la vibración de cada frecuencia individual. Se presenta normalmente como un gráfico de intensidad o presión de sonido en función de la frecuencia.

2.3.2. Intensidad de sonido y presión de sonido

La intensidad de sonido, también conocida como intensidad acústica, es definida como la energía que transportan las ondas de sonido por unidad de área[17]. Las unidades del sistema internacional para la intensidad, incluyendo intensidad acústica, es el vatio por metro cuadrado (W/m^2). Muchas medidas de intensidad acústica se hacen de forma relativa al umbral estándar de intensidad audible I_0:

$$I_0 = 10^{-12} vatios/m^2 = 10^{-16} vatios/cm^2$$

Generalmente se usa la escala en decibelios (dB) para medir la intensidad acústica. Los decibelios miden la relación entre una intensidad dada y el umbral de intensidad audible, de manera que este umbral toma el valor de 0 dB:

$$I\ (dB) = 10 \log_{10} \left[\frac{I}{I_0}\right]$$

Matemáticamente, la intensidad de sonido se define como:

$$I = pv$$

Donde:

- *p* es la presión de sonido, es decir, el cambio en la presión atmosférica causado por una onda de sonido.
- v es la velocidad de la partícula en el medio.

Tanto la intensidad acústica como la velocidad de la partícula en el medio son vectores, lo que significa que tienen dirección y sentido además de magnitud. La dirección de la intensidad es la media de la dirección en la que fluye la energía. La intensidad media de un sonido en un período T es:

$$I = \frac{1}{T} \int_0^T p(t)v(t)dt$$

O también:

$$I = 2\pi^2 n^2 A^2 \rho v$$

Donde:

- n es la frecuencia del sonido.
- A es la amplitud de la onda de sonido.
- v es la velocidad del sonido.
- P es la densidad del medio donde el sonido viaja.

La presión del sonido o presión acústica es la desviación en la presión ambiente (presión media o de equilibrio) causada por una onda de sonido. En el aire, esta presión puede ser medida por un micrófono. Sus unidades de medida en el SI son los pascales (Pa)[18].

Matemáticamente, la presión de sonido (p) se define como:

$$p_{total} = p_{estática} + p$$

2.3.3. Elementos de la percepción del sonido

2.3.3.1. Altura

La altura es una cualidad perceptiva que permite distinguir un sonido agudo de uno grave. Las ondas de sonido en sí no tienen esta propiedad, es decir, se trata de una propiedad psicoacústica; pero sus oscilaciones pueden ser medidas para obtener una frecuencia[19]. Esta propiedad es la que da lugar a la clasificación de los sonidos en las escalas de notas musicales.

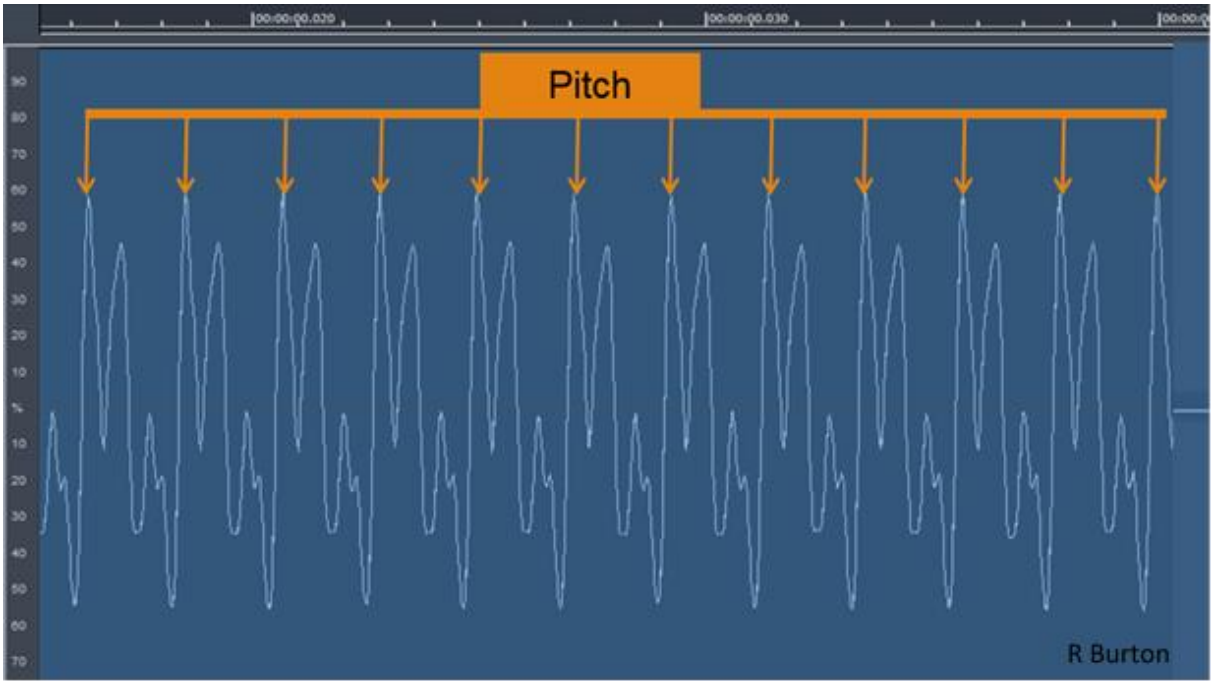


Figura 7: Percepción de la altura

2.3.3.2. Duración

La duración hace referencia a cuan corto o largo es un sonido y se relaciona con el comienzo y final de las señales creadas por las respuestas nerviosas a un sonido (Figura 8). Un sonido dura desde que es detectado hasta que se identifica que ha cambiado o finalizado. No siempre está relacionado con la duración física del sonido, ya que, por ejemplo, en ambientes ruidosos, sonidos intermitentes pueden ser percibidos como continuos ya que el ambiente ruidoso hace que las interrupciones se pierdan[20].

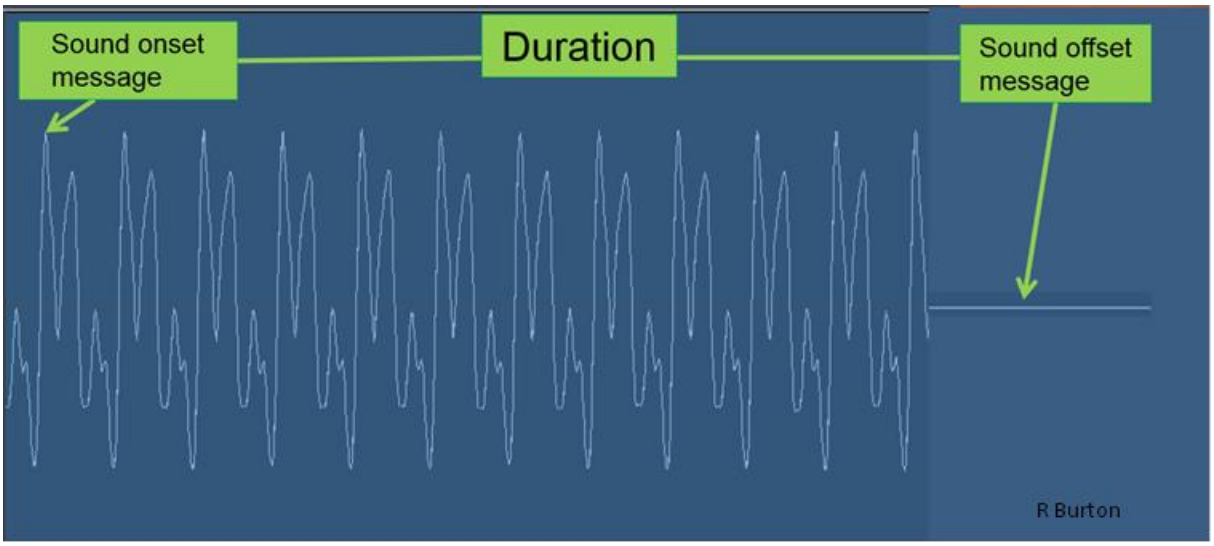


Figura 8: Percepción de la duración

2.3.3.3. Volumen

El volumen de un sonido es percibido como lo alto o bajo que resulta el mismo, y se relaciona con el número total de estimulaciones de los nervios auditivos en períodos de tiempo cíclicos cortos. Esto implica que a duraciones cortas (inferiores a 200 ms), un sonido corto puede sonar más suave que uno largo, incluso cuando tienen la misma intensidad. Las señales de alto volumen crean un mayor empuje en la membrana basilar, estimulando más nervios. Una señal compleja dispara más nervios que una simple, y por lo tanto suena más alta (Figura 9).

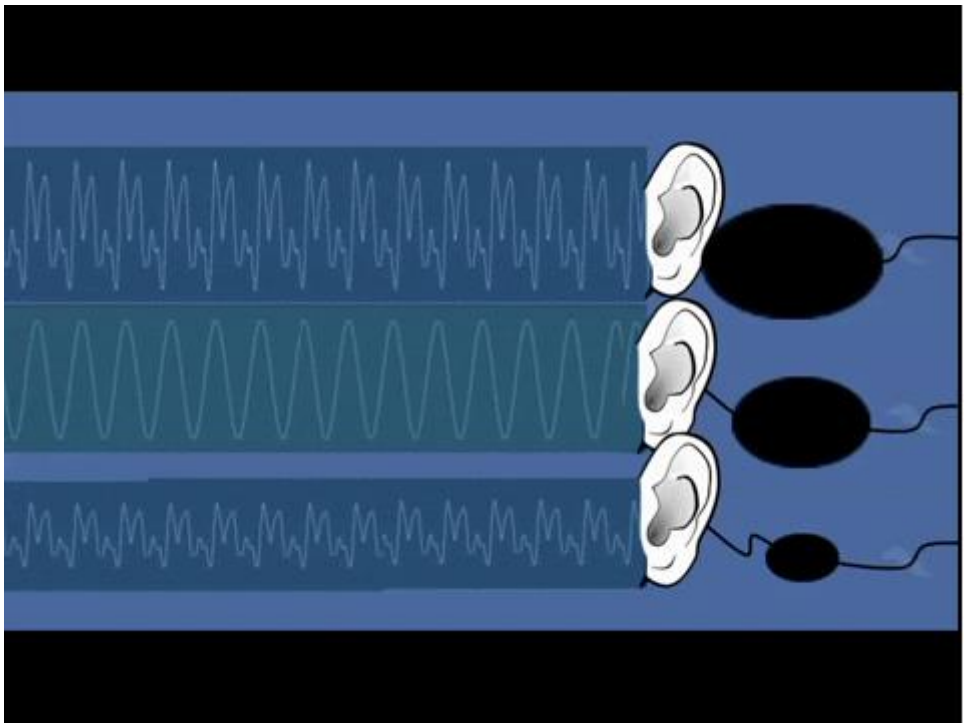


Figura 9: Percepción del volumen

2.3.3.4. Timbre

El timbre se percibe como la cualidad de los diferentes sonidos y representa la localización pre-consciente de una identidad sonora con un sonido. Esta identidad está basada en la información adquirida por los transitorios de frecuencia, el ruido, la inestabilidad, la altura percibida y la difusión e intensidad de los armónicos en una ventana de tiempo. La forma en que un sonido cambia con el tiempo (Figura 10) proporciona la mayoría de la información para la identificación del timbre. Como se ve en la figura, en un período muy corto las señales son muy similares, sin embargo, si lo ampliamos, las diferencias son notables[21].

Existen varias definiciones de lo que es un armónico, pero en acústica el armónico de una onda se refiere a un componente sinusoidal de una señal. Los armónicos son uno de los parámetros que generan el timbre de una fuente de sonido. Los armónicos más altos son inaudibles y lo que da diferentes timbres a los instrumentos (o las voces) es la amplitud y localización de los primeros armónicos.

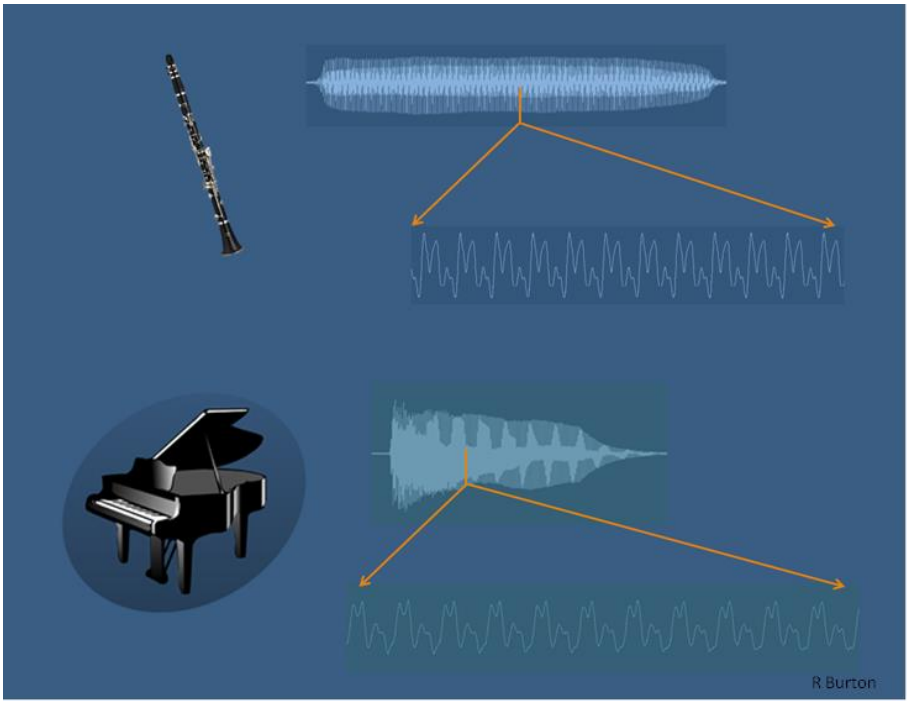


Figura 10: Percepción del timbre

2.3.3.5. Otras propiedades

- Textura sónica: relacionada con las diferentes fuentes de sonido y la interacción entre las mismas.
- Localización espacial: representa la localización cognitiva de un sonido en un ambiente determinado. Depende principalmente de la localización del emisor y el receptor.

2.4. Vocales

La palabra vocal procede del vocablo latino *vocalis* que significa relativo a la voz[22]. En castellano, esta palabra se usa tanto para los sonidos vocálicos como los símbolos que se utilizar para representarlos. En fonética, una vocal es un sonido del lenguaje verbal con dos definiciones complementarias:

- En la definición fonética, una vocal es un sonido pronunciad con el tracto vocal abierto, de forma que la lengua no toca los labios, dientes o paladar. En las consonantes, por otro lado, siempre se produce una constricción o cierre en algún punto a lo largo del tracto vocal.
- En la definición fonológica, una vocal es definida como sonido que forma el pico de una sílaba.

En los lenguajes hablados, las vocales fonéticas normalmente forman el pico (núcleo) de las sílabas. Aunque algunos lenguajes permiten que otros sonidos formen el núcleo de una sílaba, es el caso de las llamadas consonantes silábicas.

Tradicionalmente se ha dicho que las cualidades de las vocales se deben a su articulación. Daniel Jones desarrolló el sistema cardinal de vocales, que describe las mismas en términos de las características de la lengua al pronunciarlas: altura (dimensión vertical), anchura (horizontal) y redondez de los labios. Estos tres parámetros se indican en el esquema cuadrilateral IPA (figura 11), aunque la producción de vocales no sólo depende de los mismos, sino también la posición del velo nasal, el tipo de vibración de las cuerdas vocales y la posición de la raíz de la lengua. Esta concepción se ha sabido obsoleta desde principios del siglo XX, y ya el propio manual IPA dice que “el cuadrilátero de las vocales deber ser visto como una abstracción, no un mapeado directo de la posición de la lengua”[23]. No obstante, este concepto de que las cualidades de la vocal vienen determinadas por la posición de la lengua y el redondeo de los labios sigue siendo usada en pedagogía, ya que es una explicación intuitiva de cómo se distinguen las vocales.

