Contents

List of Figures

List of Tables

Part I

Introducción

Capítulo 1

Motivacion y planteamiento del problema

Las deficiencias auditivas o fonadoras afectan a millones de personas en todo el mundo. Se estima que aproximadamente el 8% de la población nacional presenta algún nivel de deficiencia auditiva. Muchos de los casos se dan a edades muy tempranas debido a anomalías genéticas, enfermedades durante el embarazo como la rubeola o algunas infecciones no controladas como la meningitis. Esta población es más sensible a perder las capacidades de comunicación oral por lo que en los países más desarrollados se han implementado diversos programas con el fin de facilitar la integración social de los mismos.

A lo largo de años se han desarrollado numerosas técnicas y estudios de intervención para facilitar la integración de los mismos derivando en dos grandes vertientes:

- Gestualismo:es la vertiente más utilizada en la actualidad, considerada como el lenguaje materno de las personas con deficiencias auditivas severas. A lo largo de los años se han mejorado las doctrinas para la enseñanza del mismo hasta llegar al nivel de que las pautas evolutivas del individuo son similares a las de uno sano. • Consiste esencialmente en un sistema de símbolos producidos por gestos corporales que mediante combinaciones transfieren información. Es un sistema no universal e icónico con diferentes versiones.

Figure 1.1: Abecedario en el lenguaje de signos Español

- Oralismo: • es la vertiente que busca la adquisición del lenguaje oral como lengua materna. Ha derivado en distintas técnicas como:
  - Sistema verbotonal: busca el aprovechamiento y optimización de los restos auditivos mediante sistemas tecnológicos avanzados que actúan como seleccionadores y amplificadores.
  - Lectura labial: consiste en el reconocimiento del habla mediante la captación visual del movimiento y posición de los órganos que intervienen en la fonación.

Como el lenguaje de signos o gesticular no es conocido por toda la población es lógico pensar que las capacidades comunicativas de estas personas serán limitadas creándose cierta tendencia a la exclusión social. Numerosos estudios

defiende el oralismo como principal solución para la integración de este segmento poblacional. Los defensores del oralismo pretenden plantear alternativas al lenguaje de signos. Los más optimistas buscan que estas personas no solo hablen sino que también se comuniquen satisfaciendo sus necesidades en todas las áreas de su desarrollo.

Mediante ciclos empíricos han demostrado que en la mayoría de los casos con una planificación y entrenamiento adecuado se logra una asimilación lingüística legible, aunque requiere de un periodo extenso ya que aún se sigue intentando buscar metodologías adecuadas para optimizar el proceso. La naturaleza del problema radica en la dificultad de la asimilación de conocimientos sobre fonemas complejos. A falta de percepción sonora se buscan otras vías de captación de conocimientos como la sensorial o la visual.

Dado que nuestro sistema visual es más sensible y capta una mayor información siempre será la mejor opción para este problema. La solución estaría en la planificación de un programa que facilite la absorción de conocimiento de señales bioacústicas de una forma visual y sencilla. En esencia, se busca sustituir el sistema auditivo periférico, encargado del proceso de captación durante la audición, por el sistema visual para de esta forma permitir una alternativa. De ser así, se conseguiría mejorar y agilizar el proceso de aprendizaje lingüístico y mejorar las capacidades comunicativas de todas estas personas.

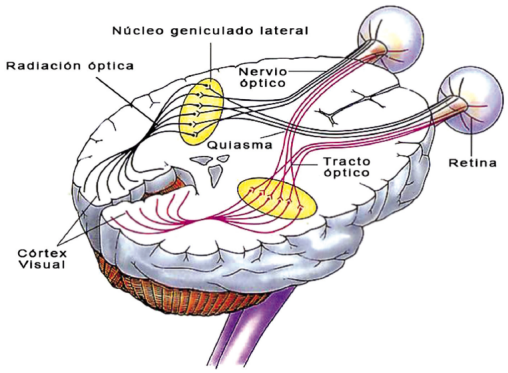

Figure 1.2: Sistema Visual Humano - Esquema genérico

El problema radica en que el sonido no transfiere información visual alguna por lo que esta debe de ser generada artificialmente. La evolución tecnológica ha permitido el estudio de las características de la señal de voz y a lo largo de los años se han desarrollado numerosas interfaces para la producción y síntesis del habla. Este proyecto nace para cubrir esta carencia. Se trata de desarrollar una herramienta que agilice el proceso de aprendizaje lingüístico mediante la representación visual de los principales parámetros característicos de la señal de voz.

Posteriormente se realizarán pruebas experimentales para garantizar el correcto funcionamiento de la misma así como posibles mejoras tanto en el rendimiento como la usabilidad. El fin último, es hacer de esta herramienta un medio cotidiano en los tratamientos de la logopedía así como en el día a día de estas personas.

1.1 Antecedentes

La tecnología ha evolucionado haciéndose cada vez más accesible, por lo que paulatinamente han ido mejorando las investigaciones sobre el procesamiento de la señal de voz. Estas investigaciones han dado lugar a una gran cantidad de herramientas diseñadas para la investigación y estudio de la señal de voz. A día de hoy, cualquier ordenador ya sea de sobremesa o portátil puede correr este tipo de aplicaciones por lo que la posibilidad de extraer información de señales acústica esta ahí presente. El inconveniente principal radica en que estas herramientas son complejas siendo desarrolladas más bien para ámbitos científicos de investigación. La necesidad de herramientas para doctrinas fonéticas ha ido creciendo a lo largo de los años pero todas ellas comparten una naturaleza compleja que requiere de conocimientos previos sobre procesado de señal de voz. Algunas de las destacadas son:

- Dr. Speech: es un software comercial diseñado para la extracción y evaluación de información de una señal de voz. Es una herramienta portable con uso muy extendido en campos de estudio de la voz, esencialmente a nivel profesional por logopedas en todo el mundo. Permite el análisis de señales en tiempo real o mediante carga de archivos con costes computacionales reducidos.

- DADiSP: es una herramienta comercial similar a Dr. Speech pero con unas características más avanzadas. Permite la extracción de la mayoría de los parámetros de una señal sonora pero requiere de una configuración exhaustiva y unos conocimientos sobre procesado de señal mínimos. No permite el análisis de señales en tiempo real. Su principal inconveniente es que sus costes de licencia que rondan los 2000$, un precio excesivo para un particular.

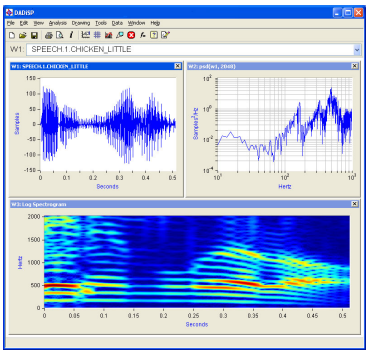

Figure 1.3: GUI principal del software comercial, DADiSP

- WaveSurfer: es una herramienta de código abierto diseñada para visualización y manipulación del sonido. Presenta una interfaz simple e intuitiva que se puede adaptar a diferentes tareas. Es una aplicación que calcula los principales parámetros de la señal de voz pero no presenta un funcionamiento adecuado para análisis en tiempo real.

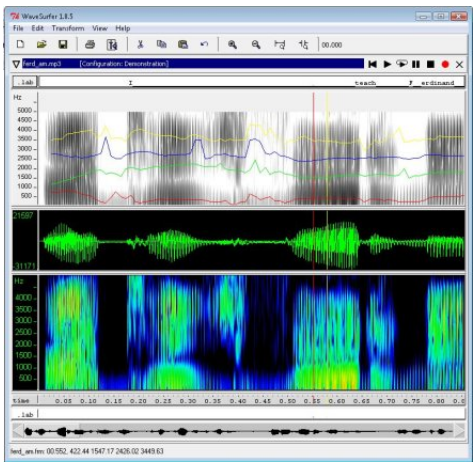

Figure 1.4: GUI principal del software libre, WaveSurfer

- Praat: es una herramienta de software libre multiplataforma diseñada para analizar, sintetizar y manipular la voz humana. Es la herramienta más utilizada para realizar análisis acústicos y anotar corpus orales.

  Permite la extracción de los parámetros más importantes de la señal de voz pero requiere de unos conocimientos mínimos sobre procesado de señal de voz. Por otro lado no permite un análisis adecuado de señales en tiempo real.

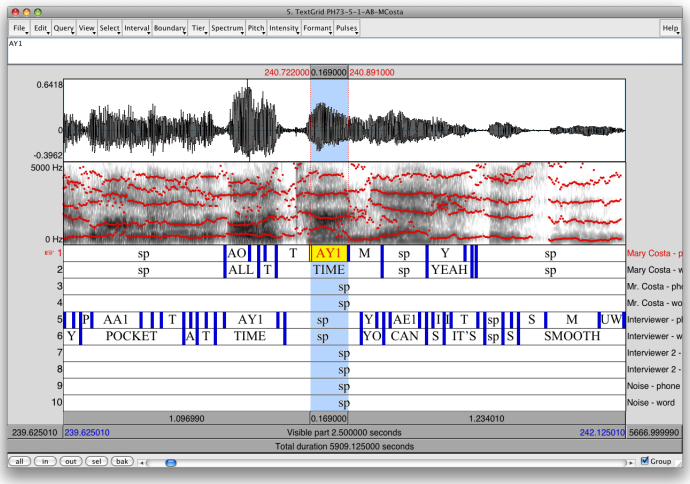

Figure 1.5: GUI principal del software libre, Praat

Todas herramientas presentan sus ventajas e inconvenientes pero ninguna de ellas fue diseñada para facilitar los procedimientos de aprendizaje para la fonación y audición correcta en personas con deficiencias auditivas. Aunque todas ellas comparten la misma naturaleza (presentar la información visual de señales acús-

ticas de forma visual) presentan una serie de complejidades que dificultan su uso para este fin. Algunas de las más destacadas:

- La complejidad de uso ya en la mayoría de ellas se requieren conocimientos sobre procesado de señal que la mayoría de la población desconoce.
- Los elevados costes de licencia.
- La información extraída en la mayoría de las herramientas es muy especializada y requiere de unos conocimientos previos para su entendimiento.
- La falta de integridad de otros idiomas: es un factor fundamental dado que la mayoría de las herramientas están construidas en inglés.
- No todas ellas permiten un análisis en tiempo real, hecho imprescindible para un aprendizaje óptimo.
- Interfaces poco intuitivas y difíciles de configurar.
- La mayoría de ellas no son multiplataforma, solo corren en sistemas Windows. Cada vez se apuesta más por el uso de sistemas operativos de uso libre en entornos educativos por lo que estas herramientas no se pueden utilizar a priori.

Viendo todas estas ocurrencias incrementa aún más la necesidad de desarrollar una aplicación que cumpla esas carencias y que optimice la visualización de resultados. Se han de definir una serie de objetivos previos para planificar el presente proyecto de forma adecuada.

1.2 Objetivos y especificaciones

El objetivo principal de este proyecto se basa en la necesidad planteada en las secciones anteriores. Se busca agilizar las capacidades de aprendizaje lingüístico de un determinado sector de la población con el diseño e implementación de una herramienta. La herramienta debe de permitir la extracción de información de señales acústicas y la presentación de la misma de una forma visual adecuada y simple. Como pautas principales para la implementación deben de cumplir una serie de especificaciones básicas:

- *Multiplataforma:* es necesario que esta esté soportada por diferentes sistemas operativos, centrándonos en Windows, Mac y Guadalinex (como principal distribución de Linux).
- *Multilenguaje:* la herramienta tiene que ser configurable para ejecutarse en diferentes idiomas. Además tiene que permitir una integración sencilla para inclusión de nuevos idiomas.
- *Software Libre:* la herramienta debe de distribuirse bajo licencia GNU. Los usuarios serán libres de descargarla y distribuirla.
- *Uso en tiempo real:* la herramienta debe de permitir el cálculo de los parámetros seleccionados en tiempo real.

- *Recursos Reducidos:* la herramienta debe de requerir unos recursos limitados para poder ser ejecutada en la mayoría de los ordenadores convencionales.

- *Uso sencillo*: el usuario debe de ser capaz de ejecutar y configurar la herramienta sin tener conocimientos previos sobre procesado de señal.

- *Configuración automática:* la herramienta debe de permitir una configuración automática adaptando sus componentes a las necesidades del entorno: detectores de actividad sonora, eliminación de ruido...

- *Personalización:* el usuario debe de poder personalizar la herramienta: colores de las gráficas, datos a mostrar, escala de colores en los espectros...

- *Exportación de resultados:* la herramienta debe de permitir exportar los resultados obtenidos a los formatos más convenientes: doc, pdf, png...

Cumplidos estos requerimientos se pondrá a prueba la aplicación en una serie de pruebas piloto y serán los propios *beta-testers* los que sugieran posibles mejoras en la herramienta.

1.3 Planificación del proyecto

La parte más engorrosa del proyecto está constituida por un problema de ingeniería de software, el desarrollo de la herramienta en sí. El proceso para el desarrollo de la misma se regirá por las pautas planteadas en el modelo de ciclo de vida del desarrollo de software en espiral. Este ciclo se engloba en tres secciones principales bien diferenciadas:

- Fase de planificación: detección y análisis de requisitos además de la elección de herramientas de desarrollo. En esta fase se barajaran cada una de las opciones disponibles en el mercado y se elegirá la más óptima para el proyecto.

- Fase de implementación: desarrollo y testeo del software. Se construirán cada uno de los bloques que conformarán el programa de forma independiente para posteriormente ser testeados y optimizados.

- Fase de despliegue y mantenimiento: liberación de la herramienta y soporte. Una vez terminada la programación y esta sea testeada se dará pie a una serie de pruebas piloto para garantizar el correcto funcionamiento de la misma.

Una vez desarrollada la herramienta se planificara un programa específico para exprimir todas sus características y de esta forma poder realizar una serie de pruebas de evaluación experimentales. La idea es que la aplicación sea testea por diferentes usuarios. Se instará a un determinado grupo poblacional conformado por hombre y mujeres de diferentes edades a realizar un programa experimental que ponga a prueba sus aptitudes para la adquisición de información acústica de forma visual en ausencia de percepción auditiva.