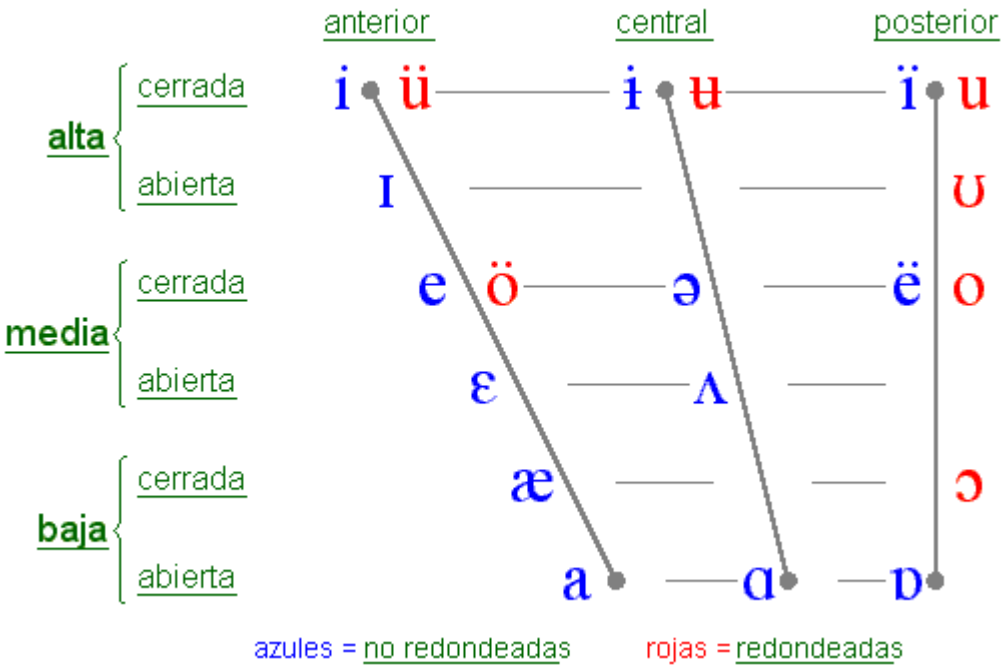


Figura 11: Tabla IPA de clasificación de las vocales

La acústica de las vocales está ampliamente estudiada. Las diferentes cualidades de las vocales son determinadas de los formantes, de los que se hablará en detalle más adelante. La acústica de las vocales se puede ver fácilmente con los espectrogramas, que muestran la energía acústica en cada frecuencia y como esta cambia con el tiempo.

Los aspectos de la prosodia (parte de la lingüística que analiza los acentos, la entonación y los tonos[24]) se asocian a las sílabas, no a la propia vocal.

Un sonido vocálico cuyas cualidades no cambian durante la duración de la vocal se llama monoptongo, también conocido como vocal pura o estable. Sin embargo, un sonido vocálico que cambia de una cualidad a otra se llama diptongo, y si cambia sucesivamente por tres cualidades, triptongo. Todos los lenguajes tienen monoptongos y muchos tienen diptongos, pero los triptongos o sonidos vocálicos con aún más cambios de cualidades son relativamente raros. En español existen los tres; el sonido vocálico en *sin* es un monoptongo /ɪ/, el sonido de *soy* es un diptongo /ɔɪ/, y el de *buey* un triptongo /ueɪ/. Los diptongos y triptongos se distinguen de secuencias de monoptongos si el sonido vocálico debe ser analizado en fonemas diferentes o no.

2.4.1. Formantes

La palabra formante cuenta con varias definiciones:

- Formante fue utilizado por James Jeans en 1938[25] para hablar de la colección de armónicos de una nota que son aumentados por resonancia.
- Fue definido por Gunnar Fant en 1960[26]: "Los picos espectrales del espectro de sonido |P(f)| se llaman formantes".
- Benade (1976)[27] escribió: "Los picos observados en la envolvente del espectro se llaman formantes".
- En terminología acústica estándar, la Acoustical Society of América definió los formantes en 1994 de la siguiente manera[28]: "En un sonido complejo, un rango de frecuencias donde hay un máximo absoluto o relativo en el espectro de sonido. Unidades, hertzios (Hz). NOTA: La frecuencia en el máximo es la frecuencia del formante.

Tras la definición de Fant, él mismo define las frecuencias de resonancia del tracto vocal en función de la ganancia T(f) del tracto vocal[26]: "La localización de la frecuencia en un máximo en |T(f)|, es decir, la frecuencia de resonancia, es muy cercana al correspondiente máximo en el espectro |P(f)| del sonido completo". Más adelante dice que[26]: ¨Conceptualmente estos deberían mantenerse separados, pero en la mayoría de los casos la frecuencia de resonancia y la frecuencia del formante pueden utilizarse como sinónimos". Aquí se plantea pues un problema: resonancia y formantes son conceptualmente diferentes, y sus frecuencias son sólo aproximadamente iguales. En cualquier caso, en la mayoría de circunstancias, el formante es la única información que se tiene sobre la resonancia. Por esta razón, en análisis de discurso estos términos se utilizan de manera indistinta.

Existe incluso un tercer significado en la investigación del habla. La acústica del tracto vocal está modelada usando un modelo matemático de un filtro[29]. Las frecuencias de los polos de este modelo son también cercanas a las frecuencias de los formantes. Como resultado, algunos investigadores se refieren a las frecuencias de los polos como formantes. De este modo, para algunos investigadores formante se refiere a un pico en la envolvente del espectro (una propiedad del sonido de la voz), para otros a la resonancia del tracto vocal (una propiedad física del tracto), mientras que para un tercer grupo se refiere a un polo en un modelo matemático de filtro (la propiedad de un modelo).

En el campo de la acústica, formante se queda con su definición original: un pico ancho en la envolvente del espectro de una onda de sonido. Cuando se está haciendo referencia a un formante sobre los 400 Hz en el sonido de un cuerno francés, es claramente un pico en la envolvente del espectro, no una de las resonancias.

La información requerida para distinguir entre sonidos de voz puede ser representada de manera puramente cuantitativa especificando picos en el espectro de amplitud/frecuencia. Muchos de estos formantes están producidos por resonancia de tubo y cámara, pero algunos tonos de derivan del efecto Venturi (por el cual un fluido disminuye su presión cuando aumenta su velocidad al pasar por una zona de sección menor)[30]. El formante de menor frecuencia es llamado F_1, el segundo F_2 y el tercero F_3. Normalmente el uso de los dos primeros es suficiente para distinguir la vocal.

Vocal (IPA)	Formante F_1 (Hz)	Formante F_1 (Hz)	Diferencia F_1-F_2 (Hz)
i	240	2400	2160
y	235	2100	1865
e	390	2300	1910
ø	370	1900	1530
ɛ	610	1900	1290
œ	585	1710	1125
a	850	1610	760
œ	820	1530	710
ɑ	750	940	190
ɒ	700	760	60
ʌ	600	1170	570
ɔ	500	700	200
ɤ	460	1310	850
o	360	640	280
ɯ	300	1390	1090
u	250	595	345

Tabla 2: Media de los formantes en vocales[31]

Las consonantes oclusivas (y en cierto grado también las fricativas) modifican la localización de formantes en las bocales alrededor de ellas. Los sonidos bilabiales (como la /b/ en “barco”) causan una bajada en los formantes; los sonidos velares (como la /k/) casi siempre muestran unos F_2 y F_3 prácticamente juntos; los sonidos alveolares (como /t/ y /d/) causan

menos cambios sistemáticos en los formantes de las vocales vecinas, aunque depende de qué vocal esté presente.

Si la frecuencia fundamental de la vibración subyacente es más alta que la frecuencia de resonancia (frecuencia característica de un cuerpo o un sistema que alcanza el grado máximo de oscilación[32]), entonces el formante generalmente impartido para esa resonancia se pierde. Para visualizar los formantes se utilizan los espectrogramas (Figura 12).

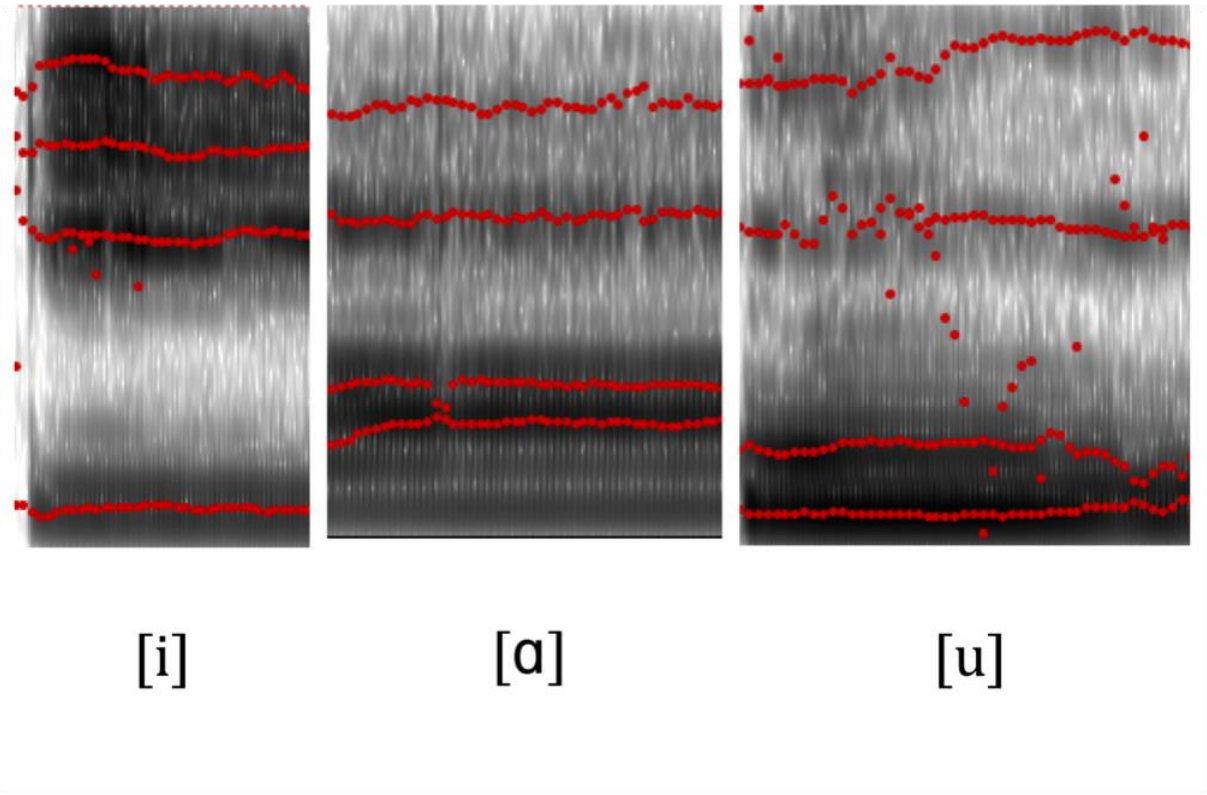


Figura 12: Formantes vistos en un espectrograma

3. Plataforma de desarrollo: Praat

3.1. Introducción

Praat es un software para análisis y síntesis de voz escrito por Paul Boersma y David Weenik del Departamento de Fonética de la Universidad de Ámsterdam[33]. Este software es compatible con diferentes sistemas operativos: UNIX, Linux, MAC y Windows. La primera versión fue lanzada en 5 de diciembre de 1995, mientras que la última versión, la que se ha usado para desarrollar este proyecto, es la 6.0.29 de mayo de este mismo año.

Figura 13: Logo de Praat

En su parte analítica, el software permite trabajar tanto con archivos de audio, de los que soporta las extensiones más comunes (MP3, WAV, AIFF, etcétera); como realizar grabaciones utilizando un micrófono integrado o micrófonos externos al ordenador, siendo posible también la modificación de los parámetros de la grabación (mono/estéreo, frecuencia de muestreo, etcétera). Una vez importado el sonido, el programa ofrece múltiples formas de manipular el mismo y anotar las diferentes partes que lo componen para facilitar su posterior estudio.

Por otra parte, Praat no es un sintetizador de texto a voz, aunque sí que permite la generación de multitud de sonidos: pueden usarse fórmulas para generar sonido, como una onda senoidal o ruido blanco; pueden crearse sonidos a partir de otro tipo de datos, como un tren de pulsos generado a partir de un contorno de tono; pueden crearse sonidos similares al habla a partir de los datos de intensidad, tono y formantes; puede hacerse síntesis articulatoria, es decir, generar un sonido resultado de la contracción de determinados

músculos; o pueden crearse nuevos sonidos a partir del filtrado o mejora de sonidos existentes.

Por último, Praat permite la creación de scripts dentro del propio software, es decir, permite programar una serie de tareas que se realicen de forma automática al ejecutar una instancia. Utiliza un lenguaje de programación de propósito general con capacidades especiales para simular elecciones en los menús.

3.2. Componentes principales

3.2.1. Ventanas principales: Praat Objects y Praat Picture

Al abrir Praat, se muestran en pantalla las dos ventanas predeterminadas: la ventana principal o ventana de objetos (bajo el nombre de Praat Objects) y la ventana de gráficos (Praat Picture). La ventana principal muestra una lista de todos los objetos con los que trabaja el programa, que serán explicados en profundidad más adelante.

Desde la ventana de objetos (Figura 14) pueden llevarse a cabo todas las operaciones que permite Praat: pueden importarse y exportarse sonidos, generar nuevos objetos a partir de los existentes, abrir o crear scripts, abrir la ventana de edición, la ayuda, etcétera.

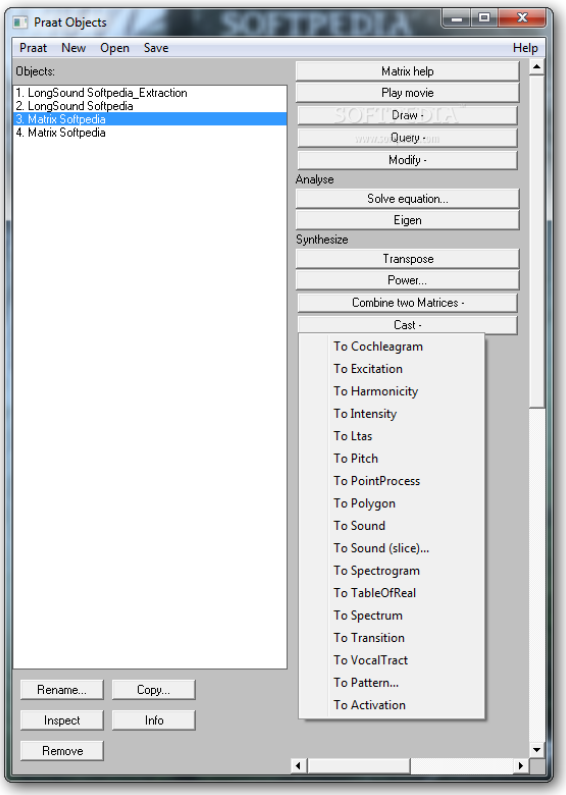


Figura 14: Ventana *Praat Objects*

Los botones de la parte superior permiten la importación o creación de objetos de audio, scripts, o lanzar la ayuda. Los de la parte inferior permiten manejar los objetos: copiarlos, borrarlos, renombrarlos, etcétera. Por último, las opciones de la parte derecha de la ventana son los correspondientes a las funcionalidades de Praat en sí, y cambiarán según el tipo de objeto que se tenga seleccionado.

Por otra parte, la ventana de gráficos (Figura 15) permite mostrar por pantalla todos los tipos de diagrama que permite elaborar Praat: desde una simple visualización de la forma de onda hasta representaciones más complejas como diagramas de formantes de vocales. También permite la edición de estos gráficos y su impresión para ser incluidos en documentos o artículos.

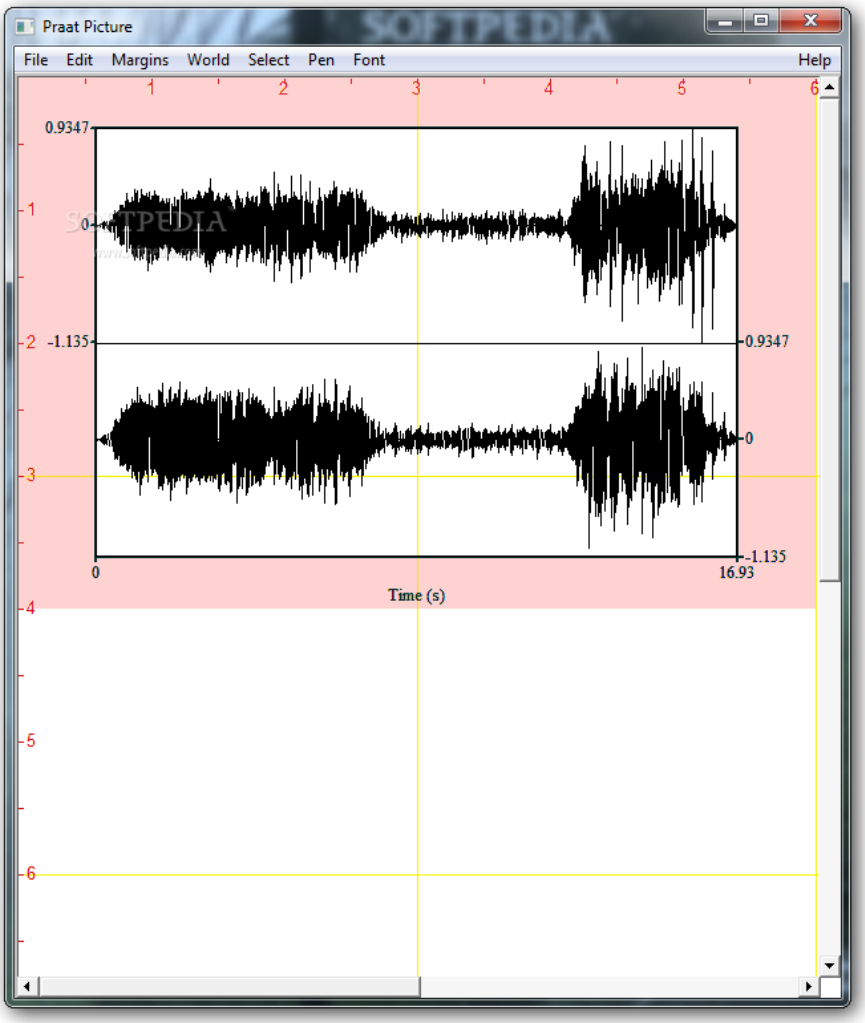


Figura 15: Ventana *Praat Picture*

3.2.2. Ventana de ayuda

La ventana de ayuda de Praat se lanza desde la ventana de objetos, seleccionando la opción Help. Esta ventana contiene un manual completo de uso de Praat de unas 800 páginas, que cubre todas las funcionalidades del programa. La ventana de ayuda cuenta con una función de búsqueda por texto, y además los artículos de la misma están interconectados con otros artículos relacionados mediante hiperlinks. La ayuda se complementa con las FAQ de la página principal y un grupo de usuarios de Praat en Internet.

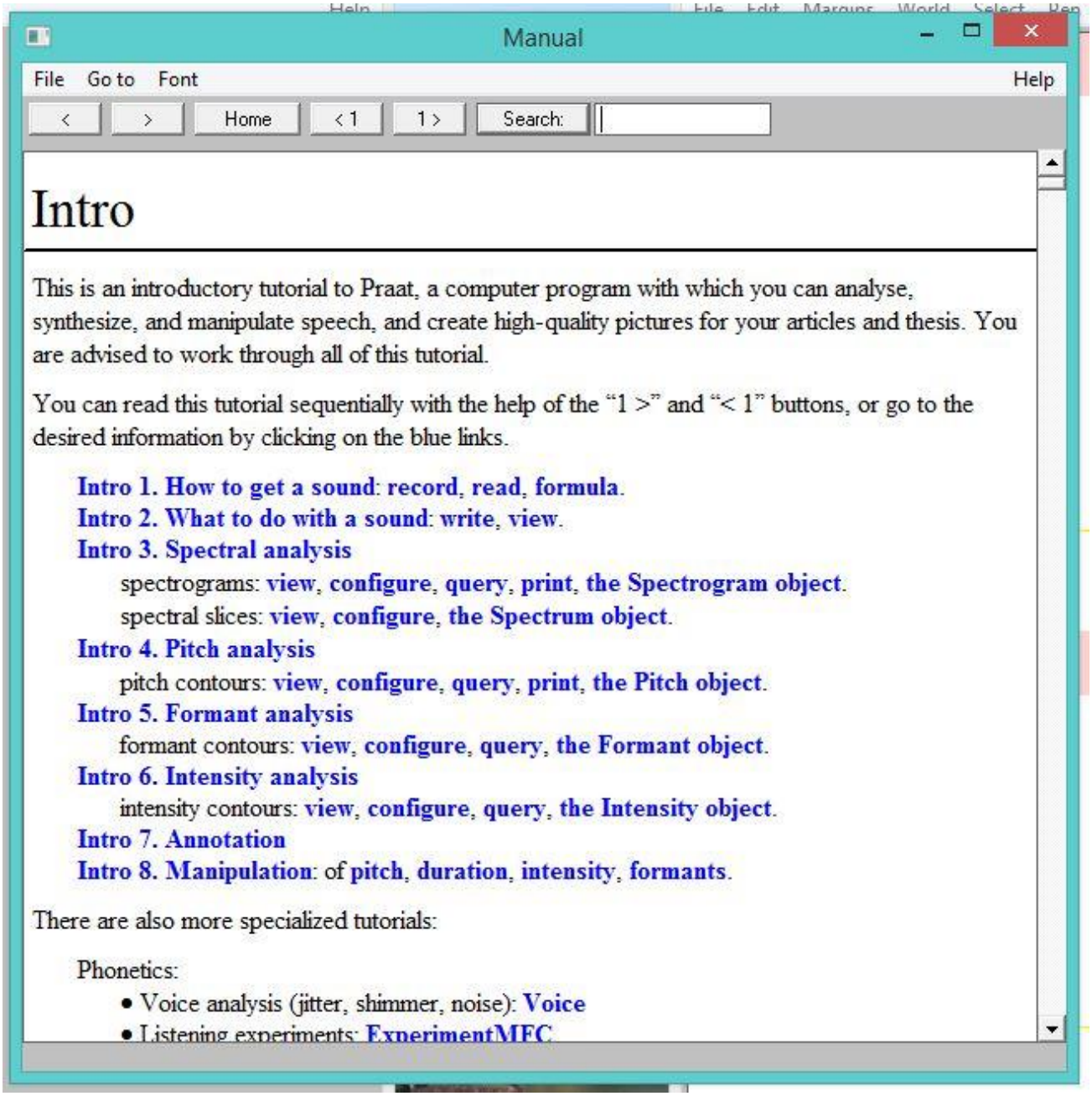


Figura 16: Ventana de ayuda de Praat

3.2.3. Ventana View & Edit

La ventana View & Edit de Praat (Figura 17) es, como su propio nombre indica, una ventana que permite visualizar y editar un objeto de sonido. Esta ventana se despliega desde la principal, cuando teniendo un objeto de sonido seleccionado se elige la opción View & Edit en el menú de la parte derecha.

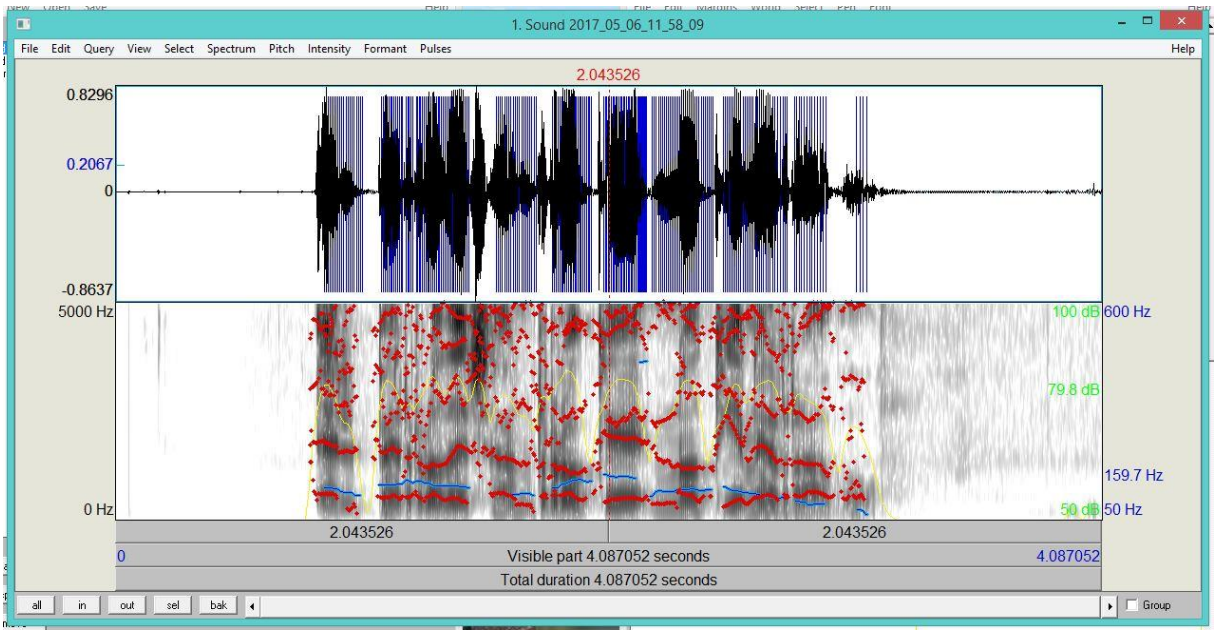


Figura 17: Ventana *View & Edit*

Esta ventana muestra la forma de onda del objeto de sonido seleccionado, y su espectrograma justo debajo. Praat permite además visualizar en esta ventana y sobre el espectrograma información gráfica de los formantes (en rojo), la intensidad (en amarillo, no visible) y el tono (en azul) de la señal de audio. Esta información puede añadirse o eliminarse de la visualización, o pueden aplicarse algoritmos o cambiar ajustes de cualquiera de las visualizaciones, utilizando los botones de la parte superior de la ventana.

Los botones de duración de la parte inferior de la pantalla, permiten aislar y reproducir partes concretas del sonido, etiquetarlas, ampliarlas o reducirlas o incluso extraerlas como nuevos objetos de sonido y guardarlas en un archivo.

3.2.4. Ventanas de scripts

La ventana de scripts se abre automáticamente cuando se selecciona la opción de abrir o crear un nuevo script en la ventana principal. Esta ventana es un editor de texto sencillo, similar al editor de texto predeterminado del sistema operativo que se esté utilizando (en caso de Windows sería el Bloc de Notas). Desde esta misma ventana se ejecutará el script (o parte de él) mediante la opción *Run* de la parte superior.

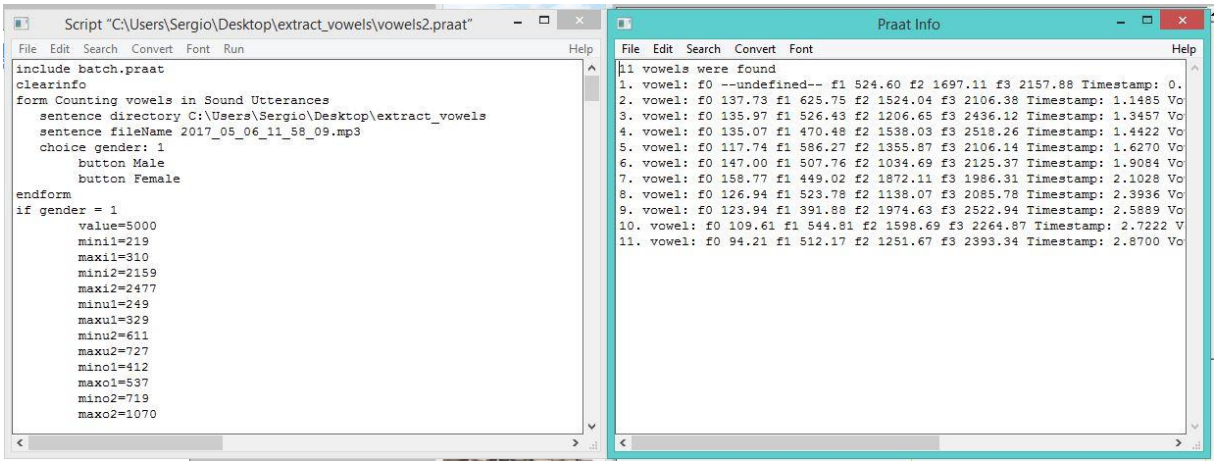


Figura 18: Ventana de scripts

Mediante los scripts también se puede requerir que el usuario introduzca información. Para ello se utilizan las ventanas de formulario, que solicitan que el usuario introduzca información en forma de texto, números, o seleccione una o varias opciones que modificarán los parámetros con los que trabaja el script.

Por último, la ventana de información permite mostrar en pantalla todo tipo de información en forma de texto: desde simplemente mensajes para el usuario, hasta los valores de variables calculados mediante el script.

3.3. Objetos

Como se ha mencionado anteriormente, Praat lleva a cabo todos sus análisis, algoritmos, etcétera, sobre los llamados objetos. Estos objetos son las instancias que aparecen listadas en la ventana principal de Praat, y pueden ser de diferentes tipos. Los objetos son muy variados, y en esta memoria solo se mencionan los que son de interés en el marco de este proyecto.

3.3.1. Propósito general

Son los objetos principales de Praat, la base sobre la que se ejecutan los análisis y se crean el resto de objetos. Entre ellos destacan los siguientes:

- Sound: objeto de sonido de Praat. Puede ser creado mediante una grabación mono o estéreo, sintetizado, o importado de un archivo que Praat pueda leer: WAV, AIFF/AIFC, FLAC, NIST o MP3. También puede ser guardado en los mismos formatos, excepto MP3.
- LongSound: es análogo al objeto de sonido, pero para grabaciones de mayor duración, es decir, varios minutos.
- Matrix: representa una función z (x, y) en el dominio [x_{min}, x_{max}] × [y_{min}, y_{max}]. El dominio es muestreado en las direcciones de x e y con intervalos de muestreo constante (dx y dy) a lo largo de cada dirección.
- Polygon: una secuencia de puntos (x_i, y_i) en un espacio bidimensional.
- PointProcess: una secuencia de puntos t_i en tiempo, definida en un dominio [t_{min}, t_{max}]. El índice i va desde 1 hasta el número de puntos. Los puntos se ordenan por tiempo, es decir, $t_{i+1} > t_i$.
- Strings: representa una lista de cadenas ordenada.
- Table, TableOfReal, Distribution, PairDistribution: objetos para cálculo estadístico en Praat.
- Permutation: objeto con un número n de elementos con un orden 1,2...n. Se usa para modificar el orden de estos elementos.
- ParamCurve: es una secuencia de puntos con marca de tiempo (x (t_i), y (t_i)) en un espacio bidimensional.

3.3.2. Análisis periódico

- Pitch: este objeto representa el tono en función del tiempo. Está divido en muestras con periodos iguales. Cada muestra contiene información acerca de los candidatos que contiene, y estos almacenan información acerca de su frecuencia en hertzios si contienen voz, o un cero si no la tienen. Por tanto, este objeto se utiliza para ver qué partes de un sonido tienen voz y cuáles no.
- Harmonicity: representa el grado de periodicidad acústica, también llamado ratio de armónicos a ruido. La armonicidad se expresa en dB: si el 99% de la energía de

la señal está en la parte periódica, y el 1% es ruido, el HNR es 10*log10(99/1) = 20 dB. Un HNR de o dB quiere decir que existe la misma energía en los armónicos y en el ruido. Esto puede utilizarse para medir la calidad de la voz.

- Intensity: representa el valor de la intensidad en puntos de tiempo separados linealmente $t_i = t_1 + (i - 1)\, dt$, con valores en dB SPL, por ejemplo, dB relativos a $2 \cdot 10^{-5}$ pascales, que es el umbral auditivo normalizado para una onda senoidal de 1000 Hz.
- IntensityTier: este objeto almacena puntos con información acerca de la intensidad y la marca de tiempo en ese punto.