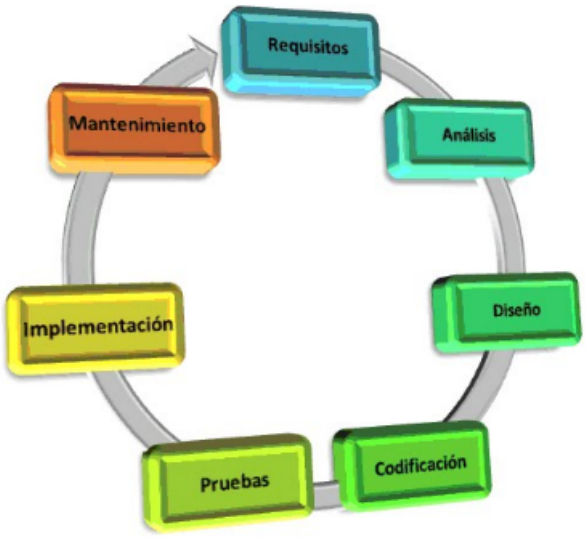

Figure 1.6: Modelo en espiral del cíclo de vida del desarrollo de software

El programa se estructura en diferentes niveles de complejidad ascendentes con el fin de buscar un compromiso entre las horas requeridas para el aprendizaje frente a las capacidades asimiladas. De esta forma mediante la confirmación de una serie de resultados poder extraer conclusiones y poder garantizar la viabilidad del proyecto.

Capítulo 2

Audición

El habla es el medio de comunicación más útil y complejo del ser humano. El hombre es capaz de emitir señales constituidas por ondas de presión a través del aparato fonador y captar estas últimas mediante del sistema auditivo. Esto implica que ambos aparatos tengan un papel fundamental en cualquier sistema de comunicación oral, por lo que conocer sus funcionamientos e implicaciones son de vital importancia para un desarrollo óptimo del presente proyecto.

La audición es un aspecto fundamental en el proceso de comunicación lingüística que afecta tanto a la percepción como la producción del sonido. Está constituida por el conjunto de procesos psico-fisiológicos que proporcionan al ser humano la capacidad de tener percepción auditiva. La audición es llevada a cabo en dos etapas:

1. Etapa fisiológica: en esta etapa el sujeto realiza el proceso de captación de sonido transformando las ondas de presión sonora en impulsos eléctricos. La captación es realizada por el subconjunto de órganos que conforman el sistema auditivo periférico.

2. Etapa psicológica: una vez captada la señal se realiza un proceso de percepción. Este proceso es realizado por el subconjunto de órganos del sistema auditivo central.

Figure 2.1: Sistema auditivo humano

Las personas con discapacidades auditivas tienden a presentar problemas en la primera de las etapas debido a disfunciones del sistema auditivo periférico. Este último está dividido en tres partes:

- Oído externo: actúa como filtro y canalizador de los estímulos sonoros. Está compuesto fundamentalmente por:

  - El pabellón auricular: pantalla de captación de sonidos formada por tejidos cartilaginosos y un tejido conectivo conocido como lóbulo.
  - El conducto auditivo externo: formado por tejidos cartilaginosos y la porción interna ósea que llega hasta la membrana timpánica.
  - La membrana timpánica: formada por tres capas, actúa como frontera entre el oído externo e interno. Es la encargada de transformar las ondas sonoras en vibraciones mecánicas.

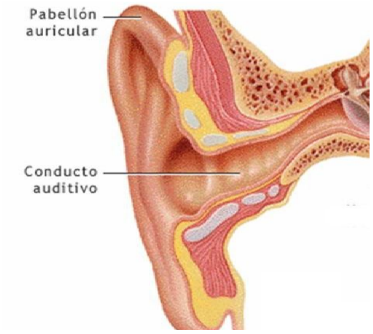

Figure 2.2: Oido externo - Esquema genérico

- • Oído medio: es el encargado de conducir y amplificar las vibraciones de la membrana timpánica. Está compuesto principalmente por dos cavidades:
  - Cavidades timpanomastoideas: están en contacto con la trompa de Eustaquio y ayudan a mantener el equilibrio de presiones entre el oído externo e interno.
  - Cadena de huesecillos: compuesta por el martillo, yunque y estribo. Configuran la palanca de segundo grado que incrementa los estímulos o los aísla como protección del oído interno.

Figure 2.3: Oido medio - Esquema genérico

- Oído interno: es el encargado de transformar los estímulos mecánicos en impulsos eléctricos mediante la estimulación de las células ciliadas. Está aislado en una capsula conocida como cápsula laberíntica que constituye el laberinto óseo. Este último está conformado por:
  - El caracol: en el mismo, la membrana basilar y de Reisner crean las escalas timpánicas y vestibular.
  - El vestíbulo: alberga el utrículo y el sáculo encargados de la percepción estática de la situación de la cabeza.

– Los conductos semicirculares: encargados del proceso de percepción de equilibrio dinámico.

El oído interno es el encargado de transcribir las frecuencias de vibración a impulsos nerviosos, distribuyéndose en una escala logarítmica conocida como escala del caracol coclear, figura2.5, un conjunto de filtros naturales que extraen las frecuencias fundamentales de los sonidos captados.

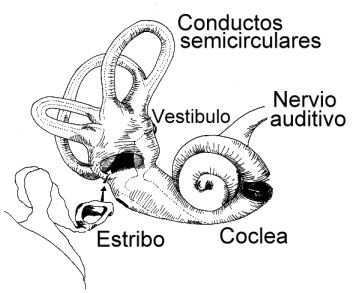

Figure 2.4: Oido interno - Esquema genérico

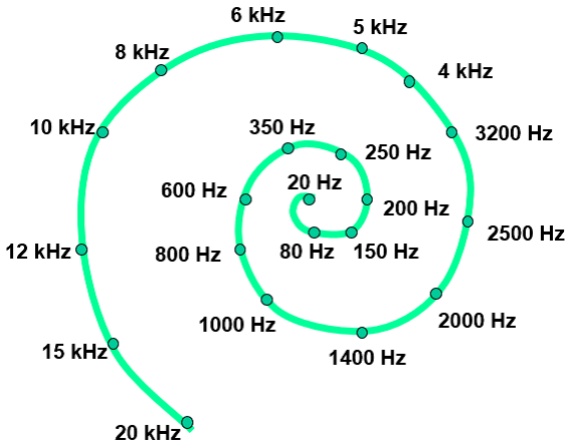

Figure 2.5: Caracol coclear: Distribuciones de frencuencias en la cóclea.

Para que se produzca una correcta audición cada una de estas partes deben de funcionar adecuadamente y de forma síncrona. Las ondas sonoras captadas conforman la energía mecánica transformada.

Esta energía será canalizada por el pabellón auricular hacia el conducto auditivo externo chocando con el tejido membranoso del tímpano que iniciará una vibración produciéndose el movimiento en cadena del conjunto de huesecillos. Posteriormente la energía es amplificada a través de dos mecanismos:

- La diferencia de superficie entre la membrana timpánica y oval.
- El efecto de amplificación de la palanca de segundo grado que constituye la cadena osicular.

Para que la amplificación se produzca de forma adecuada es necesario una integridad y unos niveles de presión atmosféricos equivalentes ambos lados. Cuando se da el caso, existen mecanismos que pueden aislar un nivel de percepción excesivo mediante la rigidez de la cadena.

El estímulo amplificado moviliza los líquidos del oído interno que estimulan las células ciliadas (albergadas en el órgano de Corti) transformando la energía mecánica en energía eléctrica. Esta última, posteriormente, es transmitida hasta la región temporooccipital de la corteza cerebral donde se generará la memoria auditiva por el sistema auditivo central.

La memoria generada permite la comunicación oral ya que este proceso de lenguaje está basado en la repetición de sonidos percibidos. La capacidad de aprendizaje es mayor cuanto menor es la mielinización de la vía neural, y a partir de los seis años gran parte de la misma está ya mielinizada. Esto implica que en los casos con problemas auditivos una rehabilitación a temprana edad favorezca el proceso de aprendizaje lingüístico.

2.1 Discapacidad auditiva y efectos en el aprendizaje lingüístico

Sociológicamente una persona presenta discapacidades auditivas cuando se le es diagnosticada una audición deficiente que afecta a ambos oídos (pérdidas auditivas bilaterales). La discapacidad auditiva deriva de pérdidas o anormalidades en la función anatómica y/o fisiológica del sistema auditivo. Las pérdidas pueden verse ocasionadas por distintos factores presentando cada una de ellas unos trastornos totalmente diferenciados. Existen distintos niveles de discapacidad auditiva, partiendo del umbral de nivel de audición medio de una población sana las podemos clasificar en estas variantes:

- Audición normal: presentan un nivel de audición de 0-20dB.
- Hipoacusia leve o ligera: presentan un nivel de audición de 20-40dB. Estos sujetos tienen problemas con la percepción de sonidos a largas distancias.
- Hipoacusia media o moderada: presentan un nivel de audición de 40-70dB. Por regla general estos sujetos presentan un retraso en el proceso de aprendizaje del lenguaje así como alteraciones articulatorias severas.
- Hipoacusia severa: presentan un nivel de audición de 70-90dB. Requieren de niveles elevados para tener percepción alguna. Estas personas comienzan a presentar un déficit comunicativo severo ya que perciben una audición deficiente.
- Hipoacusia profunda o sordera: presentan un nivel de audición de más de 90dB. Estas personas por regla general pierden sus capacidades comunicativas sin el tratamiento adecuado. Existe una audición residual ilegible.
- Cofosis o anacusia: es un problema poco frecuente que implica la pérdida total de la audición.

Los déficits derivados de estos problemas conforman la alteración sensorial más frecuente en la población general. Estos trastornos son sumamente importantes

ya que presentan implicaciones sustanciales en la evolución psico-social del individuo, esencialmente en el desarrollo lingüístico.

Es por ello que en el ámbito de la logopedia se han planteado numerosos programas de intervención que han presentado un conjunto relevante de evidencias empíricas sobre su efectividad. Los resultados de los mismos sugieren que las personas con algún trastorno auditivo severo tienden a presentar problemas acentuados en su desarrollo cognitivo, social o emocional.

El problema radical subyace en los problemas derivados de una percepción equivoca del lenguaje. Es muy común que estos sujetos presenten problemas prelocutivos que impidan la absorción instantánea fonética del habla. En los casos más radicales, sin tratamiento alguno, el individuo nunca sale de la fase prelingüística en su desarrollo reduciendo sus capacidades comunicativas a balbuceos y sonidos vocales sin significado alguno.

En algunos casos coexiste una exposición constante a un lenguaje gesticular, lenguaje de signos, que deriva en un desentrenamiento fonético. En estos casos se aprende un lenguaje truncado que les permite la comunicación de forma natural pero de forma más lenta y laboriosa.

Actualmente, se intenta dar cobijo al lenguaje oral como herramienta comunicativa, tendencia conocida como *oralismo*. Esta movimiento surge porque el lenguaje gesticular limita tanto la comunicación como la integración social de estas personas.El oralismo ha levantado numerosos estudios y programas que culminan en fases de entrenamiento con mayor o menos éxito. Los principales problemas detectados fueron:

- Problemas fonológicos: debido a las limitaciones auditivas, las personas con discapacidades auditivas tienden a presentar problemas en la discriminación de sonidos. En los casos menos marcados únicamente están presentes en sonidos consonánticos pero también son presentes los errores en sonidos vocálicos.

- Problemas léxico-semánticos y morfo-sintácticos: las limitaciones en el proceso de aprendizaje derivan en problemas léxico-semánticos acentuados. Estas personas tienden a presentar un vocabulario empobrecido además de dificultades en las relaciones semánticas.

  La pobreza gramatical se traduce en el uso de frases excesivamente simples que complican el proceso de comprensión limitando las capacidades lingüísticas.

En algunos casos, con los avances tecnológicos los dispositivos de corrección auditiva, ya sean audífonos o implantes cocleares, solventan algunos de estos problemas. Pero son de uso limitado y el proceso de adaptación no suele ser exitoso en todos los casos.

Capítulo 3

Fonación

La fonación está constituida por el conjunto de procesos psico-fisiológicos que intervienen en la producción sonora y en la articulación de palabras. Durante el proceso el ser humano hace uso de la mayoría de los órganos del sistema respiratorio por lo que a lo largo de los años se ha hecho un estudio de la morfología fonética tanto del sistema de oxigenación como del instrumentó de producción de sonidos[20], *sistema de fonador humano.* Generalmente se estructura en un sistema simplificado conocido *módelo físico del aparato fonador humano,* esquematizado en la figura 3.1, estructurado en tres partes fundamentalmente:

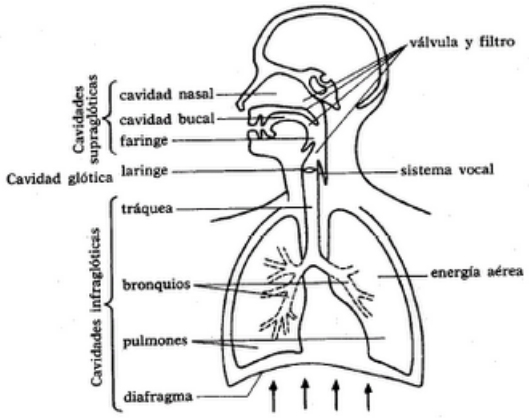

Figure 3.1: Aparato fonador humano

- El sistema subglotal: compuesto principalmente por los pulmones y la musculatura respiratoria asociada tanto a la inspiración la espiración. El funcionamiento del mismo es sencillo, durante el proceso de inspiración la energía almacenada en las paredes elásticas de los pulmones es suficiente para expulsar aire a través de la caja torácica durante la espiración [3]. .

  En la producción del habla la presión que se requiere debe de garantizar la suficiente energía como para producir sonido además de ser constante. Los músculos intercostales garantizan una dilatación adecuada de la caja torácica y una presión estable y suficiente ajustando los niveles de presión de forma automatizada y adecuándose a la situación en que se encuentre el individuo.