3.3.3. Análisis espectral

- Spectrum: representa el espectro complejo en función de la frecuencia. Si el espectro fue creado desde un sonido, los valores complejos se expresan en Pa/Hz.
- LTAS: representa la densidad espectral de potencia (La energía media en un sonido en un rango de tiempo y frecuencia concretos, expresado en Pa^2/Hz)[34] en función de la frecuencia, expresada en dB/Hz relativo a $2 \cdot 10^{-5}$ Pa.
- Spectogram, BarkSpectogram, MelSpectogram: es la representación acústica tiempo-frecuencia de un sonido: la densidad espectral de potencia. Está dividido en muestras espaciadas por el mismo tiempo t_i y frecuencia f_j. Bark y Mel modifican la escala de frecuencias.
- Formants: representa la estructura espectral en función del tiempo, el contorno de los formantes. Está dividido en muestras separadas por periodos espaciados igualmente, y cada muestra contiene información de frecuencia y ancho de banda de varios formantes.
- Cochleagram: representa el patrón de excitación de la membrana basilar del oído interno en función del tiempo.

3.3.4. Manipulación del sonido

- PitchTier y FormantGrid: almacenan la misma información que los objetos Pitch y Formant, añadiendo información de marca de tiempo.
- Manipulation: un objeto que permite cambiar los contornos de tono y la duración de otros objetos.
- Duration: contiene un número de puntos del tipo (tiempo, duración), donde duración se interpreta de manera relativa (la duración de un sonido manipulado comparada con la duración del original).

3.3.5. Etiquetado y segmentación

- TextGrid: es el objeto de Praat utilizado para anotar los sonidos. Se genera a partir de un objeto Sound o LongSound, desde cero, o mediante la unión de varios

TextGrid anteriores. Contiene un determinado número de niveles donde anotar, que pueden ser de dos tipos: nivel de intervalos, que es una secuencia de intervalos etiquetados con uniones entre ellos; o nivel de puntos, una secuencia de puntos etiquetados. Estos niveles se crean al crear el TextGrid, y pueden ser editados mediante la ventana del editor.

4. Diseño de la solución técnica

4.1. Planteamiento del problema

El problema que se presenta en este caso es el siguiente: se tiene un archivo de audio que contiene una grabación de voz, y se requiere detectar las vocales que aparecen en esta grabación de voz, así como calcular los formantes de las mismas.

La grabación puede presentarse en cualquiera de los formatos soportados por el programa, o grabarse directamente usando el mismo si se dispone de un micrófono. La información debe ser tanto presentada visualmente como almacenada para facilitar su posterior almacenamiento y procesado.

Inicialmente se planteó la opción de que fuese un análisis en tiempo real, es decir, el script iría detectando las vocales según el usuario las fuese diciendo, pero el análisis en tiempo real no está entre las funcionalidades de Praat. Por lo tanto, lo más cercano a esto sería la grabación de voz y el posterior procesado una vez la voz se haya grabado.

Como es obvio, cuanto mejor sea la calidad del audio mejor funcionará el script. Es decir, cuanto más clara sea la voz y menos ruido de fondo exista, más precisa será la detección y más ajustado el cálculo de los formantes.

Para realizar la detección de las vocales no influye la persona que está hablando, sin embargo, para el cálculo de formantes, la frecuencia de cálculo máxima deberá ser modificada si se tiene una grabación de un hombre o de una mujer.

4.2. Planteamiento de la solución

Desde el principio, y una vez descartada la idea de hacer análisis en tiempo real, se optó por escribir un script que realice la misma detección y análisis partiendo de diferentes fuentes. El primero de ellos realizará la detección y análisis desde un archivo presente en el ordenador, mientras que el segundo permitirá la grabación de voz directamente desde el programa.

Ya que como se ha dicho anteriormente el script está pensado para ser usado por personas que no necesariamente tengan amplios conocimientos del programa Praat, una ventana de diálogo al lanzar el script recogerá los datos necesarios para la ejecución. Estos datos serán: tipo de análisis a ejecutar (grabación o desde archivo), directorio de trabajo, nombre de archivo, extensión del archivo, duración de la grabación y género de la persona que habla.

La selección de modo de funcionamiento, formato de archivo y género del hablante se hará mediante botones, pudiendo seleccionar sólo una de las opciones en cada caso. La información requerida vendrá cumplimentada con unos valores por defecto, con el objetivo de ayudar a que los campos sean rellenados correctamente, ya que en caso contrario el script arrojará un error. Después del error la ventana de selección sigue estando presente, por lo que es posible modificar la información errónea y relanzar el script sin necesidad de volver a lanzarlo. Los valores por defecto son:

- Directorio: escritorio.
- Nombre del archivo: Recording. Este campo debe ser obligatoriamente modificado en el caso de trabajar desde archivo, ya que este archivo no tiene por qué existir.
- Formato de archivo: elige el formato de entrada o de salida del audio, dependiendo del tipo de análisis que se lleve a cabo. En caso de ser desde archivo, hay que seleccionar la misma extensión que tenga el archivo que se quiera analizar, y en caso de grabación, se seleccionará el formato en el que se guardará el audio grabado. Como Praat no soporta grabación a MP3, si se selecciona esta opción se guardará como WAV.
- Tiempo de grabación: 5 segundos.

Al pulsar el botón Accept de la ventana de diálogo, el script desde archivo realizará el análisis, mientras que el de grabación abrirá una ventana de información mostrando el texto "*Please speak for n seconds*" donde n será el valor seleccionado en la ventana de diálogo. Esta ventana estará abierta lo que dure la grabación, y se cerrará automáticamente al finalizar esta y comenzar el análisis, que se hará de manera automática nada más finalizar.

Una vez completado el análisis, los resultados se mostrarán del mismo modo para ambas versiones del script. Se abrirán dos ventanas automáticamente, mostrando información visual y numérica de las vocales encontradas:

- Ventana de información: mostrará los resultados del análisis en formato de texto. La primera línea indicará el número de vocales totales encontradas con la frase "*N vowels were found*". En las líneas posteriores se mostrará la siguiente información para cada una de las vocales encontradas:
    - F0: frecuencia fundamental en la vocal.
    - F1, F2 y F3: formantes de la vocal.
    - Timestamp: punto de tiempo donde se encuentra la vocal.
    - El porcentaje de probabilidad de que sea cierta vocal. Aquí se distinguen dos casos:
        - Tanto F1 como F2 están en los rangos correspondientes a una vocal: el porcentaje de que sea esa vocal es del 100%, por tanto, se mostrará la vocal correspondiente y el porcentaje del 100%.
        - Cualquier otro caso: se calculan las probabilidades que tiene de ser cada vocal, en función de la pertenencia a los rangos y la cercanía a

ellos, y se muestran los porcentajes calculados para todas las vocales.

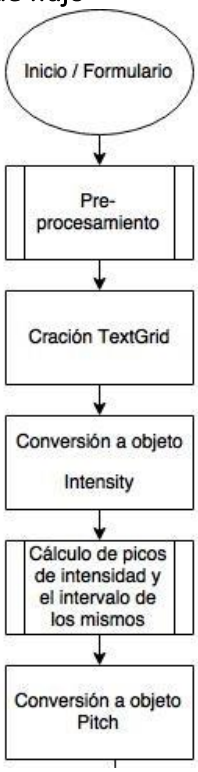
- Ventana de editor: las características de esta ventana fueron explicadas en el apartado 3.2.3. En este caso, mostrará la forma de onda del sonido que se ha analizado, y su espectrograma justo debajo. En la tercera capa, mostrará un nivel de puntos de un TextGrid llamado "*vowels*", teniendo marcados aquellos puntos donde han sido encontradas las vocales y etiquetadas con la vocal propiamente dicha, en caso de que se haya encontrado coincidencia en los formantes, o con la vocal de mayor porcentaje si no se encontró coincidencia exacta.

Por último, los scripts guardarán automáticamente los archivos necesarios para replicar los resultados obtenidos por ellos mismos en cualquier momento sin necesidad de ser ejecutados de nuevo:

- El TextGrid se guardará como un archivo .vowels.TextGrid, con el mismo nombre que el archivo analizado, o el nombre seleccionado para la grabación.
- La información mostrada en la ventana de texto se guardará en un archivo de texto (.txt en Windows), con el mismo nombre que el archivo analizado, o el nombre seleccionado para la grabación.
- La grabación de voz en el caso del script de grabación se guardará en un archivo con el nombre y la extensión seleccionados en el cuadro de diálogo.

4.3. Diseño

4.3.1. Diagrama de flujo



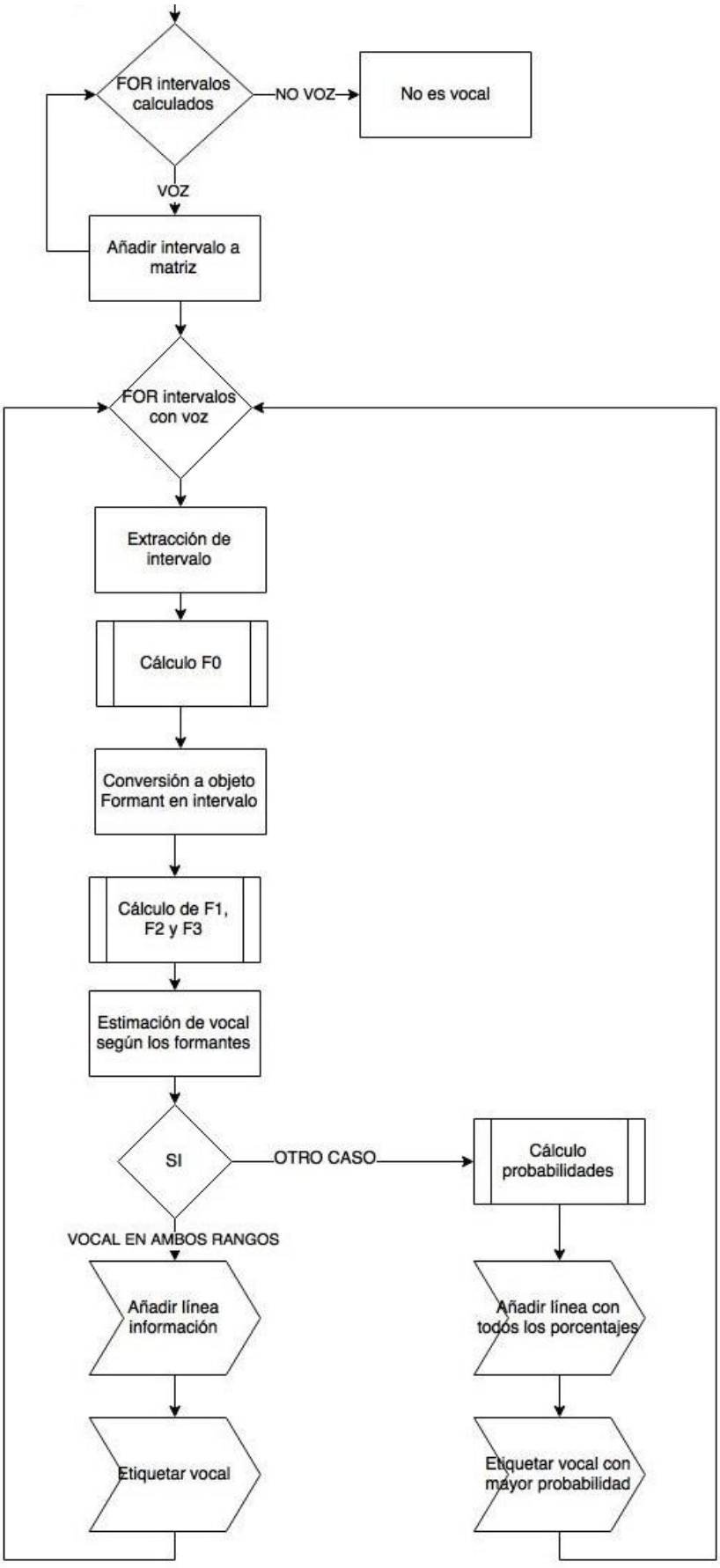


Figura 19: Diagrama de flujo

Como se dijo en el apartado anterior, nada más lanzar el script se muestra el formulario para elegir los datos. Una vez hecho esto, se trasladan los mismos a las variables del script y se selecciona la frecuencia máxima para el cálculo de formantes en función de si el hablante es hombre o mujer. A continuación, otro formulario mostrará los rangos para la estimación de la vocal detectada en función de sus formantes, con diferentes valores en función de la elección de género del hablante, permitiendo a su ver modificarlos. Si se está grabando, mostrará en pantalla el tiempo en segundos seleccionado para grabar a la vez que se inicia la grabación.

Una vez se ha hecho esto, se crea el objeto de sonido a partir del archivo de audio, o se inicia la grabación y se crea el objeto de sonido a partir de la misma. Las funciones utilizadas en ambos casos son las siguientes:

- Crear objeto a partir de archivo - Read from file: a esta función se le proporcionan el directorio, el nombre de archivo y la extensión que se quiere agregar como un objeto a Praat, y si el directorio y nombre existen y el tipo de archivo es compatible con Praat, se genera un objeto correspondiente.
- Grabar sonido – Record Sound (fixed time) y Write to file: La primera función graba un sonido desde el dispositivo de audio elegido y lo convierte en un objeto de sonido Praat. La segunda funciona de manera opuesta a la explicada en el apartado anterior, guarda el objeto seleccionado como un archivo con la extensión seleccionada. Tiene los siguientes argumentos:
    - Entrada: dispositivo de entrada, a elegir entre micrófono, línea o digital. En este caso, micrófono.
    - Ganancia, valor que debe estar entre 0 y 1, en este caso se ha elegido 0.99.
    - Equilibrio, valor que indica la distribución de sonido estéreo entre izquierda y derecha, variando entre 0 y 1. En este caso se ha elegido 0.5, de modo que será igual en ambos lados.
    - Frecuencia de muestreo: puede tomar los valores de 11025, 22050 y 44100 Hz. En este caso vale 44100 Hz.
    - Duración: número de segundos que se grabará, elegido en el formulario por el usuario.

Tras crear el objeto de sonido, se realiza el pre-procesado del mismo. Adicionalmente, se crea el objeto TextGrid para etiquetar las vocales más adelante, mediante la función Sound: To TextGrid que crea el objeto con el mismo eje de tiempo que el sonido. Tiene los siguientes argumentos:

- Nombre de los niveles: nombre que tendrá cada uno de los niveles que se mostrarán en la ventana del editor, y donde se podrán colocar puntos o intervalos. En este script solo se crea un nivel llamado *vowels*.
- Niveles de puntos: nombres de los niveles que serán de puntos, si no, por defecto son de intervalos. Aquí el nivel es de puntos.

La detección de vocales se realiza mediante la detección de picos de intensidad en el sonido, y la comprobación de que esos picos son áreas con voz. Para ello se requiere la creación de dos objetos Intensity y Pitch mediante las siguientes funciones:

- Sound: To Intensity: función que crea el objeto de intensidad a partir del de sonido. El algoritmo hace que los valores del sonido sean primero cuadrados, y posteriormente convolucionados con una ventana de análisis Gaussiana (Kaiser-20; con los lóbulos laterales por debajo de -190 dB). La duración efectiva de la ventana de análisis es 3.2 / tono mínimo, lo que garantiza que una señal periódica es analizada teniendo una sincronización intensidad-tono con un rizado no mayor a 0.00001 dB. Tiene los siguientes argumentos:
    - Tono mínimo (en Hz): la mínima frecuencia periódica en la señal. Lo ideal es seleccionarla lo más alta que permita la señal, de forma que el contorno de la intensidad sea más escarpado, facilitando la detección de picos. El tono mínimo suele estar en torno a los 40 Hz para la voz humana.
    - Paso de tiempo, en este caso se ha elegido de 0.01 segundos.
    - Eliminación de presión de sonido media: muchos micrófonos añaden un nivel constante a la presión de aire. Generalmente puede ser ignorado, pero si se elige la opción de eliminarla, el algoritmo sustraerá la presión media antes de aplicar la ventana Gaussiana. En este script se ignorará.
- Sound: To Pitch: función que crea el objeto de tono a partir del de sonido. El algoritmo hace una detección acústica de periodicidad en base a un método de auto-correlación exacto descrito por Boersma[35]. Para estimar la función de auto-correlación a corto plazo de una señal sobre la base de una señal con ventana, se divide la función de auto-correlación de la señal con ventana por la función de auto-correlación de la ventana $r_x(\tau) \approx r_{xw}(\tau) / r_w(\tau)$. Los ajustes que controlan la adición de candidatos son:
    - Paso de tiempo, del mismo valor que en la intensidad, 0.01 segundos.
    - Suelo de tono: no se añadirán candidatos por debajo de esta frecuencia. Este valor se calcula en el pre-procesado.
    - Techo de tono: no se añadirán candidatos por encima de esta frecuencia. Este valor se calcula en el pre-procesado.