- La laringe: actua como una válvula de aire que se abre durante la inspiración y que se cierra durante actos que impliquen rigidez abdominal [3]. Esta compuesta por cartilagos controlados por músculos. La *cricoide,* cartílago base, posa los cartilagos *atirenoides* y *tiroides* que sustentan las cuerdas vocales.

  Estas últimas tienen un aspecto tremendamente importante en el proceso de producción de voz ya que son las que producen impulsos cuasiperiódicos de aire durante la producción fonemas vocálicos. Las cuerdas vocales responden vibrando a las fuerzas generadas del flujo de aire pulmonar. La vibración produce una variación del volumen de aire glotal que se traduce en cambios de presión de aire en el tracto vocal [15, 20].

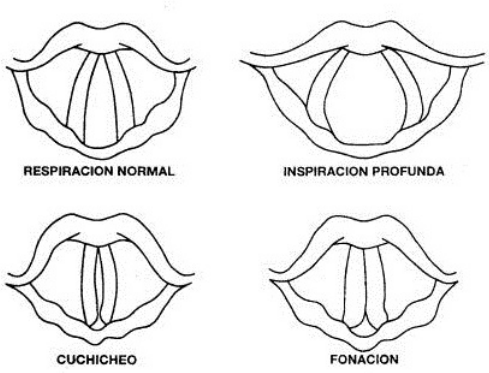

Figure 3.2: Posición de las cuerdas vocales

Fisiológicamente, las cuerdas vocales, están conformadas por un conjunto de músculos y tejidos con una determinada apertura (glotis) sujetos entre los tres cartílagos. El tamaño de los cartílagos así como la longitud de la misma definen la frecuencia de vibración.

Cuando se produce la vibración de las mismas los sonidos generados pasan a través del tracto vocal sin impedimentos con un alto contenido energético dando lugar a lo que comúnmente se conoce como sonidos sonoros. Cuando estas no vibran existen restricciones importantes al paso del aire limitando el contenido en energía dando lugar a lo que se conoce como sonidos sordos.

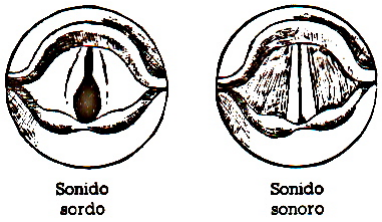

Figure 3.3: Sonidos sordos y sonoros.

- El tracto supralaríngeo: es el tramo del intrumento fonador que modifica substancialmente el flujo de aire glotal para configurar los distintos fonemas comprensibles (sonidos del habla) [27, 3]. En esta etapa interactúan principalmente:
  - • Labios: varían la apertura de la cavidad oral.
  - Cavidad nasal: separada de la cavidad oral modifican la producción mediante la bifurcación del aire a través de la misma.
  - Paladar: gobierna la apertura o cierre de la cavidad nasal.
  - Lengua: permite diferentes configuraciones del tracto vocal.
  - Faringe: conecta la laringe y el esófago con la boca.

La cantidad de músculos que conforman el tracto supra-laríngeo es considerable, es por ello que existen muchas configuraciones y cada persona hace uso de las mismas hasta cierto grado. Genéricamente cada una de ellas se realizada mediante una determinada *articulación*.

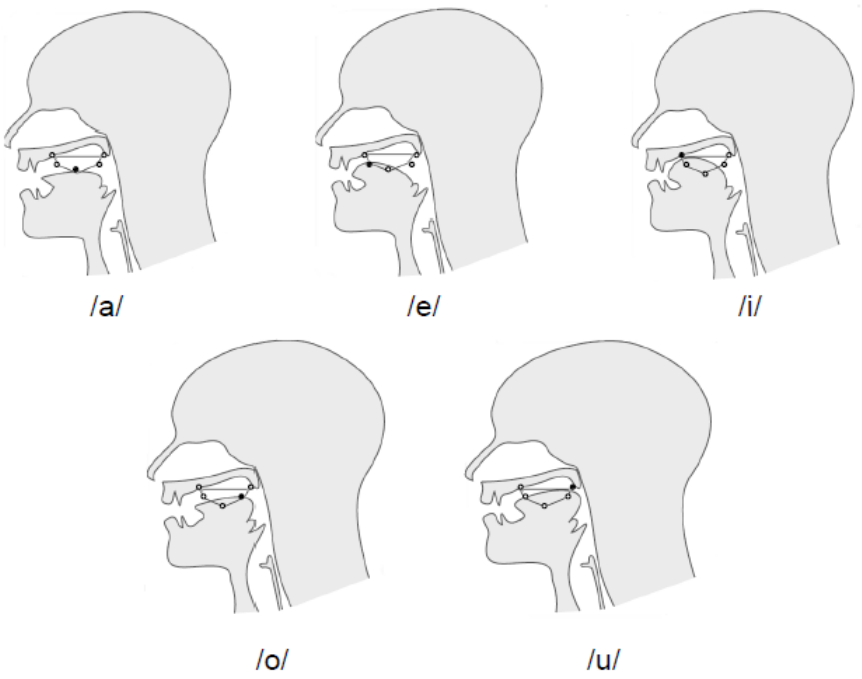

Figure 3.4: Puntos de arituclación para las vocales - Castellano.

Al igual que ocurría con el sistema auditivo es necesario que el sistema fonador funcione adecuadamente para garantizar una correcta comunicación oral. La comunicación oral se basa fundamentalmente en la repetición de sonidos percibidos por lo que se requiere de una coordinación de ambos sistemas. El proceso de aprendizaje lingüístico se acompleja en ausencia de percepción sonora y acentúa problemas derivados de una mala fonación como la dislalia.

La dislalia es un trastorno en la articulación de los fonemas por alteraciones funcionales en los organismos periféricos del habla. Se trata de una incapacidad para formar o pronunciar correctamente ciertos fonemas.

Es una de las anomalías lingüísticas más comunes que sin un tratamiento precoz adecuado puede derivar en problemas significantes en el desarrollo sociopsicológico del individuo debido a la limitación que ejerce en la capacidad comunicativa y adaptación social. Entre las dislalias más comunes podemos destacar las siguientes:

- Rotacismo: la no articulación del fonema /r/.
- Ceceo: pronunciación de /s/ por /z/.
- Seseo: pronunciación de /z/ por /s/.

- Sigmatismo: la no articulación del fonema /s/.
- Jotacismo: la no articulación del fonema /x/.
- Mitacismo: la no articulación del fonema /m/.
- Lambdacismo: la no articulación del fonema /l/.
- Numación: la no articulación del fonema /n/.
- Nuñación: la no articulación del fonema /ñ/.
- Kappacismo: la no articulación del fonema /k/.
- Gammacismo: la no articulación del fonema /g/.
- Ficismo: la no articulación del fonema /f/.
- Chuitismo: la no articulación del fonema /ch/.
- Piscismo: la no articulación del fonema /p/.
- Tetacismo: la no articulación del fonema /t/.
- Yeismo: la no articulación del fonema /ll/.
- Chionismo: sustitución de /rr/ por /l/.
- Checheo: sustitución de /s/ por /ch/.

Por regla general un individuo sano puede presentar alguno de estos síntomas durante su niñez pero dentro de una evolución normal, en la madurez, estas dificultades van siendo superadas. En los casos más radicales los problemas son persistentes, casos conocidos como *dislalia funcional*. Los síntomas más evidentes en estos casos:

- Trastornos fonéticos: alteraciones de la producción propios de aspectos motrices del aparato fonador. En estos casos la percepción y discriminación auditiva es totalmente funcional.
- Trastornos fonológicos: estos son los casos más complejos. Derivan de problemas de percepción y organización derivados de una mala discriminación auditiva.

Los casos de mayor interés son estos últimos ya que requieren de entrenamientos más complejos para su tratamiento. Los trastornos fonológicos tienden a introducir una expresión oral deficiente e incluso inteligible por las continuas desfiguraciones verbales empleadas. Entre los errores más comunes podemos destacar:

- Sustitución: errores de articulación en los que un sonido es reemplazado por otro.
- Distorsión: transmisión de sonidos de forma incorrecta, distorsionada o deformada.
- Omisión: tendencia a omitir aquellos fonemas que no son capaces de pronunciar.

- Adición: intercalación de un sonido incorrecto junto aquel que no son capaces de emitir.
- Inversión: cambios en el orden de los sonidos.

los sonidos. Al igual que ocurría con las disparidades auditivas es necesaria una detección precoz de estos problemas para su correcto tratamiento, un problema que no es trivial y que aun levanta numerosas investigaciones en estudios de fonética y fonología.

3.1 Propagación del sonido en el tracto vocal

En la física clásica la producción de ondas sonoras se debe esencialmente a la vibración de algún cuerpo en un medio elástico. El aparato fonador humano es un sistema complejo de difícil modelación que presenta la dificultades añadidas derivadas de las fuertes variaciones temporales, pérdidas caloríficas, acoplamiento acústico de la cavidad nasal etc.

A lo largo de los años se han presentado diferentes modelos basados en tubos sonoros fundamentados en la función área del tracto vocal, $A(x,t)$. Si consideramos que en el proceso de producción sonora las ondas se propagan en modo plano y no aparecen pérdidas por viscosidad ni por conducción térmica podemos aplicar las leyes de conservación de la masa, momento y energía y derivar las ecuaciones (3.1) y (3.2).

$$-\frac{\partial p\left(x,t\right)}{\partial x}=\rho\frac{\partial}{\partial t}\left[\frac{u\left(x,t\right)}{A\left(x,t\right)}\right] \tag{3.1}$$

$$-\frac{\partial u\left(x,t\right)}{\partial x}=\frac{1}{\rho c^{2}}\frac{\partial}{\partial t}\left[p\left(x,t\right)A\left(x,t\right)\right]+\frac{\partial A\left(x,t\right)}{\partial t} \tag{3.2}$$

1. $p(x,t)$ es la presión de sonido en el tubo
2. $u(x,t)$ es la velocidad del flujo en el volumen
3. ρ densidad del aire en el tubo.

La busqueda de la solución pasa por considerar un estado estacionario donde la función del área del tracto vocal permanece constante en un periodo temporal. A partir de esta suposición se han ido planteando diferentes alternativas siendo la más sencilla el modelo de tubo uniforme sin pérdidas. Partimos la solución de una onda plana longitudinal de forma que:

$$\begin{aligned} p\left(x,t\right) &= P\left(x,\Omega\right)e^{j\Omega t} \\ u\left(x,t\right) &= U\left(x,\Omega\right)e^{j\Omega t} \end{aligned}$$

Sustituyendo en la ecuación (3.1) y (3.2) tenemos que:

$$-\frac{dP}{dx}=ZU \quad siendo \quad Z=j\Omega\frac{\rho}{A} \tag{3.3}$$

$$-\frac{dU}{dx}=YP \quad siendo \quad Y=j\Omega\frac{A}{\rho c^{2}} \tag{3.4}$$

Donde los parámetros Z e Y conforman la *impedancia y admitancia acústica por unidad de longitud* respectivamente. Las soluciones a ambas ecuaciones son de la forma:

$$P(x,\Omega) = Ae^{\gamma x} + Be^{-\gamma x} \tag{3.5}$$

$$U(x,\Omega) = Ce^{\gamma x} + De^{-\gamma x} \tag{3.6}$$

$$\gamma = \sqrt{ZY} = j\frac{\Omega}{c} \tag{3.7}$$

Aplicando las condiciones de contorno en el extremo abierto, $P(L,\Omega) = 0$, así como en la glotis, $U(0,\Omega) = U_G(\Omega)$, derivamos como solución:

$$p(x,t) = jZ_0 \frac{\sin\left(\Omega\frac{L-x}{c}\right)}{\cos\left(\Omega\frac{L}{c}\right)} U_G(\Omega)\, e^{j\Omega t} \tag{3.8}$$

$$u(x,t) = \frac{\cos\left(\Omega\frac{L-x}{c}\right)}{\cos\left(\Omega\frac{L}{c}\right)} U_G(\Omega)\, e^{j\Omega t} \tag{3.9}$$

$$Z_0 = \frac{\rho c}{A}$$

siendo Z_0 la *impedancia característica del tubo.* Si caracterizamos para el caso de la velocidad del aire en los labios, tenemos que:

$$u(L,t) = U(L,t)\, e^{j\Omega t} = \frac{1}{\cos\left(\Omega\frac{L}{c}\right)} U_G(\Omega)\, e^{j\Omega t}$$

Una propiedad imporante es la relación entre la entrada y salida de las velocidad de volumen que representa la respuesta en frecuencia del sistema, ecuación (3.10)

$$V_a(j\Omega) = \frac{U(L,\Omega)}{U_G(\Omega)} = \frac{1}{\cos\left(\Omega\frac{L}{c}\right)} \tag{3.10}$$

Esta función presenta infinitos polos que se corresponden con cada una frecuencias resonantes, también conocidas como *frecuencias formantes* en términos de producción del habla.

3.2 Modelo digital

La idea fundamental es la construcción de un modelo en el que modelamos el complejo sistema fonador humano en un conjunto de bloques independientes. Una de las formas de implementar este sistema es considerar la figura3.5, donde modelamos de forma independiente cada uno de los organos más relevantes que conforman el aparato fonador humano[14]. De forma genérica vamos a suponer que el proceso fonador es emitido únicamente por la cavidad oral. Para esos casos la función de transferencia puede interpretarse como:

$$Y(z) = U(z)\, P(z)\, O(z)\, R(z) \tag{3.11}$$

- $Y(z)$: la señal de voz emitida
- $U(z)$: emula la fuente de excitación glotal.
- $P(z)$: función de transferencia que emula los efectos de la laringe.
- $O(z)$: función de transferencia que emula los efectos de la cavidad oral.
- $R(z)$: función de transferencia que emula los efectos de radiacion labial.