  El resto de valores están determinados con defecto, aunque sí podrían modificarse usando la función Sound: To Pitch (ac):

    - Muy exacto: por defecto deshabilitado, si se habilita dobla el tamaño de la ventana Gaussiana.
    - Umbral de silencio: por defecto 0.03, los intervalos que no tengan amplitudes sobre este umbral son silencio.
    - Umbral de voz: valor estándar de 0.45, la fuerza de un candidato sin voz, relativa a la máxima correlación posible.
    - Coste por octava: valor estándar de 0.01 por octava, grado de preferencia por candidatos de alta frecuencia.
    - Coste por salto de octava: valor estándar de 0.35, grado de preferencia por que no existan cambios de tono.

- Coste con/sin voz: valor estándar de 0.14, es el grado de preferencia por detectar transiciones de voz/no voz.

Cuando los intervalos pertenecientes a las vocales hayan sido localizados, se entra en un bucle que realiza exactamente las mismas tareas para cada uno. En primer lugar, se extrae la parte del sonido correspondiente al intervalo. Esto se lleva a cabo con la función Extract part con los siguientes argumentos:

- Tiempo de inicio y tiempo de fin, que coincidirán con los del intervalo, calculado con el procedimiento que se describirá en el siguiente capítulo.
- Forma de la ventana: rectangular.
- Amplitud relativa entre el sonido original y el extraído. Un valor de 1 no modifica el sonido.
- Preservar tiempos entre sonido original y extraído, con valor 0 no se preservan.

Una vez calculada la frecuencia fundamental, se calcula el objeto Formant únicamente de la parte extraída. El sonido se re-muestrea con una frecuencia del dos veces el valor del formante máximo, y luego se aplica un pre-énfasis con un valor $\alpha = \exp(-2 \pi F \Delta t)$ donde Δt es el periodo de muestreo. Para cada ventana de análisis, Praat aplica una ventana de tipo Gaussiano, que computa los coeficientes LPC con el algoritmo Burg. El número de polos que computa este algoritmo es el doble del número máximo de formantes, siendo esta la razón por la que se puede seleccionar un valor de número máximo de formantes que sea múltiplo de 0.5. El algoritmo inicialmente encuentra el número máximo de formantes en el rango entre 0 y el formante máximo. Para ayudar a encontrar mejor F1 y F2, no se tienen en cuenta los valores en los rangos de 0 a 50 Hz, o de 50 Hz por debajo del formante máximo al propio formante máximo. Los atributos de esta función son los siguientes:

- Paso de tiempo, en este caso se ha seleccionado 0, por lo que el algoritmo lo computará como un 25% de la ventana de análisis.
- Número máximo de formantes: para la mayoría de análisis de habla, se extraen 5 formantes por paso.
- Formante máximo: es el techo en el rango de búsqueda, en hercios. Este valor se basa en la selección del principio, y será de 5000 Hz para hombres y 5500 Hz para mujeres.
- Longitud de la ventana: duración efectiva de la ventana de análisis. El valor que se usa en realidad es el doble del seleccionado. Con el valor seleccionado de 0.0025, por tanto, la ventana durará 0.005 segundos.
- Pre-énfasis: el punto +3 dB para un filtro paso bajo invertido con una pendiente de +6 dB por octava. Si este valor es de 50 Hz, entonces los valores por debajo de 50 Hz no se amplifican, los valores cerca de 100 Hz se amplifican 6 dB, los valores cerca de 200 Hz se amplifican 12 dB, y así sucesivamente. Esto se debe a que los espectros de las vocales suelen caer a 6 db por octava, y de este modo, se crea un espectro más plano de forma que se mejora el análisis de formantes, ya que se busca que los formantes coincidan con máximos locales, no con la pendiente espectral global.

El proceso de cálculo de formantes y frecuencia fundamental es sencillo: se utiliza la función Get value at time en el punto medio del intervalo de la vocal. En el caso de F0, sobre el objeto Pitch, y en caso de los formantes, en el objeto de formantes, teniendo además en este caso que indicar para qué formante (1,2,3…).

Tras calcular los formantes y detectar las vocales (proceso detallado en el siguiente capítulo), queda mostrar en pantalla la información recabada. Para la ventana de información, se utiliza la función printline, que funciona de manera similar a la función printf en C. Para insertar el punto en el TextGrid, se selecciona el mismo y se utiliza la función Insert Point, eligiendo el nivel vowel para insertar el punto, y el texto con el que se quiere etiquetar (esto es, la vocal de mayor porcentaje).

Por último, se guardan los archivos para poder replicar los resultados. Para el TextGrid se utiliza la función Write to text file, que guarda el archivo como texto, en el directorio especificado y con el mismo nombre que el sonido analizado y la extensión .vowels.TextGrid. Para la información en texto, se utiliza la función appendFile, que agrega la información seleccionada a un archivo de texto: se selecciona un archivo de texto con el mismo nombre que el sonido y extensión .txt, y la información será todo el texto de la ventana de información.

4.3.2. Procesos

4.3.2.1. Pre-procesamiento

El pre-procesado se lleva a cabo apoyándose en otros scripts con tareas bien diferenciadas, la función de cada uno de ellos se detalla a continuación:

- Workpre: calcula el punto de inicio y la duración total del sonido mediante las funciones Get start time y Get total Duration. Con estos datos, extrae la parte del sonido añadiendo un margen de 0.025 segundos. De esta forma, se elimina la parte de la grabación que no contiene información sonora, haciendo el análisis más efectivo. Por úlitmo, llama al script Fixdc.
- Fixdc: en primer lugar, resta la media del sonido, y una vez hecho esto crea un nuevo objeto de sonido mediante el filtrado paso alto de Hann desde 60 Hz. A continuación, llama al script declip y por último escala los tiempos del objeto recortado a los del original.
- Declip: busca el máximo absoluto de presión de sonido. En caso de que no pueda hallarlo, toma el valor 0. Si es mayor o igual a 1, se escala el sonido a 0.99 ya que Praat por defecto recorta lo que esté por encima de 1, resultando en distorsión.
- Minmaxf0: calcula la frecuencia fundamental, mediante la conversión a objeto pitch y la comprobación de los intervalos con voz. Si encuentra uno o más intervalos con voz, extrae los valores de F0 en el cuartil 25 y 75%, siendo estos

respectivamente el mínimo y máximo F0. En caso de no hallarlo, los valores por defecto son 40 y 600 Hz.

Para finalizar, ya en el script principal, se lee la frecuencia de muestreo y se re-muestrea a 11025 Hz si es menor que esta, y se filtra con un solo formante a 1000 Hz con un ancho de banda de 500 Hz.

4.3.2.2. Detección de intervalos correspondientes a vocales

Una vez han sido creado los objetos de intensidad y tono, se busca su máximo absoluto mediante la función Get maximum utilizando una interpolación cúbica, de forma que se obtiene el valor del máximo y el tiempo en el que se produce. Se convierte el objeto Intensity a un objeto matriz para obtener los parámetros, y posteriormente se crea una nueva matriz con los parámetros de dominio, número de columnas, distancia entre columnas, valor de x asociado a la primera columna, dominio en y, número de filas, distancia entre filas y valor de y asociado a la primera fila.

Por último, se crean dos vectores de tipo PointProcess a partir de los datos de la matriz: uno de ellos contendrá los intervalos correspondientes a los máximos y el otro los máximos en sí. Según el número de picos calculados, un bucle for va pico por pico comprobando si tiene voz o no. Si la tiene, y está por encima del umbral de intensidad se guardan los datos de tiempo de inicio y tiempo de final del máximo, esto es, de la vocal, y se incrementa la variable de cuenta en 1.

De esta forma, al terminar se tienen una serie de variables que almacenan los puntos de tiempo de inicio y fin de las vocales, y una variable que almacena el número total de vocales encontradas.

4.3.2.3. Detección de vocales y cálculo de probabilidades

Para la detección de las vocales una vez calculados sus formantes, se han utilizado los rangos de F1 y F2 calculados en el Anexo 3. Existen diferentes datos de rangos para hombres y mujeres, por lo que la asignación de los valores de estos rangos depende de la elección de género que se hizo en el formulario. Adicionalmente se permite la elección de los datos mediante el formulario de inicio del script.

Al comparar los datos calculados de F1 y F2 con los rangos de formantes, pueden darse dos casos diferenciados:

- Ambos formantes están dentro del rango para cualquier vocal: en este caso puede afirmarse que se ha detectado la vocal con toda seguridad. La vocal detectada se etiqueta en el TextGrid y en la ventana de información se muestra la vocal y entre paréntesis un porcentaje del 100%.
- Ambos formantes no están dentro de algún rango, o uno lo está, pero el otro no: se calculan las probabilidades que tiene de ser cada vocal.

El cálculo de las probabilidades se realiza de la siguiente manera:

$$\%\ de\ vocal\ F1 = \left(1 - \frac{Distancia\ al\ extremo\ del\ rango\ de\ F1\ más\ cercano}{Máxima\ distancia\ al\ extremo\ del\ rango\ F1\ más\ cercano}\right) * 50$$

$$\%\ de\ vocal\ F2 = \left(1 - \frac{Distancia\ al\ extremo\ del\ rango\ de\ F2\ más\ cercano}{Máxima\ distancia\ al\ extremo\ del\ rango\ F2\ más\ cercano}\right) * 50$$

De esta manera, el porcentaje será mayor cuanto más cerca esté el formante de alguno de los intervalos. Si uno de los dos formantes está dentro de uno de los rangos, automáticamente el porcentaje para esa vocal en ese rango será del 50%. El porcentaje total se calcula como:

$$\%\ de\ vocal = (\%\ de\ vocal\ F1) + (\%\ de\ vocal\ F2)$$

Por último, se comparan los porcentajes obtenidos para cada una de las vocales, y se etiqueta la vocal correspondiente al mayor en el TextGrid. En la ventana de información se añaden a la línea los porcentajes para cada una de las 5 vocales.

5. Resultados y evaluación

Las tres siguientes figuras muestran las ventanas que abre el script al ser lanzado. La figura 20 muestra el formulario de inicio, y las figuras 21 y 22 son un ejemplo de cómo quedan las ventanas de información y del editor que se abren automáticamente al terminar de ejecutarse el script.

Figura 20: Formulario de inicio

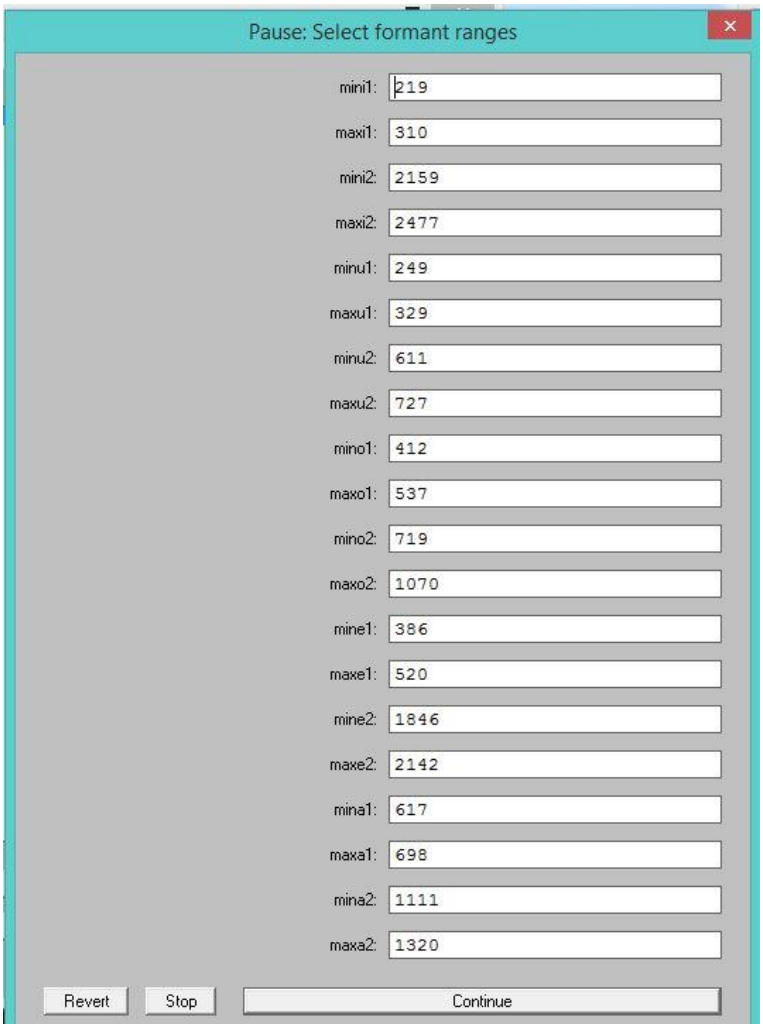


Figura 21: Formulario de rangos de formantes

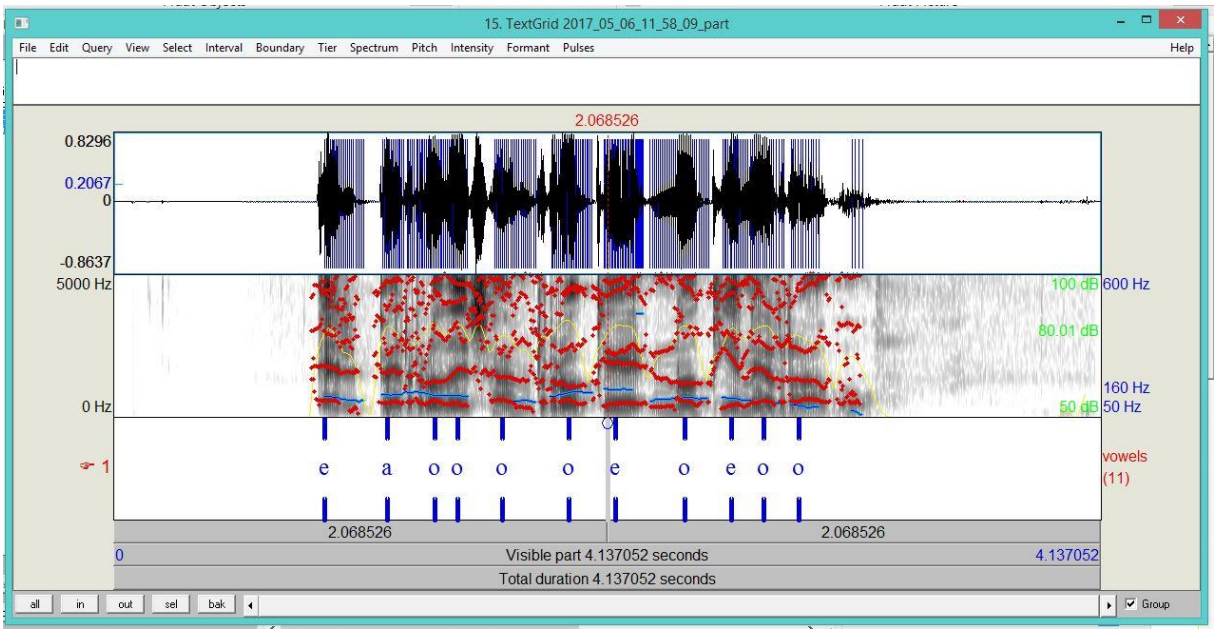


Figura 22: Resultado, ventana del editor

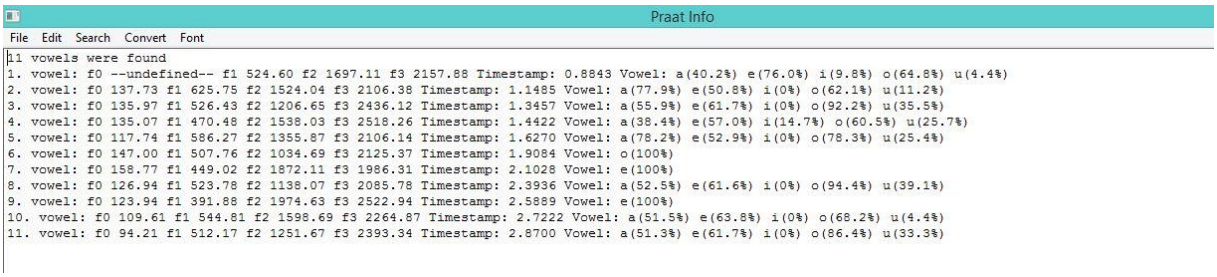


Figura 23: Resultado, ventana de información

Para medir la exactitud de los datos obtenidos, se han analizado cinco grabaciones de voz diferentes, todas del mismo hablante, masculino y nativo español: una de ellas contiene solo las vocales, tres contienen palabras con las mismas consonantes y variando las vocales, y la última de ellas un pasaje de El Quijote.