Agrupando los efectos de la laringe así como de la cavidad oral en una única función denotada como $V(z)$ [3, 27, 14] , función de transferencia del tracto vocal. De esta forma:

$$Y(z) = U(z)\,V(z)\,R(z) \tag{3.12}$$

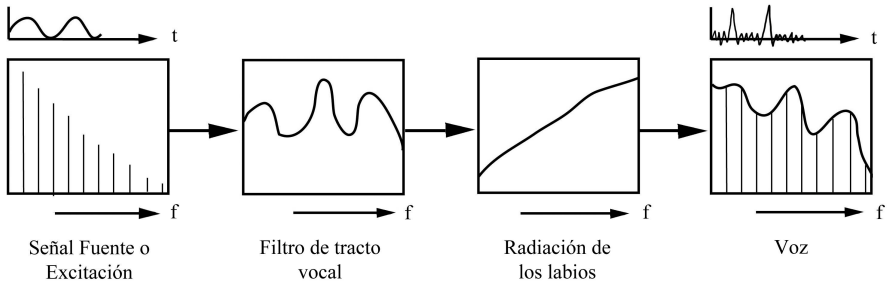

Figure 3.5: Distintos bloques para el modelado del aparato fonador humano

Cada uno de los bloques que la conformás con de vital importancia por lo que realizaremos un análisis independiente de cada uno de ellos:

- Fuente de excitación glotal: • El estudio de la fuente de excitación es de vital importancia ya que define la principal diferencia entre sonidos sordos o sonoros. En el caso más sencillo, la emisión de fonemas no vocálicos, las cuerdas vocales permanecen abiertas por lo que podemos considerar la excitación como un ruido aleatorio con parámetro de gananciaA_N.

  En el caso de fonemas sonoros las cuerdas vocales entran en excitación generando pulsos cuasi-periódicos de la forma presentada en la figura3.6 en los que se pueden distinguir tres etapas:

  - Fase de apertura o abierta: en esta etapa la glotis es abierta
  - Fase de retorno: en esta etapa la glotis comienza a cerrarse debido a la tensión ejercida por el tracto vocal.
  - Fase de cierce: en esta etapa la glotis permanece cerrada.

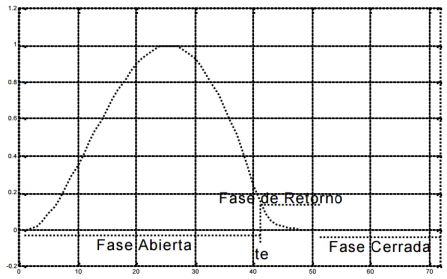

Figure 3.6: Forma de onda volumen-velocidad del pulso glotal

Para la constucción del modelo tendremos que diseñar una fuente que se aproxime a esta excitación. A lo largo de los años se han propuesto diversas soluciones pero una de las más extendidas es la planteada en la ecuación (3.13) [24, 3].donde N_1 marca la mitad de la fase de apertura y N_2 el final del ciclo de retorno.

$$g(n) \begin{cases} \frac{1}{2}\left[1-\cos\left(\frac{\pi n}{N_1}\right)\right] & 0 \le n \le N_1 \\ \cos\left(\pi \frac{n-N_1}{2N_2}\right) & N_1 \le n \le N_1+N_2 \\ 0 & resto \end{cases} \tag{3.13}$$

- Modelado del tracto vocal: En la sección anterior construimos un modelo de tubos sin pérdidas que nos daba las pautas a seguir para comprender el sistema fonador humano y a partir del mismo construir una respuesta en frecuencia que emulara el comportamiento del tracto vocal, ecuación (3.10). Para la implementación digital buscaremos la función de transferencia equivalente a la respuesta en frecuencia del modelo que mejor se adapte [14, 27]:

$$V(z) = \frac{G}{1-\sum_{k=1}^{N} \alpha_k z^{-k}} \tag{3.14}$$

Al ser reales los coeficientes del denominador los polos de la función o bien son reales o son pares complejos conjugados y dado que el sistema fonador humano es un sistema estable, se cumple que $\alpha_k > 0$ y $|z_k| < 1$.

$$z_k = e^{-\alpha_k T \pm j2\pi F_k T}$$

Considerando además los efectos de la radiación labial $R(z)$, tenemos que:

$$P_L(z) = R(z)\,U_L(z) \quad siendo \quad R(z) \approx R_0\left(1-z^{-1}\right) \tag{3.15}$$

- Modelado de los efectos de radiación labial: en el proceso de captura de sonido muchos de los micrófonos utilizados operan a una distancia muy cercana a los labios del hablante. En esta circunstancia la señal se ve afectada por una impedancia de radiación producida por la misma

naturaleza de los labios. Mediante el modelado de tubos sin pérdidas se puede demostrar que esta vendrá dada de la forma:

$$R(z) = \frac{U_L(z)}{P_L(z)}$$

Generalmente el modelado exacto se hace tremendamente complicado por lo que es común hacer uso de un modelo simplificado que aproxime adecuadamente los efectos de la radiación. Generalmente es modelada mediante un filtro FIR, ecuación (3.16). El parámetro α es de vital importancia, se tiende a tomar $\alpha = 0.97$ que simula adecuadamente la cadida de 6dB por octava típica en señales de voz [3, 14].

$$R(z) = 1 - \alpha z^{-1} \tag{3.16}$$

Part II

Metodología

Capítulo 4

Procesado de la señal de voz

4.1 Pre-Procesado de la señal de voz

En el proceso de codificación de ondas sonoras el objetivo fundamental es la obtención de una representación digital óptima de estas últimas. El paso inicial es la obtención de alguna representación primaria mediante un sensor de captación, de ahora en adelante el micrófono de nuestro dispositivo, para un posterior post-proceso. La señal eléctrica obtenida, $x(t)$, se someterá a un agregado de procesos interpretativos previos para obtener una representación comprimida [2, 29] que facilitará el proceso de extracción de parámetros. Por regla general todo este proceso se suele desglosar en dos bloques genéricos:

- **Muestreo y codificación**: la señal temporal es muestreada a una determinada frecuencia, F_s. Posteriormente es cuantizada por un codificador aproximando cada uno de los valores a un conjunto de intervalos conocidos como *niveles de cuantización*.
- **Restauración y realce**: técnicas aplicadas para eliminar ruidos y compensar la degradación producida por la comprensión de la señal. Generalmente se suele utilizar operadores invariantes en el tiempo.

Cada uno de estos pasos es englobado por distintos autores en diferentes etapas. Por regla general, para un correcto análisis de la señal de voz se hace necesario un procesado previo para solventar los problemas derivados de la comprensión de la señal captada por el micrófono[29, 21]. Todo este ciclo es abarcado en varias pautas bien diferenciadas [22], generalmente:

1. Etapa de compensación DC.
2. Etapa de realce de altas frecuencias de la señal de voz.
3. Etapa de fragmentación y ventanado de la señal de voz.

En este capitulo desglosaremos los conocimientos requeridos para una correcta implementación a nivel computacional de cada una de ellas.

4.1.1 Compensación DC

La electrónica utilizada en el sistema de captación de sonido no es ideal y es común que en cada una de las etapas que la conforman se introduzca una pequeña tensión continua no deseada denominada *tensión offset*. Para el correcto análisis es necesario eliminar esta componente. El método más extendido para compensar esta componente es el uso de la media para la estimación de la tensión offset. Esta es calculada y posteriormente restada a cada una de las componentes. En nuestro caso, este proceso se implementa mediante un filtro FIR, ecuación (4.1) , que presenta un comportamiento más eficiente[32].

$$H_{of}(z) = \frac{1 - z^{-1}}{1 - \alpha z^{-1}} \qquad \alpha \in [0.991, 0.999] \tag{4.1}$$

El comportamiento en frecuencia de este filtro, figura 4.1, es plano prácticamente en la totalidad del espectro. Sí observamos en la región de interés, frecuencias extremadamente bajas, podemos distinguir una caída en la amplitud pronunciada que atenúa considerablemente las componentes de baja frecuencia.

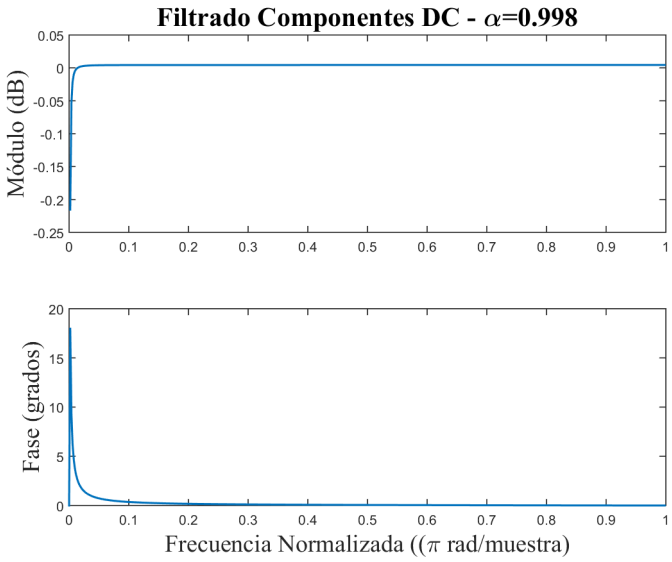

Figure 4.1: Filtro para la compensación de la tensión DC

4.1.2 Realce de altas frecuencias

Paralelamente a los errores introducidos por la tensión offset siguen coexistiendo numerosos que afectan esencialmente a las componentes de alta frecuencia. Estos se ven afectados tanto por el proceso de muestreo y cuantización como por la propia naturaleza de la señal. Algunos de los más relevantes:

1. Los segmentos sonoros presentan una pendiente espectral negativa -20dB/dec [23]

2. La audición es más sensible para frecuencias superiores a 1KHz.

3. El habla humano experimenta una caída de -6dB por octava al pasar a través del tracto vocal por el efecto del pulso glotal y la radiación labial [23, 1, 2, 9]

4. El dispositivo de captura de datos presenta un comportamiento paso baja.

Concluimos entonces que las componentes de alta frecuencia son fuertemente atenuadas por lo que se hace necesario una etapa de realce de altas frecuencias (pre-enfasis). En la práctica se suele utilizar un filtro digital FIR de primer orden que aproxima la implementación de una derivada [21], ecuación (4.2).

$$H(z) = 1 - \alpha z^{-1} \quad 0.95 \leq \alpha \leq 0.98 \tag{4.2}$$

El comportamiento en frecuencia de este filtro, figura 4.2, suaviza y realza el espectro en aproximadamente 20dB/dec. De esta forma podremos reducir la inestabilidad de los cálculos aritméticos y amplificar las componentes de alta frecuencia desglosando un proceso de extracción de parámetros más óptimo.

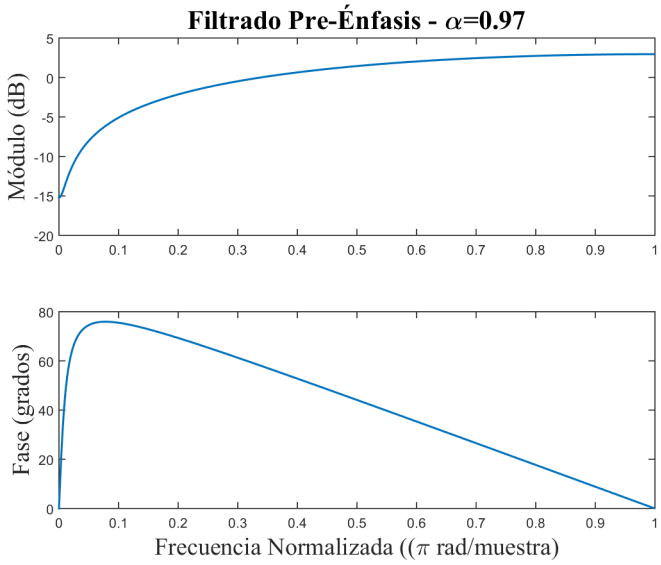

Figure 4.2: Filtro para realce de altas frecuencias

4.1.3 Fragmentación y ventanado

Las señales de voz presentan el inconveniente de que varía temporalmente en periodos cortos de tiempo por lo que el proceso de extracción de parámetros no es eficiente analizando la señal en su totalidad, análisis en tiempo extenso *(Long Time Analysis, STA)*. La clave está en que debemos de buscar un método que nos permita promediar/estimar parámetros de forma fiable mediante pequeños fragmentos temporales, técnica comúnmente llamada como análisis en tiempo corto *(Short Time Analysis , STA) [22, 23]*.

Nos encontramos entonces con el paradigma del dimensionado del fragmento temporal. Teniendo en cuenta que las características de una señal de voz varían temporalmente de forma lenta (el ser humano solo es capaz de emitir de 5-10 sonidos diferentes por segundo) podríamos abarcar distintas situaciones:

- Intervalos de 5-10 ms: producen incertidumbre en la extracción de parámetros ya que en tramas tan pequeñas se producirían cambios tanto en la frecuencia fundamental como en la amplitud de la señal.
- Intervalos de 10-100 ms: es la más recomendada. Produce cierta incertidumbre debido a los cambios de calidad en la señal de voz y los tránsitos entre señales sonoros y sordas.
- Intervalos de 100-500 ms: la gran extensión de esta no permite analizar los cambios en la señal de voz, produce una gran incertidumbre temporal.

En definitiva, independientemente del tamaño de ventana siempre existirá un margen de incertidumbre. La idea es reducir este al mínimo para obtener los parámetros de forma eficiente. La técnica STA intenta solventar estos problemas aislando y procesando los fragmentos (*tramas*) de forma periódica [22]. Para evitar que el análisis de las tramas sea independiente unas a otras, estas suelen

ser solapadas, coexistiendo así cierta correlación entre tramas adyacentes. Para ello se define un determinado tiempo de desplazamiento, T_d, frente a un determinado tiempo de ventana, T_v [23, 29, 39], de esta forma se distinguen distintas situaciones:

1. $T_v < T_d$: en este caso no hay solapamiento entre tramas sucesivas, perdiéndose parte de la señal. Por ello, no se suele utilizar.

2. $T_v = T_d$: si bien no hay pérdida de señal, la inexistencia de correlación en los valores espectrales obtenidos de tramas consecutivas suele ser desaconsejable.

3. $T_v > T_d$: es el caso más habitual. Las tramas adyacentes se solapan, por lo que el análisis espectral posterior tendrá una cierta correlación entre tramas consecutivas. De hecho, si $T_v \gg T_d$, las variaciones entre los valores obtenidos de sucesivos análisis espectrales son muy suaves.

La selección de T_v resulta de un compromiso que es preciso establecer debido a la relación inversa entre resolución espectral y resolución temporal, ecuación (4.3) [1].

$$\Delta f = \frac{\Delta\omega}{2\pi T} = \frac{2\pi}{2\pi T_s N_v} = \frac{f_v}{N_v} \quad siendo \quad T = N_v T_d \tag{4.3}$$

Definiendo un tamaño de ventana extenso obtenemos una buena resolución espectral pero a la par no se observarán los cambios bruscos de la señal de voz debido a la pérdida de resolución temporal. Un valor óptimo de T_v sería el correspondiente al período de la señal, el inconveniente estaría en que dicho período no solo es variable de un locutor a otro sino que es variable en el tiempo para un mismo locutor por lo que se hace complejo el uso de esta medida. Un valor de T_v menor que el período de la señal produciría un análisis distorsionado ya que la porción de señal analizada está truncada de forma artificial, por lo que esta opción queda también descartada.

El valor óptimo comprende varios períodos de señal (los menos posibles). Dado que los mínimos posibles valores del período de la señal varían aproximadamente entre los 40Hz (tono más grave de un hombre) y los 100 Hz (tono menos agudo de una mujer), un compromiso para la selección de T_v [23] así como para la longitud de salto es:

$$10ms \leq T_v \leq 25ms \tag{4.4}$$

$$T_d = K \cdot T_v \qquad K = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}... \tag{4.5}$$

En la práctica el tamaño e incremento puede ser cualquiera, no existe regla que a priori los limite, dandose en ocasiones el caso en que las muestras finales no abarquen el suficiente tamaño de ventana. En estos casos las muestras quedarán despreciadas [9, 29].

Al procesar cada una de las ventanas y finalizar el análisis tendremos una función con N_v valores, uno por cada ventana, el rango de valores calculados dependerá tanto del tamaño de la ventana como de la operación en cuestión. Para normalizar respecto del tamaño, suelen dividirse los valores por el tamaño de la longitud de ventana, [9]. De esta forma tendremos una nueva distribución

temporal, definida como el numero de ventanas por segundo conocida como *framerate*.

$$N_v = \frac{F_s}{R_v} = \frac{1}{T_d} = \frac{1}{KT_v} \tag{4.6}$$

El proceso de fragmentación se realiza mediante un proceso de ventanado. En este punto la elección de la ventana a utilizar determinará las características espectrales de las muestras a analizar ya que aísla una porción de la señal de voz del resto de la señal. La ventana ideal presentaría un único lóbulo principal sin secundarios, hecho que no se puede implementar en la práctica [13]. La solución pasa por buscar un compromiso entre el ancho de banda deseado y la longitud temporal buscada que mejor se adapte a la aplicación en cuestión, por lo que existe una gran variedad de configuraciones posibles. Algunas de las más extendidas son las planteadas en las figuras 4.3, 4.4, y 4.5.

$$w(n) = 0.54 - 0.46\cos\left(2\pi\tfrac{n}{N}\right) \qquad 0 \leq n \leq N$$

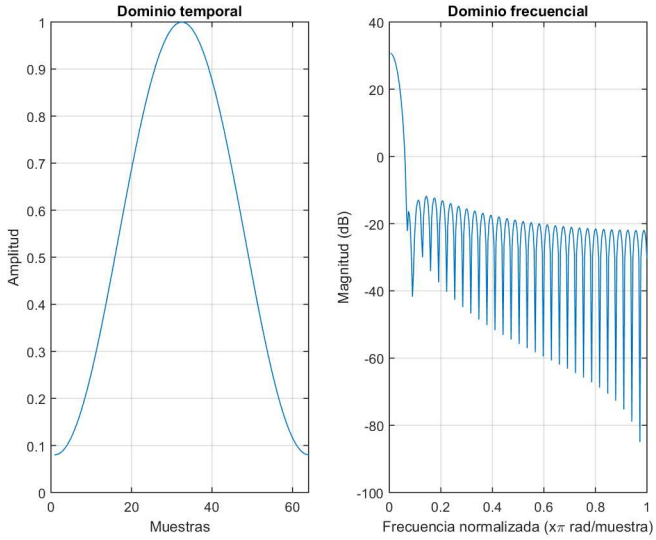

Figure 4.3: Ventana Hamming

Implementación

Procedemos finalmente a observar los resultados obtenidos tras la implementación tomando como ejemplo el fragmento sonoro de la figura 4.6. Le aplicaremos un ventanado de tipo Hamming y posteriormente realizaremos un filtrado para compensación de tensión DC así como realce de componentes de alta frecuencia obteniendo los resultados mostrados de la figura 4.7. Podemos comprobar que las componentes de baja frecuencia, esencialmente las componentes DC, son fuertemente atenuadas así como paralelamente se realiza nos 20dB/dec las componentes de alta frecuencia.

$$w(n) = e^{-\frac{1}{2}\left(\alpha \frac{2n}{N-1}\right)^2} \qquad -\frac{N-1}{2} \leq n \leq \frac{N-1}{2}$$

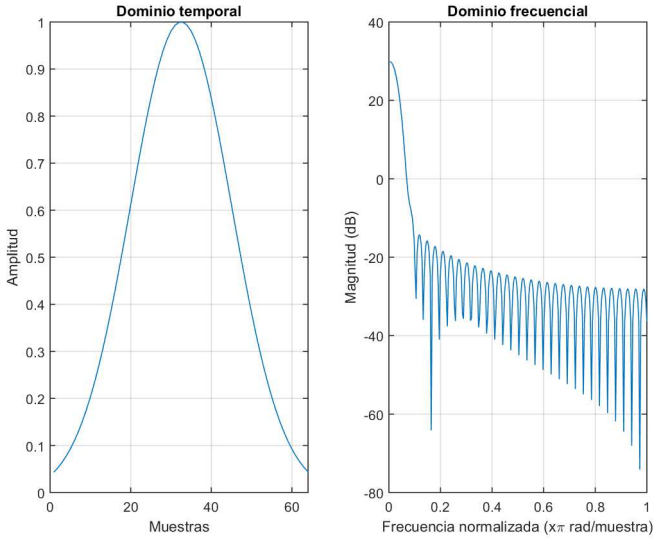

Figure 4.4: Ventana Hanning

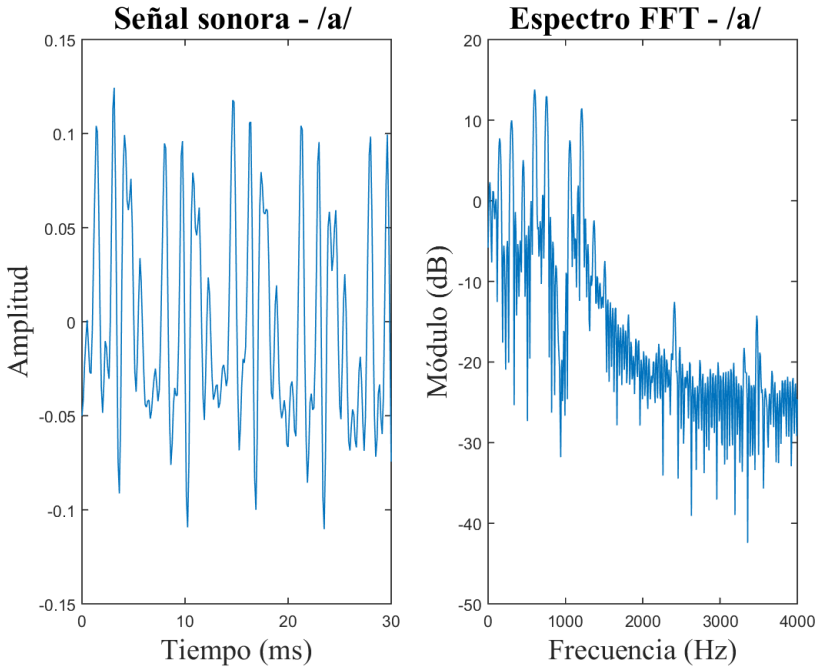

Figure 4.6: Espectro de un fragmento sonoro sin preprocesamiento.

$$w(n) = 0.5\left[1 - \cos\left(2\pi\frac{n}{N}\right)\right] \qquad 0 \leq n \leq N$$

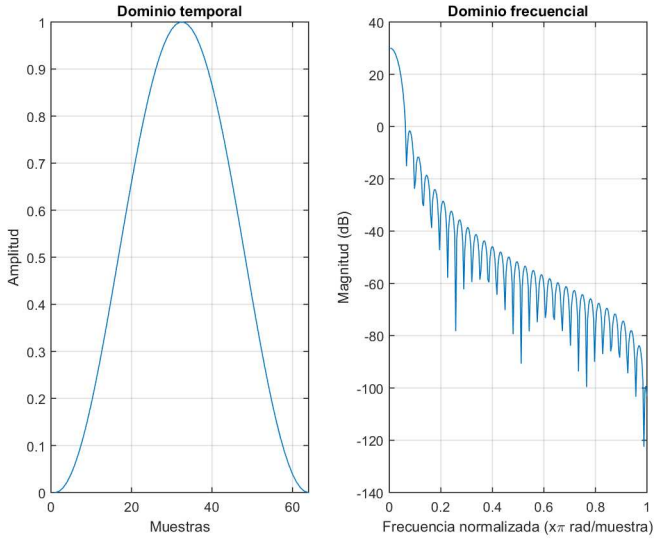

Figure 4.5: Ventana de Gauss

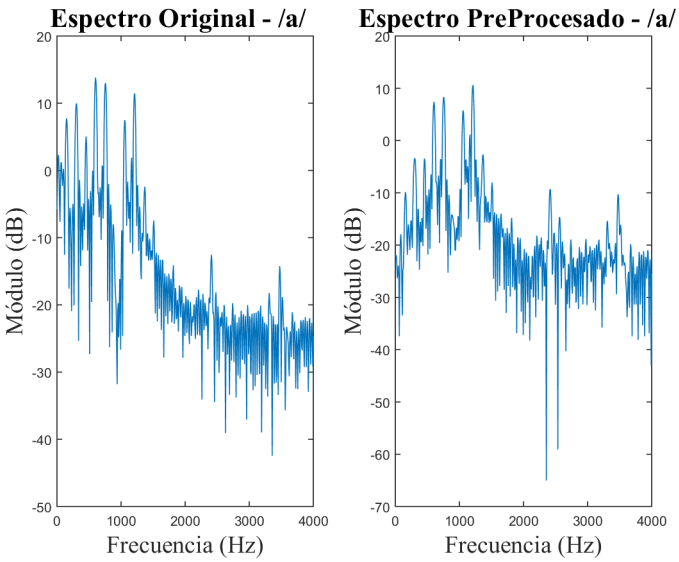

Figure 4.7: Espectro de un fragmento sonoro con y sin preprocesamiento.

4.2 Análisis temporal

Aunque en la práctica el proceso de parametrización de una señal de voz se realiza en su amplia totalidad mediante un análisis espectral es recomendable el estudio temporal de la señal de voz para la extracción de información relevante como los niveles de potencia o el comportamiento oscilatorio de la misma. En la sección anterior se determinó que la señal de voz debía de ser previamente tratada para un correcto análisis. Posteriormente esta era segmentada y ventanada para facilitar el proceso de extracción de parámetros. La mayoría de las técnicas de extracción de parámetros se derivan de la expresión (4.7), [2, 5]. La operación $T[\cdot]$ representa una transformación (lineal o no lineal) invariante en el tiempo que representa una característica de la señal en su totalidad temporal.

$$Q = \int_{-\infty}^{\infty} T[x(t)]\, dt \tag{4.7}$$

Debido a la naturaleza cambiante de la voz, es conveniente el análisis por fragmentos de señal para así observar la evolución de los distintos parámetros calculados [29]. Por ello se utilizan ventanas, $w(t)$, que limitan el espacio temporal de análisis, obteniendo así mediante la función Q característica instantáneas de la señal para cada instante, análisis STA [2].

$$Q(t) = \int_{-\infty}^{\infty} T[x(\tau)\, w(\tau - t)]\, d\tau$$

Esta operación puede interpretarse como una convolución por lo que la señal ventanada es considerada la salida de un filtro paso baja (por la naturaleza de la ventana). Además al ser esta muestreada, para la parametrización de señales discretas, derivamos la expresión (4.8), donde $Q(\hat{n})$se interpreta como los valores medios ponderados de la secuencia en el instante $n = \hat{n}$. La operación en cuestión realiza saltos sobre la señal original de forma que se trata ed un proceso de diezmado que no produciría pérdidas bajo elecciones de ventana adecuadas.

$$Q(\hat{n}) = \sum_{m=-\infty}^{\infty} T[x(m)\, w(n-m)]\Big|_{n=\hat{n}} \tag{4.8}$$

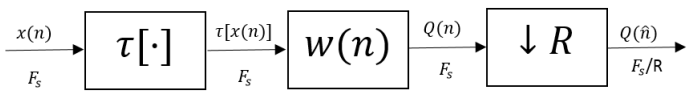

Figure 4.8: Short Time Analysis

4.2.1 Energía en tiempo corto

La energía sonora o energía acústica se puede interpretar como la energía que transmiten o transportan las ondas sonoras. Es un parámetro de uso extendido ya que sirve como primera aproximación para distinguir entre fragmentos sonoros y sordos dado que estos últimos son propensos a presentar menos energía. Analíticamente la energía de una señal se puede obtener según la expresión (4.9) para señales continuas, derivada de esta última la expresión (4.10) utilizada para la estimación de este parámetro en señales discretas.

$$E = \int_{-\infty}^{\infty} x^2(t)\, dt \tag{4.9}$$

$$E = \sum_{n=-\infty}^{\infty} x(n)^2 \tag{4.10}$$

Esta expresión presenta el inconveniente de que tiene poca utilidad para señales de voz ya que realiza un análisis global de la señal despechando todos los cambios temporales. Se hace necesario entonces una adaptación para un análisis en tiempo corto. Entendiendo esta operación como una función Q, (4.8), determinamos que la energía de un fragmento de señal muestreada viene dado según la ecuación (4.11).

$$E(\hat{n}) = \sum_{m=-\infty}^{\infty} T[x(m)\, w(n-m)]\Big|_{n=\hat{n}} = \sum_{m=-\infty}^{\infty} x(m)^2\, w(n-m)^2\Big|_{n=\hat{n}} \tag{4.11}$$

La ventana realiza saltos sobre la señal cuadrática por lo que es un proceso de diezmado que no produciría pérdidas si se elige un valor correcto de dimensión de ventana. Esto ocurre por la propia naturaleza de la ventana que actúa como un filtro paso baja para la operación $T[x(m)]$, por lo que podemos extender la definición considerando un filtro de salida, tal que $h(n) = w(n)^2$, de forma que:

$$E(\hat{n}) = \sum_{m=-\infty}^{\infty} x(m)^2\, h(n-m)\Big|_{n=\hat{n}} \tag{4.12}$$

La definición del filtro $h(n)$ es considerada por muchos autores como un filtro FIR de primer orden para la estimación de la misma, $\hat{w}(n)$. De esta forma, podemos obtener la energía como:

$$E(\hat{n}) = \sum_{m=-\infty}^{\infty} x(m)^2\, w(n-m)^2\Big|_{n=\hat{n}} = \sum_{m=-\infty}^{\infty} x(m)^2\, \hat{w}(n-m)\Big|_{n=\hat{n}}$$

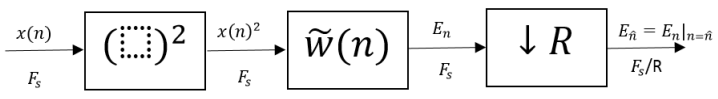

Figure 4.9: Diagrama para el cálculo de la energía en tiempo corto.