- Primera grabación: sólo vocales. Contiene cinco vocales en orden a, e, i, o y u.
- Segunda grabación: vocales con consonantes oclusivas nasales. Contiene las siguientes cinco palabras mama, mema, mina, momia y mundo.
- Tercera grabación: vocales con consonantes oclusivas orales sonoras. Contiene las siguientes cinco palabras: baba, beba, viva, boba y buba.
- Cuarta grabación: vocales con consonantes oclusivas orales sordas. Contiene las siguientes cinco palabras: papa, Pepa, pipa, popa, pupa.
- Quinta grabación: extracto de El Quijote. Contiene la siguiente cita: “La libertad, Sancho, es uno de los más preciosos dones que a los hombres dieron los cielos; con ella no pueden igualarse los tesoros que encierra la tierra ni el mar encubre”[36].

Los resultados obtenidos al analizar estas grabaciones se muestran en la siguiente tabla. Los resultados detallados pueden encontrarse en el Anexo 4.

Número total de vocales	97
Número total de vocales encontradas	81
Número total de vocales etiquetadas correctamente	56
Falsos positivos	12
Porcentaje de éxito en detección	74,31%
Porcentaje de éxito en etiquetado	60,22%

Tabla 3: Resultados de detección de vocales

Analizando los resultados de la detección:

- Los falsos positivos se deben en su totalidad a vocales con una pronunciación extendida en el tiempo, lo que causa que se detecten varias veces, o a la aparición de consonantes oclusivas nasales, que en ocasiones presentan un pequeño pico de intensidad que se reconoce como una vocal.

- Los fallos de detección se deben principalmente a dos factores:
    - Vocal cerrando una frase: en el final de una frase la intensidad de la voz baja, por lo que no detecta el pico de intensidad y, por tanto, la vocal.
    - Vocal en un diptongo o triptongo: en este caso sí que se detecta vocal, pero únicamente con un pico de intensidad, por lo que si existen dos o tres vocales no son detectadas.

En cuanto al etiquetado, al usarse datos medios de formantes es difícil establecer una mejor correlación. Para ello sería necesario ajustar los rangos para cada hablante en particular, ya que varían ligeramente en función del tono de voz y el acento; y las consonantes que acompañan a la vocal también producen que las frecuencias de los formantes aumenten o disminuyan.

6. Presupuesto y planificación

En este apartado se realiza un desglose de las tareas realizadas a lo largo de este Trabajo de Fin de Grado, lo que facilitará finalmente un cálculo aproximado de su coste.

6.1. Planificación del trabajo

Se ha dividido el trabajo en varias fases diferenciadas entre sí:

Fase 1: Documentación inicial

I. Familiarización con la plataforma de desarrollo Praat (10 horas).
II. Estudio de conceptos técnicos asociados (20 horas).

Fase 2: Desarrollo de la detección de vocales y el cálculo de formantes

I. Estudio de soluciones de detección ya implementadas (10 horas).
II. Implementación de la solución de detección (10 horas).
III. Estudio de soluciones de cálculo ya implementadas (10 horas).
IV. Implementación de la solución de cálculo (10 horas).

Fase 3: Desarrollo de la presentación de resultados

I. Diseño del método de etiquetado en TextGrid e información por pantalla (20 horas).
II. Diseño del método de estimación de vocales (5 horas).

Fase 4: Integración

I. Implementación de la introducción de datos por parte del usuario (5 horas).
II. Integración de los elementos anteriores (5 horas).

Fase 5: Elaboración de la memoria

I. Redacción de la memoria (55 horas).
II. Corrección y maquetación (10 horas).

Fases	Horas empleadas
Documentación inicial	30
Desarrollo de detección de vocales y cálculo de formantes	40
Desarrollo de la presentación de resultados	25
Integración	10
Elaboración de la memoria	65
Total	170

Tabla 4: Desglose de tareas

6.2. Presupuesto

6.2.1. Costes materiales

Los materiales necesarios han sido un ordenador donde realizar el desarrollo de la aplicación. Praat no requiere de mucha capacidad de computación, por lo que un procesador de 1000MHz, 512 Mb de RAM y 8 Gb de disco duro son suficientes. Adicionalmente, debe contar con micrófono integrado para poder grabar voz, por lo que se ha elegido un ordenador portátil, que integra micrófono. La licencia de Praat es gratuita y puede descargarse directamente desde su página web.

Concepto	Precio (€)
Ordenador	279,90
Total	279,90

Tabla 5: Costes materiales

6.2.2. Costes de personal

Para la realización de este trabajo, ha sido necesaria la presencia de un jefe de proyecto y un ingeniero.

Ocupación	Precio por hora(€)	Horas	Importe (€)
Jefe de proyecto	50	30	1500
Ingeniero	30	170	5100
Total	**280**	**200**	**6600**

Tabla 6: Costes de personal

6.2.3. Costes de totales

Concepto	Precio (€)
Costes materiales	279,90
Costes de personal	6600
Costes indirectos (15%)	1031,98
Subtotal	**7911,88**
IVA (21%)	1661,52
Total	**9573,4**

Tabla 7: Costes totales

El coste total del proyecto es de nueve mil quinientos setenta y tres euros con cuarenta céntimos.

7. Conclusiones y líneas futuras

7.1. Conclusiones

En este trabajo se ha creado un script que permite la detección de vocales en una secuencia de grabación de voz, así como el etiquetado de las mismas en la visualización de audio. El diseño final permite detectar las vocales tanto desde un archivo de audio, como grabar directamente desde el propio script. Se trata de un script sencillo, y que puede ser usado por usuarios que no tengan un conocimiento profundo ni del software en concreto, ni conocimientos avanzados de fonética. Todo el desarrollo se ha llevado a cabo usando software libre.

La mayor dificultad encontrada en la realización de este trabajo ha sido familiarizarse con un lenguaje de programación desconocido, que se asemeja a otros, pero no se basa en ninguno, y que cuenta con multitud de funciones específicas. Afortunadamente, la documentación y la ayuda son extensas, y cubren casi todos los temas necesarios para implementar las soluciones.

Otras dificultades han surgido debido a particularidades a la hora de escribir el código del script. Concretamente que, si hay un espacio después de una línea de código, esto puede causar que arroje un error y el script al completo no funcione, y no es algo que sea visible; o que las variables del programa no puedan empezar por mayúsculas, y el error asociado no lo especifique como tal.

Por último, aunque el funcionamiento del script es bastante satisfactorio y cumple con los objetivos planteados, la correlación a la hora de diferenciar de qué vocal se trata muestra un porcentaje de éxito no demasiado alto.

En conjunto se puede concluir que el alumno partía de un conocimiento básico de procesado de audio y programación y tras el desarrollo de la solución estos conocimientos proporcionan al alumno las capacidades necesarias para poder desarrollar código en lenguajes de programación distintos a los estudiados, así como de comprender y trabajar mejor los conceptos relacionados con las ondas de audio en general, y la voz humana en particular.

7.2. Líneas futuras

Como ya se ha mencionado en el apartado anterior, uno de los problemas principales que presenta el script es la baja correlación a la hora de diferenciar de qué vocal se trata en el punto detectado como una vocal. Esto podría corregirse modificando los datos de los rangos que se usan para el cálculo. Para ello, sería necesaria una caracterización con una muestra de usuarios mayor que los 16 que se ha tomado aquí (ver anexo 3). La parte positiva es que este

script permitiría automatizar gran parte de esta tarea, ya que puede realizar la grabación y el análisis, y sólo sería necesario registrar y procesar los resultados. Además, como los rangos no son fijos si no que pueden ser modificados, un usuario que obtenga nuevos valores podría probar el script mejorado por sí mismo sin necesidad de tener conocimientos de programación.

Otro de los problemas que podría mejorarse es la detección de las dos vocales en un diptongo o triptongo, que en la situación actual es detectado como una única vocal. Para ello sería necesario el análisis y comparación de los formantes en dos o tres puntos del intervalo correspondiente a la vocal, cuando actualmente sólo se hace en el punto intermedio. Estos dos puntos podrían ser al 25% y el 75% del mencionado intervalo. Con los datos de estos tres formantes, podría comprobarse si las diferencias entre los valores indican que hay una, dos o tres vocales diferentes. Esta solución se intentó anteriormente, pero el script arrojaba errores a la hora de analizar los puntos del 25% y 75% del intervalo vocálico, lo que imposibilitaba finalmente la ejecución y fue por tanto descartado.

Inicialmente se planteó la intención de que el análisis se desarrollase en tiempo real, pero Praat no permite esta funcionalidad de momento, por lo que también esta opción quedó descartada. Sí que sería posible hacerlo utilizando programas como Matlab, o algunas librerías de C++ y Python, pero son utilidades que no permiten tanta precisión en análisis fonético como Praat, que está diseñado ad-hoc para esta función.

8. Bibliografía

1. O'Grady, William; “Contemporary Linguistics: An Introduction” (5th ed.); Bedford/St. Martin's.
2. T.V.F. Brogan; “English Versification”; Baltimore: Johns Hopkins University Press.
3. “Proceedings of the Royal Society of London”; Series B. Vol. 91, No. 641
4. Willis, Robert; “On the vowel sounds, and on reed organ pipes”; Cambridge: [publisher not identified].
5. Bickford, Anita; “Articulatory Phonetics: Tools For Analyzing The World's Languages” (4th ed.); Summer Institute of Linguistics.
6. Titze, I.R.; “Principles of Voice Production”; Prentice Hall.
7. Thurman, Leon & Welch, ed., Graham; “Body mind & voice: Foundations of voice education” (revised ed.); Collegeville, Minnesota: The Voice Care Network et al.
8. Boldrey, Richard; “Guide to Operatic Roles and Arias”; Caldwell Publishing Company.
9. Shewan, Robert; "Voice Classification: An Examination of Methodology"; The NATS Bulletin. 35: 17–27.
10. Rothenberg, M; “The Breath-Stream Dynamics of Simple-Released Plosive Production”, Vol. 6; Bibliotheca Phonetica, Karger, Basel.
11. Greene, Margaret; Lesley Mathieson (2001); “The Voice and its Disorders”; John Wiley & Sons; 6th Edition.
12. http://pages.jh.edu/~virtlab/ray/acoustic.htm
13. https://www.merriam-webster.com/dictionary/frequency
14. Hecht, Eugene; “Optics” (2nd ed.); Addison Wesley.
15. Knopp, Konrad; Bagemihl, Frederick; “Theory of Functions Parts I and II”; Dover Publications. p. 3.
16. http://hyperphysics.phy-astr.gsu.edu/hbase/permot3.html
17. http://hyperphysics.phy-astr.gsu.edu/hbase/Sound/intens.html
18. http://www.engineeringtoolbox.com/sound-pressure-d_711.html
19. Anssi Klapuri; "Introduction to Music Transcription"; New York: Springer.

20. Jones, S.; Longe, O.; Pato, M. V.; "Auditory evoked potentials to abrupt pitch and timbre change of complex tones: electrophysiological evidence of streaming?". Electroencephalography and Clinical Neurophysiology.

21. Handel, S.; “Timbre perception and auditory object identification”. Hearing, 425-461.

22. http://www.etymonline.com/index.php?term=vowel

23. IPA; “Handbook of the IPA”, p. 12.

24. De Cantero, F. J.; “Teoría y análisis de la entonación”; Barcelona: Universidad de Barcelona.

25. Jeans, J.H.; “Science & Music”; Dover, pp 104, 148.

26. Fant, G.; “Acoustic Theory of Speech Production”; Mouton & Co, The Hague, Netherlands.

27. Benade, A. H.; “Fundamentals of musical acoustics”; Oxford University Press, London.

28. Standards Secretariat, Acoustical Society of America; “American National Standard Acoustical Terminology”; Acoustical Society of America, Melville, NY.

29. Atal, B. S. and Hanauer, S. L. ; "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave"; J. Acoust. Soc. Am., 50, 637-655.

30. "The Venturi effect"; Wolfram Demonstrations Project.

31. Catford, J.C.; “A Practical Introduction to Phonetics”; Oxford University Press, p. 161.

32. Katsuhiko Ogata; “System Dynamics” (4th ed.); University of Minnesota. p. 617.

33. Boersma, Paul & Weenink, David; “Praat: doing phonetics by computer [Computer program]”; Version 6.0.29, retrieved 24 May 2017 from http://www.praat.org/

34. http://www.fon.hum.uva.nl/praat/manual/power_spectral_density.html

35. Paul Boersma (1993): "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound." Proceedings of the Institute of Phonetic Sciences 17: 97–110. University of Amsterdam.

36. Miguel de Cervantes; El Ingenioso Hidalgo Don Quijote de La Mancha; Parte II, Capítulo LVIII.

37. Quilis, A. y Esgueva, M. (1983). Realización de los fonemas vocálicos españoles en posición fonética normal. In M. Esgueva y M. Cantarero (Eds.), Estudios de fonética I. (pp. 137-252). Madrid: Consejo Superior de Investigaciones Científicas.