Los valores estimados dependerán esencialmente de las dimensiones de ventana seleccionadas por lo que se hace necesario entonces definir un tamaño adecuado de ventana. Este debe de ser capaz de percibir los cambios temporales a la par que suministrar suficientes muestras como para estimar un valor correcto de energía. No existe un valor perfecto de tamaño de ventana por lo que será necesario la estimación de un rango aceptable.

Utilizaremos como señal la representada en la Figura 4.10. Si observamos la figura 4.11 podemos distinguir como para dimensiones de ventana muy extensas no se observan con total nitidez las fluctuaciones temporales. Por otro lado, para periodos inferiores al periodo de la frecuencia fundamental se observan fluctuaciones abruptas que no son deseadas debido a un análisis distorsionado de la señal. Un compromiso razonable se encuentra entorno a los uso de ventanas de aproximadamente 20 ms con tiempos de salto que garanticen unas 100 ventanas por segundo [22].

Una vez determinada una longitud de ventana adecuada determinamos los resultados de la figura 4.12. Es fácilmente apreciable que las tramas sonoras frente a las sordas presente unos niveles de energía más elevados por lo que su uso resulta interesante tanto para la detección de actividad sonora como para la discriminación entre fonemas.

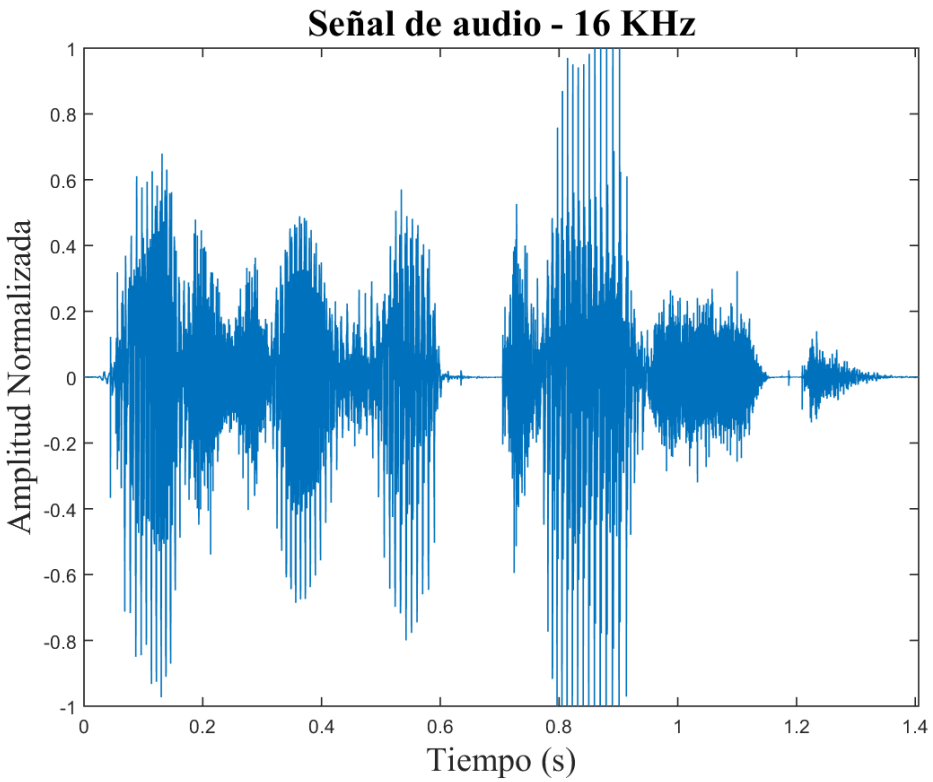

Figure 4.10: Señal de audio de análisis - $F_s = 16KHz$

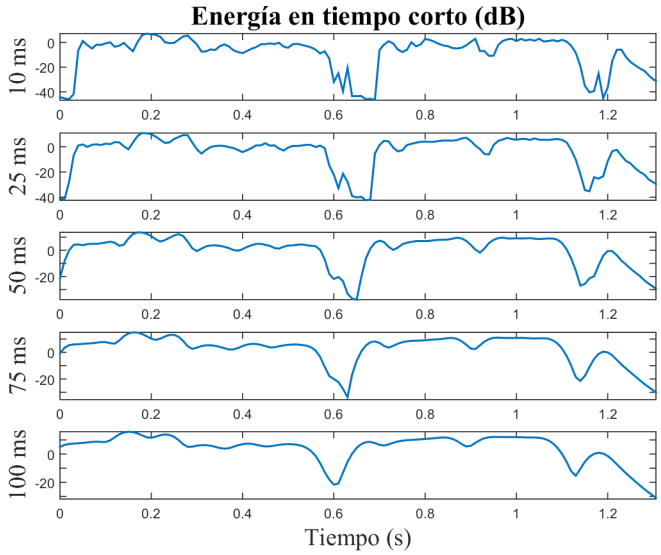

Figure 4.11: Energía en tiempo corto en dB para diferentes tamaños de ventana

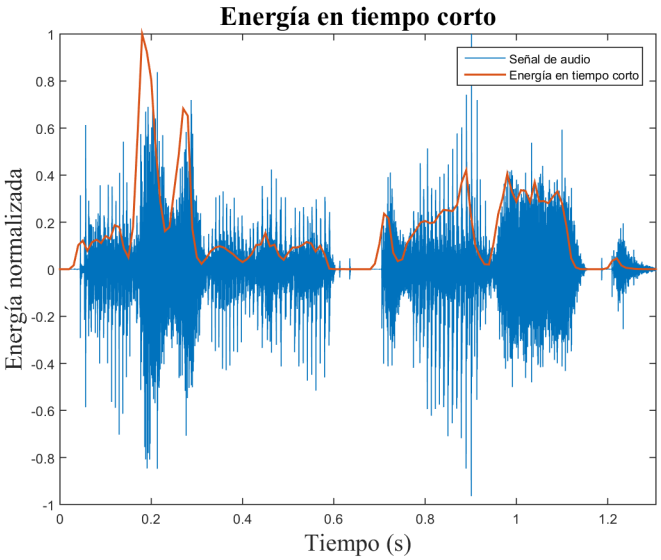

Figure 4.12: Energia en tiempo corto normalizada

Magnitud en tiempo corto

El inconveniente de la estimación de energía es que este parámetro presenta una alta sensibilidad a la amplitud de la señal (aparece elevada al cuadrado). Esto implica un amplio margen dinámico de análisis requiriendo en la mayoría de los casos una transformación logarítmica. Para evitar estas dificultades se extiende el uso de la amplitud media o STAM, *(STAM, Short Time Average*

Magnitude), ecuación (4.13) y (4.15) tanto para señales continua como discretas respectivamente [29, 2, 5].

$$M = \int_{-\infty}^{\infty} |x(t)|\, dt \tag{4.13}$$

$$M = \sum_{m=-\infty}^{\infty} |x(m)| \tag{4.14}$$

Nuevamente tendremos que adaptar la función para su correcto análisis en tiempo corto. Siguiendo las pautas de la sección anterior y partiendo de la función Q, podemos obtener la magnitud de un señal como:

$$M(\hat{n}) = \sum_{m=-\infty}^{\infty} |x(m)\, w(n-m)|\Big|_{n=\hat{n}} = \sum_{m=-\infty}^{\infty} |x(m)|\, \hat{w}(n-m)\Big|_{n=\hat{n}} \tag{4.15}$$

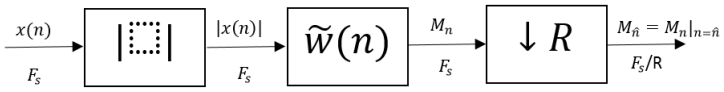

Figure 4.13: Short Time Magnitude -Esquematico

Para la estimación de este parámetro nuevamente toma vital importancia las dimensiones de la ventana seleccionada por lo que nuevamente analizaremos el comportamiento de la función frente a distintos tamaños, figura 4.14 Se observan resultados muy similares al caso de la energía con fluctuaciones menos marcadas pero existentes. Elegimos nuevamente como tamaño adecuado valores entorno a los 20 ms obteniendo los resultados de la figura 4.15.

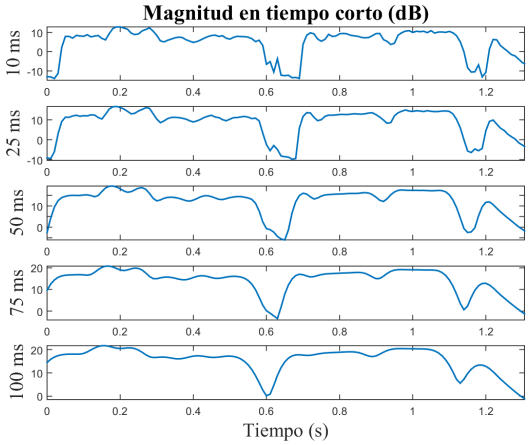

Figure 4.14: Magnitud en tiempo corto en dB para diferentes tamaños de ventana

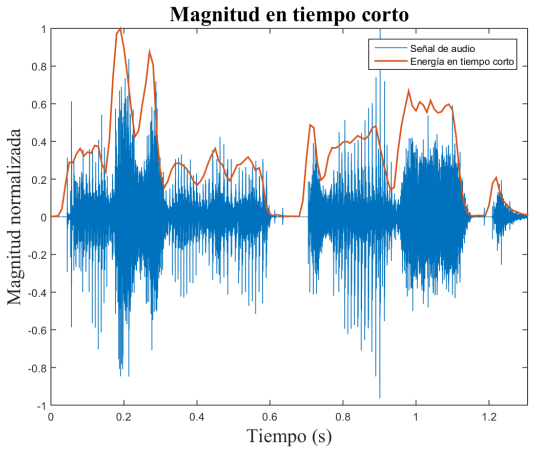

Figure 4.15: Magnitud en tiempo corto normalizada

4.2.2 Tasa de cruces por cero

La tasa de cruces por cero ha sido objeto de numerosos estudios teóricos y prácticos. Mediante esta última se puede obtener una interpretación aproximada del espectro de una señal ya que se presenta como una relación entre las componentes de alta y baja frecuencia, esencialmente si esta es de banda estrecha [22]. Analíticamente para una señal continua en el tiempo viene definida según la expresión (4.16), frente a esta, la expresión derivada para una señal muestreada se puede obtener según la ecuación (4.17).

$$Z = \lim_{T\to\infty}\left(\frac{1}{T}\int_{-T/2}^{T/2}\left|\frac{d}{dt}\left[sgn\left\{x\left(t\right)\right\}\right]\right|dt\right) \tag{4.16}$$

$$Z = \frac{1}{L}\sum_{m=-\infty}^{\infty}\left|sgn\left\{x\left(m\right)w\left(m-n\right)\right\} - sgn\left\{x\left(m-1\right)w\left(m-1-n\right)\right\}\right| \tag{4.17}$$

Generalmente se tiende a utilizar como herramienta de uso extendido para la detección de segmentos fricativos, (señal de baja energía y alta AZCR), para localizar formantes o para una clasificación tosca de la señal de voz. El inconveniente principal de la misma es su alta sensibilidad a componentes adyacentes DC (*offset*) así como la poca robustez al ruido. Esta sensibilidad se acentúa aún más en señales de voz donde se producen cambios bruscos temporalmente por lo que nuevamente se hace necesario una adaptación para el análisis en tiempo corto, ecuación (4.18).

$$Z(\hat{n}) = \frac{1}{N}\sum_{m=0}^{N-1}\left|\frac{sgn\left\{x\left(m\right)w\left(n-m\right)\right\} - sgn\left\{x\left(m-1\right)w\left(n-m-1\right)\right\}}{2}\right|\Bigg|_{n=\hat{n}} \tag{4.18}$$

Si volvemos a reinterpretar la ventana como una operación de filtrado paso baja, podremos simplificar la expresión, de forma que:

$$Z(\hat{n}) = \frac{1}{2N} \sum_{m=0}^{N-1} \left| sgn\left\{x\left(m\right)\right\} - sgn\left\{x\left(m-1\right)\right\} \right| \hat{w}\left(n-m\right) \Big|_{n=\hat{n}}$$

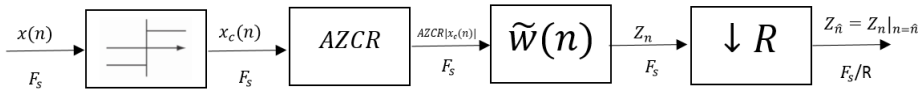

Figure 4.16: Short Time AZCR

Procedemos entonces a distinguir la influencia del tamaño de ventana seleccionado con los resultados obtenidos, figura 4.17. En el mejor de los casos se obtuvieron los resultados de la figura 4.18, donde podemos observar como la periodicidad de las tramas sonoras hace que la tasa de cruces por cero disminuya durante dichos tránsitos e incremente en tramas sordas.

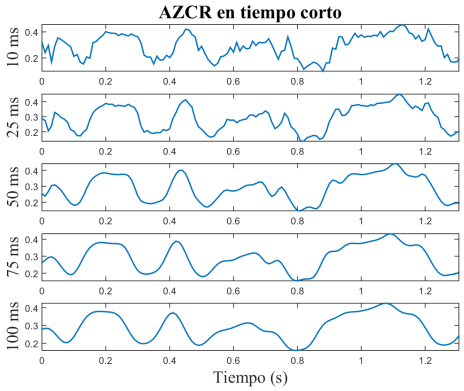

Figure 4.17: Tasa de cruces por cero para diferentes tamaños de ventana

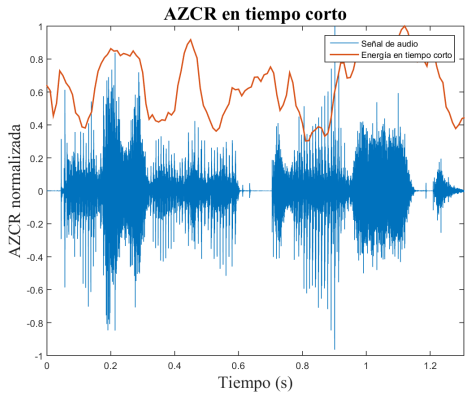

Figure 4.18: Tasa de cruces por cero normalizada

4.3 Frecuencia Fundamenta

La frecuencia fundamental se puede interpretar como la correlación acústica de la tasa de vibración de las cuerdas vocales [40]. Este parámetro es considerado la característica prosódica más importante del habla de ahí que la estimación correcta del mismo sea una labor de vital importancia, esencialmente en aplicaciones de codificación de voz o desarrollo de sistemas para discapacitados (entorno de entrenamiento para personas con problemas auditivos), tema en cuestión de interés. El inconveniente está en que su estimación se ve delimita por distintos factores, esencialmente:

- La onda de excitación glotal no sigue una estructura estrictamente periódica. El ser humano tiene la capacidad de poder expresar más información lingüística de la que expresaría sin modular la frecuencia fundamental (información prosódica) dando más robustez al entendimiento del habla. Esto hace que además de la propia forma de onda, la frecuencia fundamental no solo cambie entre locutores sino que también en un mismo locutor temporalmente [14].

- El conjunto de formantes del tracto vocal tiende a modificar la onda de excitación especialmente cuando se producen cambios bruscos en el habla [16].

- La onda de excitación glotal solo se produce en periodos temporales sonoros por lo que se requiere de un algoritmo robusto para distinguir entre tramos sonoros y tramos sordos. Las transiciones entre estos segmentos son difícilmente detectables por lo que se tiende a estimar la frecuencia fundamental de tramas excitadas por ruido aleatorio [14].

- El ruido de ambiente afecta así como el canal de comunicación delimitan drásticamente la estimación de la misma[33, 30, 19, 40, 16, 14].

A lo largo de los años se han desglosado distintos algoritmos llamados *PDA (Pitch Determination Algorithm)* que presentan ciertos niveles de robustez bajo una serie de caracteristicas concretas. que presentan ciertos niveles de robustez bajo una serie de características concretas. Esto implica que a día de hoy no exista ningún algoritmo que garantice una total fiabilidad y estabilidad en la estimación de este parámetro por lo que sigue siendo fruto de constantes investigaciones y desarrollos. Actualmente podremos agrupar cada uno de estos algoritmos en tres grandes conjuntos: :

1. PDAs basados en las propiedades temporales de la señal de voz.

2. PDAs basados en las propiedades espectrales de la señal de voz.

3. PDAs basados tanto en las propiedades temporales como espectrales de la señal de voz.

Para la aplicación que nos concierne nos centraremos esencialmente en el estudio de algoritmos basados en propiedades temporales de la señal de voz. La propiedad fundamental de las señales periódicas es la similitud en la forma de la propia señal cada cierto periodo temporal. Los PDAs basados en las propiedades temporales de la señal de voz hacen uso de esta última propiedad, implementan

algoritmos que estiman la frecuencia fundamental comparando las similitudes entre la señal original y la señal cambiante [14].

Si la distancia de comparación se aproxima al periodo de la frecuencia fundamental las similitudes existentes se maximizan. Los PDAs se implementan minimizando una función coste cuadrática debido a su tratabilidad matemática. Entre las más extendidas tenemos la función coste de distancia directa, a la cual se le añade los efectos de no estacionalidad de la señal sonora dando como resultados la ecuación (4.20) donde el factor β (ganancia de la frecuencia fundamental) controla los cambios en los niveles de señal [14].

$$E(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} [s(n) - s(n+k)]^2 \tag{4.19}$$

$$E(k,\beta) = \frac{1}{N} \sum_{n=0}^{N-1-k} [s(n) - \beta s(n+k)]^2 \tag{4.20}$$

4.3.1 PDA basado en la función de autocorrelación

La autocorrelación para valores de desplazamiento τ próximos al periodo de señal tiende a presentar un máximo local por lo que podemos concluir que la autocorrelación de una señal periódica es otra señal periódica. En señales de voz este periodo se corresponde con la frecuencia fundamental por lo que no es raro en pensar en esta función como primer método para la estimación de la misma. Asumiendo que la señal de voz analizada presenta características periódicas el criterio de error de la función de coste de la ecuación (4.19) [14] se puede reescribir según la ecuación (4.22) en función de la autocorrelación, ecuación (4.21) [30, 12, 19, 5, 8].

$$r(k) = \sum_{m=0}^{N-1-k} s(m)\, s(m+k) \tag{4.21}$$

$$E(k) = r(0) - r(k) \tag{4.22}$$

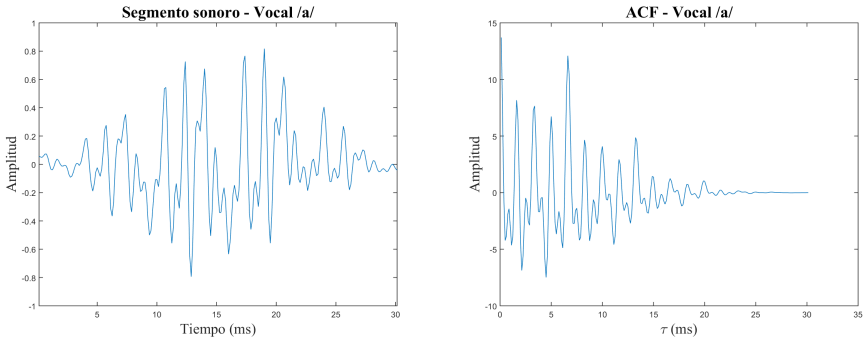

Figure 4.19: Ventana sonora - Señal original y autocorrelación de esta última.

La minimización del error estimado es equivalente a la maximización de la función de autocorrelación. Observando la figura 4.19 distinguimos el primer

máximo local en en $\tau = 0$ característico de la función de autocorrelación. Tomando este como punto de partida la estimación de la frecuencia fundamental se basará estrictamente en la estimación del segundo máximo local.

El inconveniente está en que la señal de autocorrelacion contiene muchos picos atribuidos al amortiguamiento oscilado que procede de la propia señal de voz y que hace que no siempre el primer máximo local se corresponde con el periodo de la frecuencia fundamental. Este puede aparecer desplazado a distintas octavas, figura 4.20. Esto ocurre porque las componentes sonoras son muy ricas en armónicos y aparecen errores asociados a los mismos y a subarmónicos, destacando:

- Detección de medio tono: la frecuencia fundamental detectada es la mitad de la original.
- Deteccuón de doble tono: la frecuencia fundamental detectada es el doble de la original.

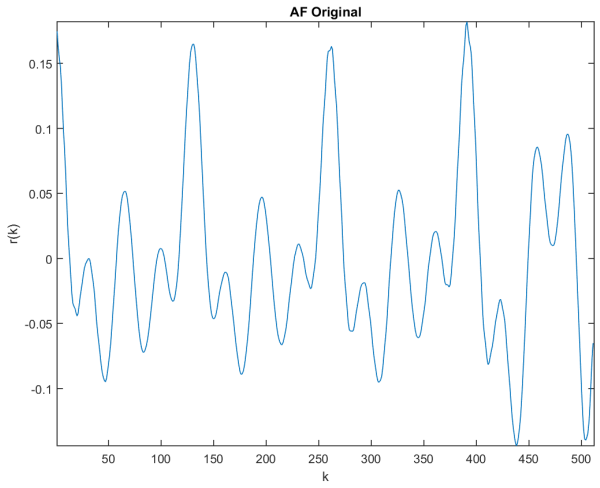

Figure 4.20: Autocorrelación con máximo localizado en subarmónicos de la frecuencia fundamental

Dentro del conjunto de armónicos, el primer formante es el más influyente ya que se haya en regiones muy cercana al margen dinámico de la frecuencia fundamental. Una de las soluciones al problema de los armónicos es la consideración de la condición de no estacionalidad de la señal de voz. La señal de voz presenta cambios temporales que la convierten en una señal no estacionaria hecho que fuerza a grandes errores en el criterio de similitudes [14] por lo que es común hacer uso de un criterio de error normalizado que considera las cualidades no estacionarias de la señal de voz, ecuación (4.20). Si realizamos un proceso de optimización:

$$\frac{\partial E\left(k,\beta\right)}{\partial \beta} = 0 \tag{4.23}$$

encontramos que la ganancia de la frecuencia fundamental óptima puede estimarse como:

$$\beta = \frac{\sum_{n=0}^{N-1-k} s(n)\, s(n+k)}{\sum_{n=0}^{N-1-k} s^2(n+k)} \tag{4.24}$$

que deriva en la función de criterio de error siguiente:

$$E(k,\beta) = \sum_{n=0}^{N-1-k} s^2(n) - \frac{\left[\sum_{n=0}^{N-1-k} s(n)\, s(n+k)\right]^2}{\sum_{n=0}^{N-1-k} s^2(n+k)} \tag{4.25}$$

Por consiguiente, la minimización de esta función se puede interpretar como la maximización de la función de autocorrelación cuadrática normalizada, ecuación (4.26). Como esta función presenta ambigüedades entre los máximos para valores negativos de la función se tiende a eliminar este error eliminando el exponente cuadrático, ecuación (4.27). Podemos observar como para el mismo tramo sonoro de la figura 4.20 se realza el segundo máximo y se reduce los niveles de las componentes armónicas, figura 4.21.

$$r^2(k) = \frac{\left[\sum_{n=0}^{N-1-k} s(n)\, s(n+k)\right]^2}{\sum_{n=0}^{N-1-k} s^2(n+k)} \tag{4.26}$$

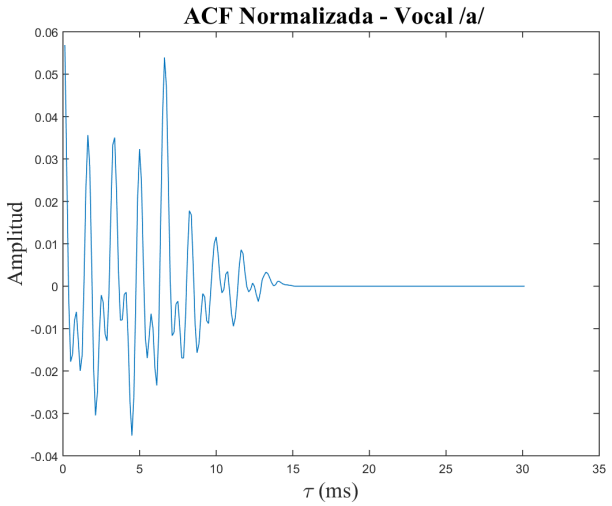

Figure 4.21: Autocorrelación normalizada con máximo localizado en la frecuencia fundamental

$$r(k) = \frac{\sum_{n=0}^{N-1-k} s(n)\, s(n+k)}{\sqrt{\sum_{n=0}^{N-1-k} s^2(n+k)}} \tag{4.27}$$

En la práctica se utilizan ventanas que albergan varios periodos de la frecuencia fundametnal, por lo que parte de la función estimada tiene poca utilidad. Para reducir el impacto de los armonicos secundarios y realzar los maximos locales, se tiende a utilizar la técnica de *media ventana cambiante (half-frame shifting)* que estima la mitad de las muestras pero de una forma más compacta, ecuacion (4.28), figura 4.22, [8]. La versión normalizada viene dada según la ecuación (4.29).

$$r(k) = \sum_{m=0}^{N/2} s(m)\, s(m+k) \qquad k = 0, ..., N/2 \tag{4.28}$$

$$r(k) = \frac{\sum_{m=0}^{N/2} s(m)\, s(m+k)}{\sqrt{\sum_{n=0}^{N/2} s^2(n+k)}} \qquad k = 0, ..., N/2 \tag{4.29}$$

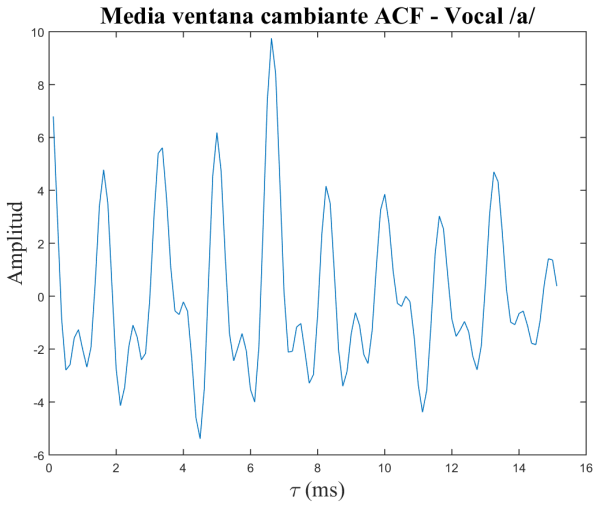

Figure 4.22: Autocorrelación normalizada implementada mediante media ventana cambiante

Tecnicas de pre-procesado

Ligado a todos los problemas anteriores encontramos que el propio tono y estado de emoción del hablante altera los valores de esta función y la existencia de

hablantes secundarios o ruido de fondo puede dar lugar a máximos no deseados. Para minimizar la influencia de los mismos existen un conjunto de técnicas que optimizan el proceso de detección de la frecuencia fundamental. Una técnica muy común es el uso de una función de recorte, *clipping*, que elimina toda la señal de entrada que no sobrepase un determinado umbral. La idea del clipping es atenuar los máximos temporales de la señal para poder discriminar con mayor naturalidad el tramo analizado, ecuación (4.30).