Anexo 1: Código de los scripts

Detect_vowels.praat

```
clearinfo
form Counting vowels in Sound Utterances
	choice type: 1
		button Record
		button Use file
	sentence directory C:\Users\Desktop
	sentence fileName Recording
	choice fileExtension: 1
		button WAV
		button AIFF
		button AIFC
		button FLAC
		button NIST
		button MP3
	sentence recordTime 5
	choice gender: 1
		button Male
		button Female
endform
if type=1
	printline Please speak for 'recordTime$' seconds
	Record Sound (fixed time)... Microphone 0.99 0.5 44100 'recordTime$'
	clearinfo
	if fileExtension = 1
		Write to WAV file... 'directory$'/'fileName$'.wav
	elsif fileExtension=2
		Write to AIFF file... 'directory$'/'fileName$'.aiff
	elsif fileExtension=3
		Write to AIFC file... 'directory$'/'fileName$'.aifc
	elsif fileExtension=4
		Write to FLAC file... 'directory$'/'fileName$'.flac
	elsif fileExtension=5
		Write to NIST file... 'directory$'/'fileName$'.nist
	else
		printline MP3 not supported, saving as WAV
		Write to WAV file... 'directory$'/'fileName$'.wav
	endif
else
	Read from file... 'directory$'/'fileName$'.'fileExtension$'
endif

if gender = 1
	value=5000
	beginPause: "Select formant ranges"
		real: "mini1", 219
		real: "maxi1", 310
		real: "mini2", 2159
```

```
                real: "maxi2", 2477
                real: "minu1", 249
                real: "maxu1", 329
                real: "minu2", 611
                real: "maxu2", 727
                real: "mino1", 412
                real: "maxo1", 537
                real: "mino2", 719
                real: "maxo2", 1070
                real: "mine1", 386
                real: "maxe1", 520
                real: "mine2", 1846
                real: "maxe2", 2142
                real: "mina1", 617
                real: "maxa1", 698
                real: "mina2", 1111
                real: "maxa2", 1320
        endPause: "Continue", 1
else
        value=5500
        beginPause: "Select formant ranges"
                real: "mini1", 210
                real: "maxi1", 270
                real: "mini2", 2586
                real: "maxi2", 3083
                real: "minu1", 206
                real: "maxu1", 280
                real: "minu2", 546
                real: "maxu2", 711
                real: "mino1", 440
                real: "maxo1", 581
                real: "mino2", 832
                real: "maxo2", 1130
                real: "mine1", 445
                real: "maxe1", 538
                real: "mine2", 2085
                real: "maxe2", 2421
                real: "mina1", 598
                real: "maxa1", 729
                real: "mina2", 1054
                real: "maxa2", 1282
        endPause: "Continue", 1
endif

s$ = selected$("Sound")
s = selected("Sound")
execute workpre.praat
wrk = selected("Sound")
To TextGrid... "vowels" vowels
textgridid = selected("TextGrid")
select wrk
sr = Get sample rate
```

```
include minmaxf0.praat

pitch = To Pitch... 0.01 minF0 maxF0

threshold = 21

select wrk

if sr > 11025
	downsampled = Resample... 11025 1
else
	downsampled = Copy... tmp
endif
Filter with one formant (in-line)...  1000 500
framelength = 0.01
int_tmp = To Intensity... 40 'framelength' 0
maxint = Get maximum... 0 0 Cubic
t1 = Get time from frame... 1
matrix_tmp = Down to Matrix
endtime = Get highest x
ncol = Get number of columns
coldist = Get column distance
h=1
newt1 = 't1'+('h'*'framelength')
ncol = 'ncol'-(2*'h')
matrix_intdot = Create Matrix... intdot 0 'endtime' 'ncol' 'coldist' 'newt1' 1 1 1 1 1
(Object_'matrix_tmp'[1,col+'h'+'h']-Object_'matrix_tmp'[1,col]) / (2*'h'*dx)
temp_IntDot = To Sound (slice)... 1
temp_rises = To PointProcess (extrema)... Left yes no Sinc70
select temp_IntDot
temp_peaks = To PointProcess (zeroes)... Left no yes
npeaks = Get number of points
select downsampled
plus matrix_tmp
plus matrix_intdot
plus temp_IntDot
Remove


cnt = 1


for pindex from 1 to 'npeaks'
	select temp_peaks
	ptime = Get time from index... 'pindex'
	select int_tmp
	pint = Get value at time... 'ptime' Nearest

	select pitch
	voiced = Get value at time... 'ptime' Hertz Nearest
	if pint > (maxint-threshold) and voiced <> undefined
		select temp_rises
```

```
                rindex = Get low index... 'ptime'
                if rindex >= 1
                        rtime = Get time from index... 'rindex'
                        otime = ('rtime'+'ptime')/2
                        otime_'cnt' = otime
                        otime2 = otime + 0.05
                        otime2_'cnt' = otime2
                        cnt += 1
                endif
        endif
endfor

select int_tmp
plus temp_rises
plus temp_peaks
plus pitch
Remove

cnt=cnt-1
printline 'cnt' vowels were found

for ifile from 1 to cnt

        otime = otime_'ifile'
        otime2 = otime2_'ifile'
   vowel = ((otime+otime2)/2)
        select wrk

        ids'ifile' = Extract part... 'otime' 'otime2' Rectangular 1 yes
   select ids'ifile'

   pitchid1 = noprogress To Pitch... 0 75 600
   vowelf0 = Get value at time... 'vowel' Hertz Linear
        #barkFreq = 13*arctan(0.00076*vowelf0) + 3.5*(arctan((vowelf0/7500)^2))
        barkFreq = vowelf0

        select ids'ifile'
   formantid = noprogress To Formant (burg)... 0 6 'value' 0.025 50
   #this sets the formant Units
   formantUnit$ = "Hertz"
   #get f1, f2, f3 values
        vowelf1 = Get value at time... 1 'vowel' 'formantUnit$' Linear
        vowelf2 = Get value at time... 2 'vowel' 'formantUnit$' Linear
        vowelf3 = Get value at time... 3 'vowel' 'formantUnit$' Linear

        #printline 'ifile'. vowel: f0 'barkFreq:2' f1 'vowelf1:2' f2 'vowelf2:2' f3 'vowelf3:2' Timestamp:
'vowel:4'
        #select textgridid
        #Insert point... 1 otime 'ifile'
        #Insert point... 1 otime2

        if (vowelf1>mini1 && vowelf1<maxi1 && vowelf2>mini2 && vowelf2<maxi2)
```

```
        case=1
elsif (vowelf1>minu1 && vowelf1<maxu1 && vowelf2>minu2 && vowelf2<maxu2)
        case=2
elsif (vowelf1>mino1 && vowelf1<maxo1 && vowelf2>mino2 && vowelf2<maxo2)
        case=3
elsif (vowelf1>mine1 && vowelf1<maxe1 && vowelf2>mine2 && vowelf2<maxe2)
        case=4
elsif (vowelf1>mina1 && vowelf1<maxa1 && vowelf2>mina2 && vowelf2<maxa2)
        case=5
else
        case=6
        disti1=abs(min(('vowelf1'-'mini1'),('vowelf1'-'maxi1')))
        distu1=abs(min(('vowelf1'-'minu1'),('vowelf1'-'maxu1')))
        disto1=abs(min(('vowelf1'-'mino1'),('vowelf1'-'maxo1')))
        diste1=abs(min(('vowelf1'-'mine1'),('vowelf1'-'maxe1')))
        dista1=abs(min(('vowelf1'-'mina1'),('vowelf1'-'maxa1')))
        disti2=abs(min(('vowelf2'-'mini2'),('vowelf2'-'maxi2')))
        distu2=abs(min(('vowelf2'-'minu2'),('vowelf2'-'maxu2')))
        disto2=abs(min(('vowelf2'-'mino2'),('vowelf2'-'maxo2')))
        diste2=abs(min(('vowelf2'-'mine2'),('vowelf2'-'maxe2')))
        dista2=abs(min(('vowelf2'-'mina2'),('vowelf2'-'maxa2')))
        maxdis1=max(disti1,distu1,disto1,diste1,dista1)
        maxdis2=max(disti2,distu2,disto2,diste2,dista2)
        pctgi1=(1-('disti1'/'maxdis1'))/2
        pctgu1=(1-('distu1'/'maxdis1'))/2
        pctgo1=(1-('disto1'/'maxdis1'))/2
        pctge1=(1-('diste1'/'maxdis1'))/2
        pctga1=(1-('dista1'/'maxdis1'))/2
        pctgi2=(1-('disti2'/'maxdis2'))/2
        pctgu2=(1-('distu2'/'maxdis2'))/2
        pctgo2=(1-('disto2'/'maxdis2'))/2
        pctge2=(1-('diste2'/'maxdis2'))/2
        pctga2=(1-('dista2'/'maxdis2'))/2
        if (vowelf1>mini1 && vowelf1<maxi1 && vowelf2<mini2 && vowelf2>maxi2)
                pctgi1=0.5
        elsif (vowelf1<mini1 && vowelf1>maxi1 && vowelf2>mini2 && vowelf2<maxi2)
                pctgi2=0.5
        elsif (vowelf1>minu1 && vowelf1<maxu1 && vowelf2<minu2 && vowelf2>maxu2)
                pctgu1=0.5
        elsif (vowelf1<minu1 && vowelf1>maxu1 && vowelf2>minu2 && vowelf2<maxu2)
                pctgu2=0.5
        elsif (vowelf1>mino1 && vowelf1<maxo1 && vowelf2<mino2 && vowelf2>maxo2)
                pctgo1=0.5
        elsif (vowelf1<mino1 && vowelf1>maxo1 && vowelf2>mino2 && vowelf2<maxo2)
                pctgo2=0.5
        elsif (vowelf1>mine1 && vowelf1<maxe1 && vowelf2<mine2 && vowelf2>maxe2)
                pctge1=0.5
        elsif (vowelf1<mine1 && vowelf1>maxe1 && vowelf2>mine2 && vowelf2<maxe2)
                pctge2=0.5
        elsif (vowelf1>mina1 && vowelf1<maxa1 && vowelf2<mina2 && vowelf2>maxa2)
                pctga1=0.5
        elsif (vowelf1<mina1 && vowelf1>maxa1 && vowelf2>mina2 && vowelf2<maxa2)
```

```
                pctga2=0.5
            endif
            pctgi=('pctgi1'+'pctgi2')*100
            pctgu=('pctgu1'+'pctgu2')*100
            pctgo=('pctgo1'+'pctgo2')*100
            pctge=('pctge1'+'pctge2')*100
            pctga=('pctga1'+'pctga2')*100
            pctg=max(pctgi,pctgu,pctgo,pctge,pctga)
            if pctg=pctgi
                detected=1
            elsif pctg=pctgu
                detected=2
            elsif pctg=pctgo
                detected=3
            elsif pctg=pctge
                detected=4
            elsif pctg=pctga
                detected=5
            endif
        endif

        if case=1
            printline 'ifile'. vowel: f0 'barkFreq:2' f1 'vowelf1:2' f2 'vowelf2:2' f3 'vowelf3:2'
Timestamp: 'vowel:4' Vowel: i(100%)
            select 'textgridid'
            Insert point... 1 vowel i
        elsif case=2
            printline 'ifile'. vowel: f0 'barkFreq:2' f1 'vowelf1:2' f2 'vowelf2:2' f3 'vowelf3:2'
Timestamp: 'vowel:4' Vowel: u(100%)
            select 'textgridid'
            Insert point... 1 vowel u
        elsif case=3
            printline 'ifile'. vowel: f0 'barkFreq:2' f1 'vowelf1:2' f2 'vowelf2:2' f3 'vowelf3:2'
Timestamp: 'vowel:4' Vowel: o(100%)
            select 'textgridid'
            Insert point... 1 vowel o
        elsif case=4
            printline 'ifile'. vowel: f0 'barkFreq:2' f1 'vowelf1:2' f2 'vowelf2:2' f3 'vowelf3:2'
Timestamp: 'vowel:4' Vowel: e(100%)
            select 'textgridid'
            Insert point... 1 vowel e
        elsif case=5
            printline 'ifile'. vowel: f0 'barkFreq:2' f1 'vowelf1:2' f2 'vowelf2:2' f3 'vowelf3:2'
Timestamp: 'vowel:4' Vowel: a(100%)
            select 'textgridid'
            Insert point... 1 vowel a
        elsif case=6
            printline 'ifile'. vowel: f0 'barkFreq:2' f1 'vowelf1:2' f2 'vowelf2:2' f3 'vowelf3:2'
Timestamp: 'vowel:4' Vowel: a('pctga:1'%) e('pctge:1'%) i('pctgi:1'%) o('pctgo:1'%) u('pctgu:1'%)
            select 'textgridid'
            if detected=1
                Insert point... 1 vowel i
```

```
                elsif detected=2
                        Insert point... 1 vowel u
                elsif detected=3
                        Insert point... 1 vowel o
                elsif detected=4
                        Insert point... 1 vowel e
                elsif detected=5
                        Insert point... 1 vowel a
                endif
        endif

        select ids'ifile'
   plus 'pitchid1'
        plus 'formantid'
   Remove
endfor

select wrk
plus textgridid
Edit

Write to text file... 'directory$'/'fileName$'.vowels.TextGrid
appendFile: "'fileName$'.txt", info$ ( )
```

Workpre.praat

```
dur = Get total duration
stt = Get start time
Extract part... stt-0.025 stt+dur+0.025 rectangular 1 no
execute fixdc.praat
```

Declip.praat

```
procedure action
        clip = Get absolute extremum... 0 0 None
        if clip = undefined
                clip = 0
        endif
        clip = 'clip:4'
        if clip >= 1
                Scale... 0.9999
        endif
endproc
```

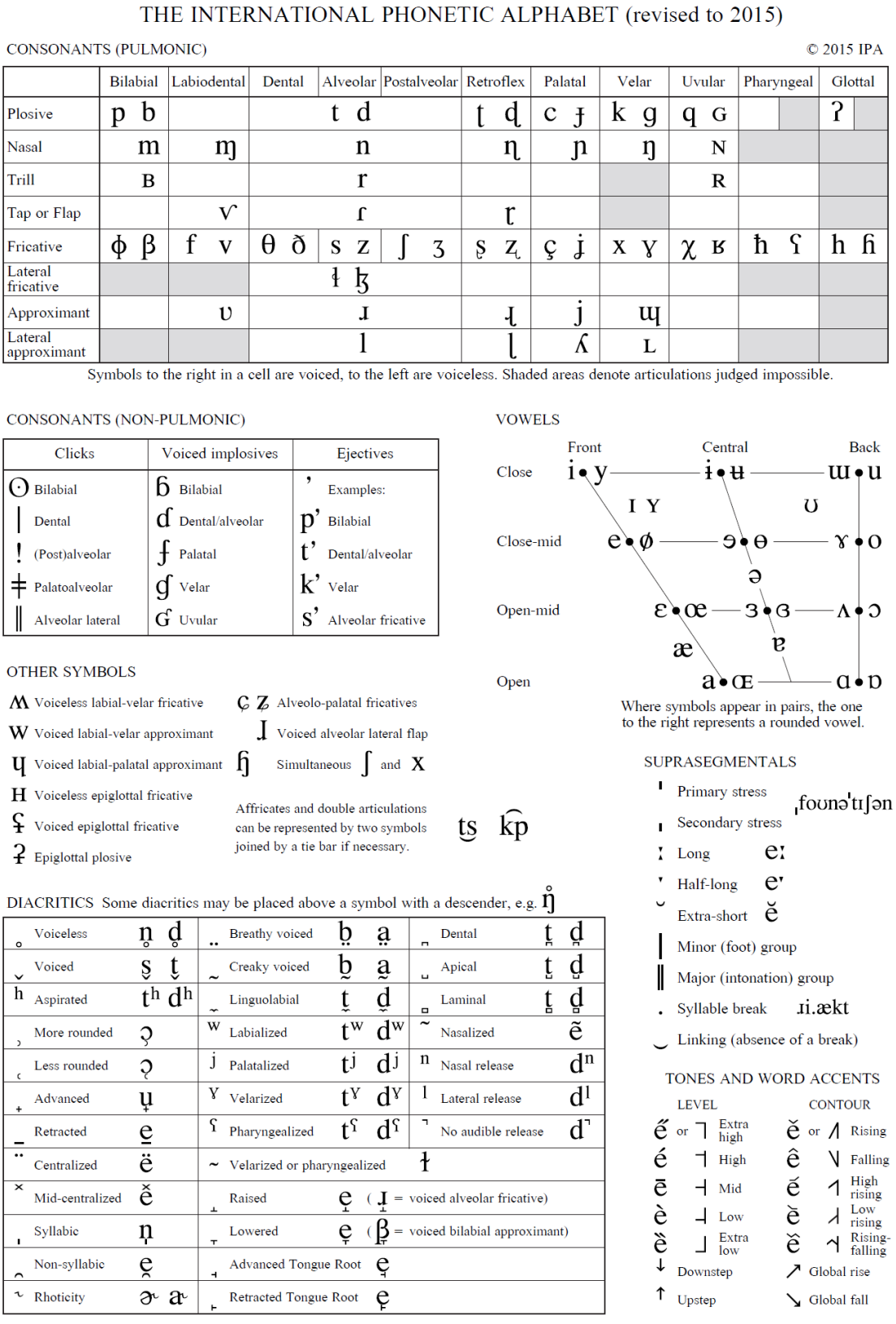
Minmaxf0.praat

```
selsnd_m = selected("Sound")
nocheck To Pitch... 0 40 600
fullName$ = selected$()
type$ = extractWord$(fullName$, "")
if type$ = "Pitch"
	voicedframes = Count voiced frames
	if voicedframes > 0
		q25 = Get quantile... 0 0 0.25 Hertz
		q75 = Get quantile... 0 0 0.75 Hertz
		minF0 = round(q25 * 0.75)
		maxF0 = round(q75 * 1.5)
	else
		minF0 = 40
		maxF0 = 600
	endif
	Remove
else
	minF0 = 40
	maxF0 = 600
endif
select selsnd_m
```

Fixdc.praat

```
procedure action
	s = selected("Sound")
	Subtract mean
	tmp = Filter (pass Hann band)... 60 0 20
	execute declip.praat
	select s
	Formula... Object_'tmp'[]
	select tmp
	Remove
	select s
	stt = Get start time
	if stt <> 0
		dur = Get total duration
		Scale times to... 0 dur
	endif
endproc
```

Anexo 2: Alfabeto Fonético Internacional



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	ɓ Bilabial	’ Examples:
ǀ Dental	ɗ Dental/alveolar	pʼ Bilabial
ǃ (Post)alveolar	ʄ Palatal	tʼ Dental/alveolar
ǂ Palatoalveolar	ɠ Velar	kʼ Velar
ǁ Alveolar lateral	ʛ Uvular	sʼ Alveolar fricative

OTHER SYMBOLS

ʍ Voiceless labial-velar fricative
ɕ ʑ Alveolo-palatal fricatives
w Voiced labial-velar approximant
ɺ Voiced alveolar lateral flap
ɥ Voiced labial-palatal approximant
ʢ Simultaneous ʃ and x
ʜ Voiceless epiglottal fricative
ʡ Voiced epiglottal fricative
ʔ Epiglottal plosive

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.

t͡s k͡p

VOWELS

Where symbols appear in pairs, the one to the right represents a rounded vowel.

SUPRASEGMENTALS

ˈ Primary stress ˌfoʊnəˈtɪʃən
ˌ Secondary stress
ː Long eː
ˑ Half-long eˑ
˘ Extra-short ĕ
| Minor (foot) group
‖ Major (intonation) group
. Syllable break ɹi.ækt
‿ Linking (absence of a break)

TONES AND WORD ACCENTS

LEVEL		CONTOUR	
e̋ or ˥	Extra high	ě or ˩˥	Rising
é ˦	High	ê ˥˩	Falling
ē ˧	Mid	e᷌ ˦˥	High rising
è ˨	Low	e᷄ ˨˩	Low rising
ȅ ˩	Extra low	e᷅ ˩˥˧	Rising-falling
↓	Downstep	↗	Global rise
↑	Upstep	↘	Global fall

DIACRITICS Some diacritics may be placed above a symbol with a descender, e.g. n̥ŋ̊

◌̥ Voiceless	n̥ d̥	◌̤ Breathy voiced	b̤ a̤	◌̪ Dental	t̪ d̪
◌̬ Voiced	s̬ t̬	◌̰ Creaky voiced	b̰ a̰	◌̺ Apical	t̺ d̺
◌ʰ Aspirated	tʰ dʰ	◌̼ Linguolabial	t̼ d̼	◌̻ Laminal	t̻ d̻
◌̹ More rounded	ɔ̹	◌ʷ Labialized	tʷ dʷ	◌̃ Nasalized	ẽ
◌̜ Less rounded	ɔ̜	◌ʲ Palatalized	tʲ dʲ	◌ⁿ Nasal release	dⁿ
◌̟ Advanced	u̟	◌ˠ Velarized	tˠ dˠ	◌ˡ Lateral release	dˡ
◌̠ Retracted	e̠	◌ˤ Pharyngealized	tˤ dˤ	◌̚ No audible release	d̚
◌̈ Centralized	ë	~ Velarized or pharyngealized	ɫ		
◌̽ Mid-centralized	e̽	◌̝ Raised	e̝ (ɪ̝ = voiced alveolar fricative)		
◌̩ Syllabic	n̩	◌̞ Lowered	e̞ (β̞ = voiced bilabial approximant)		
◌̯ Non-syllabic	e̯	◌̘ Advanced Tongue Root	e̘		
◌˞ Rhoticity	ɚ a˞	◌̙ Retracted Tongue Root	e̙		

Figura 24: IPA

Anexo 3: Caracterización de las vocales en español según sus formantes

El idioma español tiene 5 sonidos vocálicos bien diferenciados entre ellos: /a/, /e/, /i/, /o/ y /u/. Aunque en algunas variantes el número de fonemas es mayor, en este proyecto sólo se analizan los anteriormente mencionados. En las posteriores tablas se muestran los datos de media y variación típica de los formantes F1 y F2 en las 5 vocales del español para 16 hombres y 16 mujeres nativos[37].

	F1	σ	F2	σ
[í]	240,75	23,2	2841,7	248
[i]	240,58	29,9	2828,25	236,42
/i/	240,75	24	2834,9	241,23
[é]	501,75	41,3	2292	167,5
[e]	481,5	45,7	2214	111,3
/e/	491,6	38,5	2252,08	134,1
[á]	661,5	24,1	1156,4	107,58
[a]	665,9	64,8	1179,16	113,4
/a/	663,75	43,4	1167,8	106,1
[ó]	510,75	70	967,5	149
[o]	510,75	60	994,5	101,9
/o/	510,75	68,5	981	112,8
[ú]	249,75	36,9	630	82,1
[u]	236,25		627,75	
/u/	243	29,5	628,8	63,5

Tabla 8: Datos de formantes en mujeres

	F1	σ	F2	σ
[í]	268,28	45,3	2342,15	158,22
[i]	260,68	36,4	2294,09	154,6
/i/	264,5	37,8	2317,5	154,3
[é]	449,71	66,6	2052,7	147,6
[e]	454,96	63,9	1935,28	107
/e/	453,8	60,8	1995,01	113,2
[á]	665,68	39,8	1220,4	103,7
[a]	648,84	38,3	1211,59	81,8
/a/	657,28	38,4	1215	87,5
[ó]	475,8	58,4	900,46	175,4
[o]	473,3	61,6	895,18	95,7
/o/	474,5	52	888,4	103
[ú]	291,09	39,4	685,12	57,5
[u]	283,5	23,1	653,06	40,6
/u/	293,5	37,7	669,08	44,2

Tabla 9: Datos de formantes en hombres

Con los datos de estas tablas, se han calculado los datos de media de los formantes F1 y F2, y se ha tomado la desviación típica máxima para cada vocal.

	Media F1	σ	Media F2	σ
i	240,6933	29,9	2834,95	248
e	491,6167	45,7	2252,693	167,5
a	663,7167	64,8	1167,787	113,4
o	510,75	70	981	149
u	243	36,9	628,85	82,1

Tabla 10: Media y desviación estándar de formantes en mujeres

	Media F1	σ	Media F2	σ
i	264,4867	45,3	2317,913	158,22
e	452,8233	66,6	1994,33	147,6
a	657,2667	39,8	1215,663	103,7
o	474,5333	61,6	894,68	175,4
u	289,3633	39,4	669,0867	57,5

Tabla 11: Media y desviación estándar de formantes en hombres

Con estos datos se han calculado los rangos de F1 y F2 para hombres y mujeres.

	F1 MIN	F1 MAX	F2 MIN	F2 MAX
Rango i	210,7933	270,5933	2586,95	3082,95
Rango e	445,9167	537,3167	2085,193	2420,193
Rango a	598,9167	728,5167	1054,387	1281,187
Rango o	440,75	580,75	832	1130
Rango u	206,1	279,9	546,75	710,95

Tabla 12: Rango de F1 y F2 en mujeres

	F1 MIN	F1 MAX	F2 MIN	F2 MAX
Rango i	219,1867	309,7867	2159,693	2476,133
Rango e	386,2233	519,4233	1846,73	2141,93
Rango a	617,4667	697,0667	1111,963	1319,363
Rango o	412,9333	536,1333	719,28	1070,08
Rango u	249,9633	328,7633	611,5867	726,5867

Tabla 13: Rango de F1 y F2 en hombres

Las siguientes figuras muestran los rangos de F1 y F2 para las diferentes vocales en hombres y mujeres, mostrando que los rangos no se superponen en ningún momento. Los rangos no son esféricos, se ha tomado la mayor longitud de rango en cada caso para la representación:

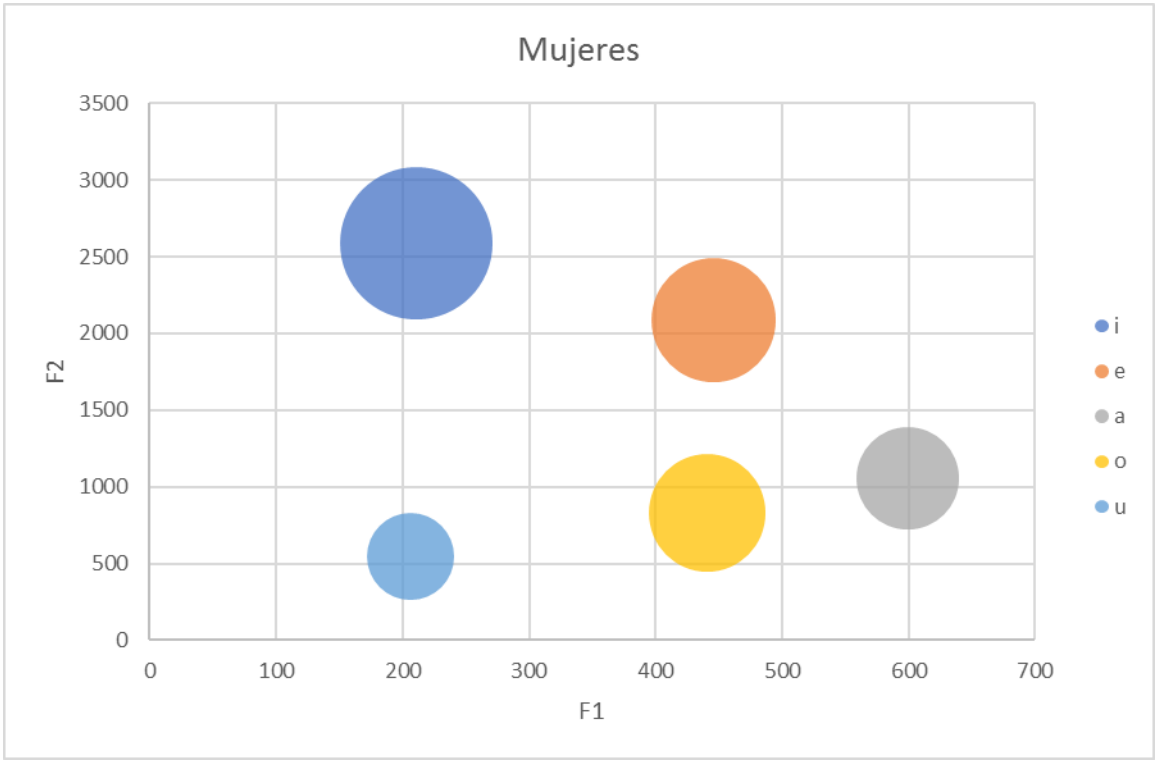


Figura 25: Rangos de F1 y F2 en mujeres

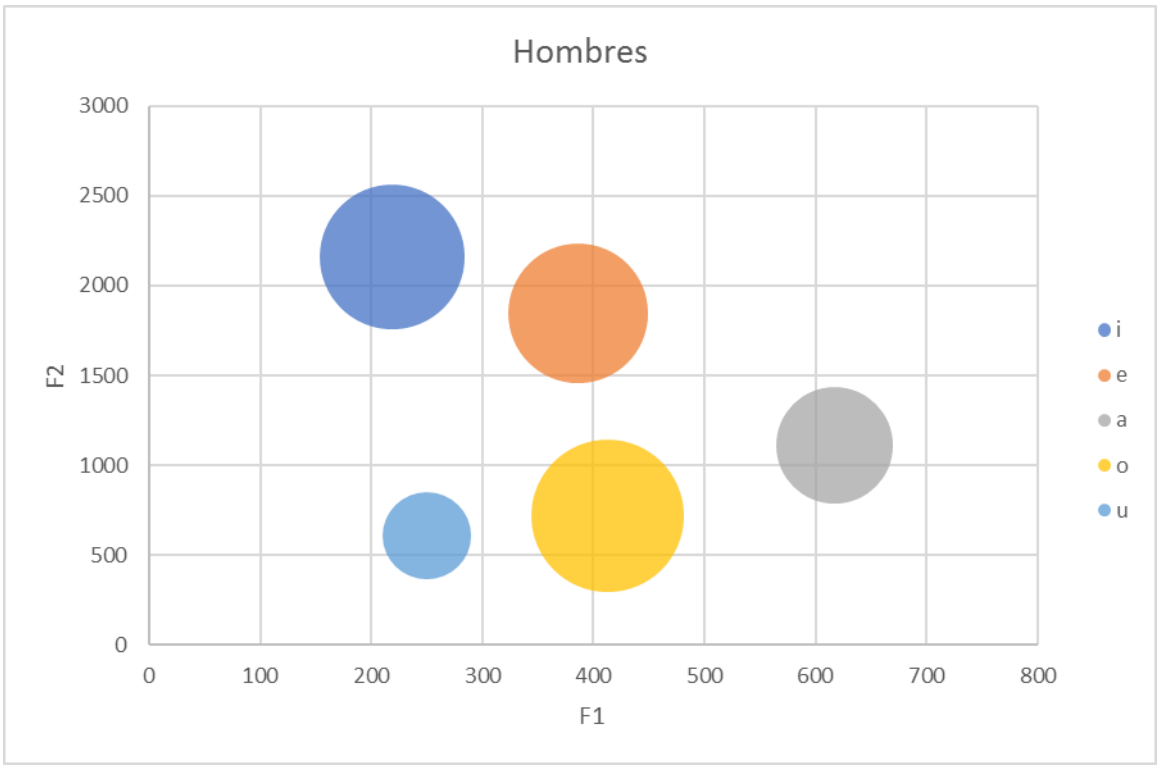


Figura 26: Rangos de F1 y F2 en hombres

Anexo 4: Tablas de resultados

Vocal	Detectada	Etiquetada correctamente
a	Sí	Sí
e	Sí	Sí
i	Sí	Sí
o	Sí	No
u	Sí	Sí
Falsos positivos	3	

Tabla 14: Resultados vocales

Vocal	Detectada	Etiquetada correctamente
a	Sí	Sí
a	Sí	Sí
e	Sí	Sí
a	Sí	Sí
i	Sí	No
a	Sí	Sí
o	Sí	Sí
i	No	No
a	Sí	Sí
u	Sí	Sí
o	Sí	Sí
Falsos positivos		1

Tabla 15: Resultados vocales con consonantes oclusivas nasales

Vocal	Detectada	Etiquetada correctamente
a	Sí	Sí
a	Sí	Sí
e	Sí	Sí
a	Sí	Sí
i	Sí	Sí
a	Sí	Sí
o	Sí	No
a	Sí	Sí
u	Sí	Sí
a	No	No
Falsos positivos		0

Tabla 16: Resultados vocales con consonantes oclusivas orales sordas

Vocal	Detectada	Etiquetada correctamente
a	Sí	Sí
a	Sí	Sí
e	Sí	Sí
a	Sí	Sí
i	Sí	No
a	Sí	Sí
o	Sí	No
a	Sí	No
u	Sí	Sí
a	Sí	No
Falsos positivos		6

Tabla 17: Resultados vocales con consonantes oclusivas orales sonoras

Vocal	Detectada	Etiquetada correctamente
a	Sí	Sí
i	Sí	Sí
e	Sí	No
a	Sí	Sí
a	Sí	Sí
o	Sí	Sí
e	Sí	Sí
u	Sí	No
o	Sí	Sí
e	Sí	No
o	Sí	Sí
a	Sí	No
e	Sí	No
i	No	No
o	Sí	No
o	Sí	Sí
o	Sí	Sí
e	Sí	Si
e	No	No
a	No	No
o	Sí	Sí
o	Sí	Sí
e	Sí	No
i	No	No
e	Sí	Sí
o	Sí	Sí
o	Sí	Sí
i	Sí	Sí

e	Sí	No
o	No	No
o	Sí	Sí
e	Sí	Sí
a	Sí	Sí
o	Sí	Sí
u	Sí	No
e	No	No
e	No	No
i	Sí	No
u	No	No
a	Sí	No
a	Sí	No
e	No	No
o	Sí	Sí
e	No	No
o	Sí	Sí
o	Sí	Sí
e	Sí	Sí
e	No	No
i	No	No
e	Sí	Sí
a	Sí	No
a	Sí	Sí
i	No	No
e	Sí	No
a	Sí	No
i	No	No
e	Sí	No
a	Sí	Sí
e	Sí	Sí
u	Sí	Sí
e	Sí	No
Falsos positivos		2

Tabla 18: Resultados vocales en cita de El Quijote