El problema radica en que la elección del nivel de clipping es muy complejo y una elección incorrecta desviaría drásticamente la estimación de la frecuencia fundamental, esencialmente en ambientes ruidos [19, 30]. Por ello se utiliza una alternativa más sencilla y que en la práctica presenta mejores resultados para reducir el impacto de este último, ecuación (4.31).

$$s\left(n\right)\begin{cases} s\left(n\right)-C^{+} & s\left(n\right)\geq C^{+} \\ 0 & C^{-}<s\left(n\right)<C^{+} \\ s\left(n\right)-C^{-} & s\left(n\right)\leq C^{-} \end{cases} \tag{4.30}$$

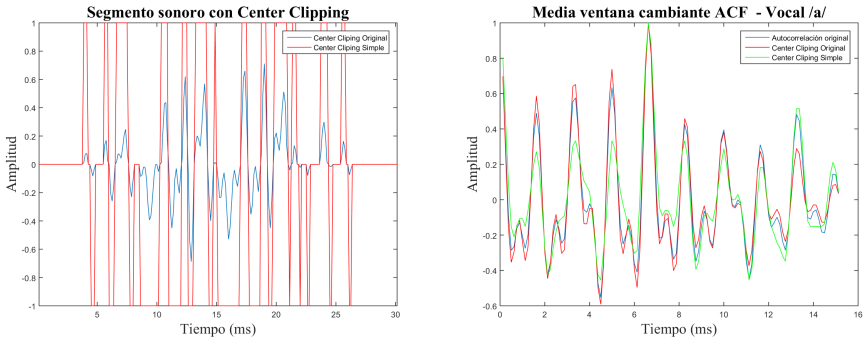

Figure 4.23: Autocorrelación de media ventana cambiante con clipping

$$s\left(n\right)\begin{cases} 1 & s\left(n\right)\geq C^{+} \\ -1 & s\left(n\right)\leq C^{-} \end{cases} \tag{4.31}$$

Podemos observar como los máximos secundarios se ven reducidos frente a los locales que son amplificados. Una vez optimizado el proceso de detección de los máximos secundarios únicamente faltaria por detectar estos últimos y obtener lo valores de la frecuencia fundamental, figura 4.24

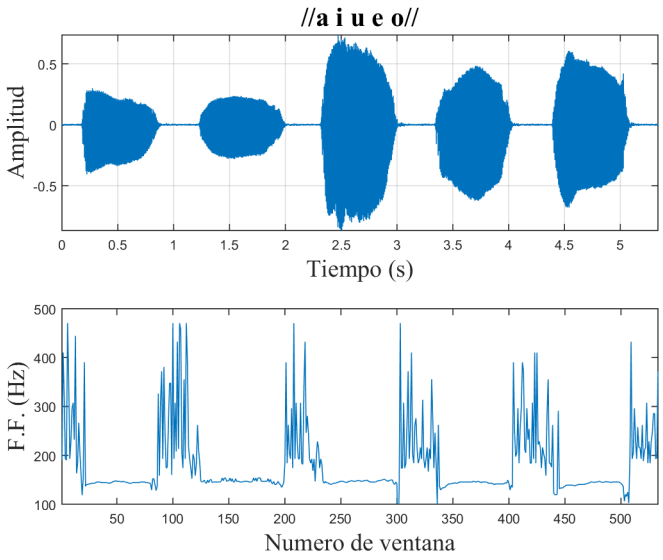

Figure 4.24: Estimación de la frecuencia fundamental mediante algortimo ACF

Si observamos los resultados en secciones sonoras la predicción tiende a ser constante y entre las distintas vocales se estima prácticametne la misma frecuencia fundamental. Para mejorar este sistema vamos a proceder a diferenciar entre tramas sordas y sonoras introduciendo el detector de sonoridad. Posteriormente aplicaremos un filtro mediana (smooth) para suavizar la estimación y observamos los resultados optimizados frente a los originales [30, 34], figura 4.25.

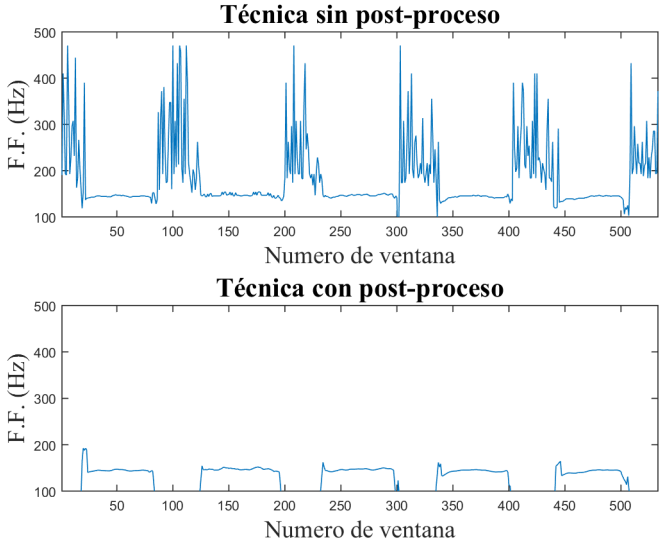

Figure 4.25: Estimación de la frecuencia fundamental mediante algoritmo ACF con detector de actividad sonora.

Concluimos que el algoritmo implementado presenta buenos resultados para señales en ambientes poco ruidosos. Los costes computacionales bajos además de su fácil implementación hacen que el uso de este último sea altamente recomendable, esencialmente en procesos en tiempo real que requieren de costes computacionales bajos.

4.3.2 PDA basado en la función AMD

El método AMDF *(Average Magnitude Difference Function)* es una variante de un análisis de autocorrelación donde en lugar de correlacionar la voz de entrada en varios retrasos, se forma una señal de diferencia entre el retraso en el habla y el original y, en cada retardo, se toma la magnitud absoluta de la diferencia por lo que se puede considerar un criterio de similitud [14]. Al igual que en el caso de la función de autocorrelación, la detección de la frecuencia fundamental se basa en un proceso de optimización. La señal de diferencia tiende a exhibir nulos profundos en los retrasos correspondientes al periodo de señal por lo que es una alternativa que nos permite evaluar y calcular la frecuencia fundamental de una señal de voz[40, 12, 25]. Generalmente viene definida según la ecuación (4.32), figura 4.26.

$$AMDF(\tau) = \sum_{n=0}^{N-1-\tau} |s(n) - s(n+\tau)| \tag{4.32}$$

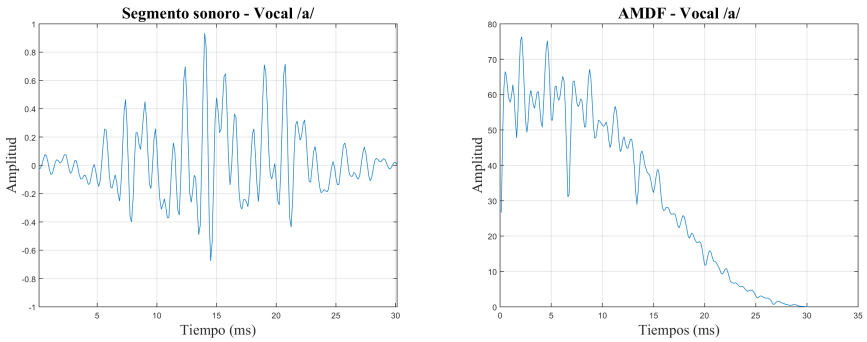

Figure 4.26: Ventana sonora - Señal original y de esta última.

Este es el algoritmo más extendido para el procesado en tiempo real debido a su alta precisión y su bajo coste computacional ya que a diferencia del método de la autocorrelación este no requiere de sumatoria de productos para su estimación [18]. Como desventaja este presenta una alta vulnerabilidad a cambios bruscos en la amplitud de la señal, muy comunes en señales acústicas, debidos a cambios de intensidad sonora o a ruidos externos por lo que su uso únicamente es recomendable en ambientes poco ruidosos. Por otro lado presenta el inconveniente de que no conversa la naturaleza periódica de la señal original más allá de la mitad de la trama donde se produce una caída abrupta. Estos problemas se pueden solventar realizando una implementación alternativa normalizada con media ventana cambiante, ecuación (4.34), que garantiza la conservación de la periodicidad y facilita la detección de los mínimos [14], figura 4.28.

$$AMDF(\tau) = \sum_{n=0}^{N-1-\tau} \frac{|s(n) - s(n+\tau)|}{N-\tau} \tag{4.33}$$

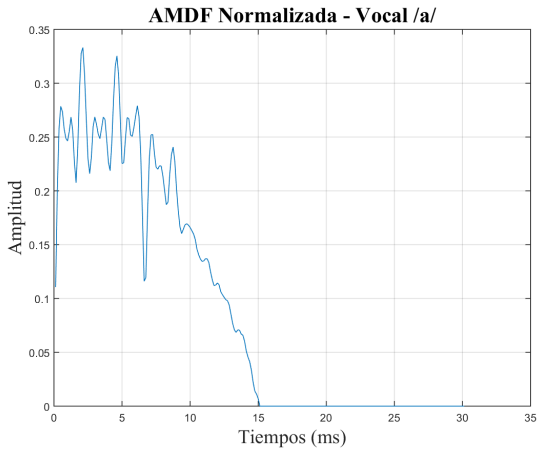

Figure 4.27: AMDF normalizada

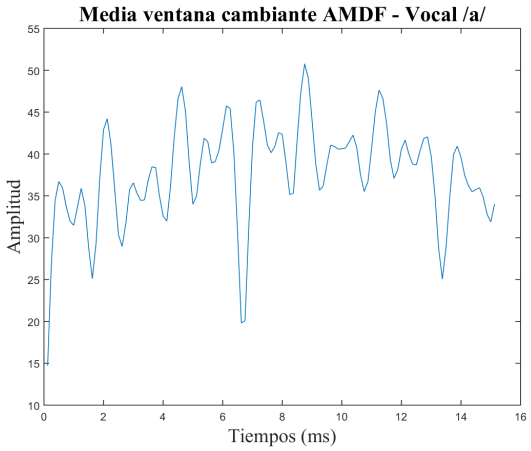

Figure 4.28: Función AMDF implementada mediante media ventana cambiante

Podemos observar como ahora se distingue perfectamente la periodicidad de la señal de voz. La ventaja frente al algoritmo convencional es que esta técnica reduce el inconveniente de la detección del primer mínimo local ya que aminoriza los efectos de los armónicos y por otro lado supone una reducción considerable de los tiempos de computo [40, 8].

$$AMDF(\tau) = \sum_{n=0}^{N/2} |x(n) - x(n+\tau)| \tag{4.34}$$

Lo único que nos quedaría es observar la robustez de este algoritmo entre tramas sonoras y sordas para el cual nuevamente utilizaremos un detector de sonoridad ligado a un posterior proceso de smoothing, figura 4.29. Los resultados obtenidos nuevamente son satisfactorios.

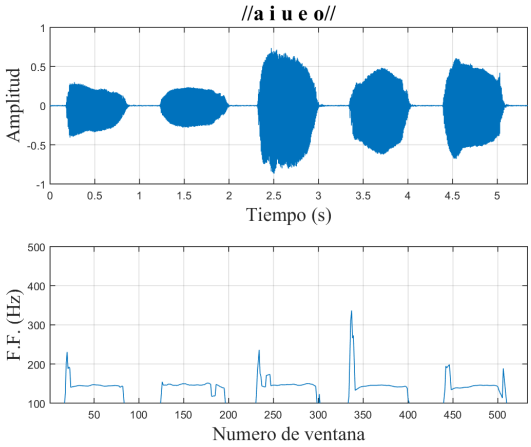

Figure 4.29: Estimación de la frecuencia fundamental mediante algoritmo AMDF con detector de actividad sonora.

4.4 Analisis espectral

En los capítulos anteriores se engloba el proceso de extracción de parámetros centrándonos estrictamente en un análisis temporal. El análisis temporal es simple e intuitivo pero está limitado en lo que a información respecta. En esta sección introducimos los conceptos previos requeridos para realizar un análisis espectral de la señal. Las señales acústicas, tanto en fragmentos sonoras como sordas, presentan una energía infinita por lo que no existe a priori una DTFT (Discrete-Time Fourier Transform) calculable [11].

Nuevamente tendremos que realizar un análisis espectral en tiempo corto. Siguiendo las pautas marcadas en las secciones adaptaremos la función genérica de la DFT para un análisis STA e introduciremos las distintas alternativas para la estimación adecuada de la envolvente compleja del espectro buscando un compromiso entre la resolución temporal y espectral. Esta nueva transformada recibe el nombre de DTFT en tiempo corto (stDTFT, short-time Discrete-Time Fourier Transform) y se define segun la ecuación (4.35). Es una transformada multidimensional que presenta tanto dependencia de la variable discreta temporal como de la variable continua de la frecuencia [5]. Bajo ciertas circunstancias se puede considerar como una transformada de Fourier normal por lo que conservará todas las propiedades de esta última.

$$\overline{S}(\omega) = \frac{1}{N} \sum_{n=0}^{N-1} s(n)\, e^{-j\omega n} \tag{4.35}$$

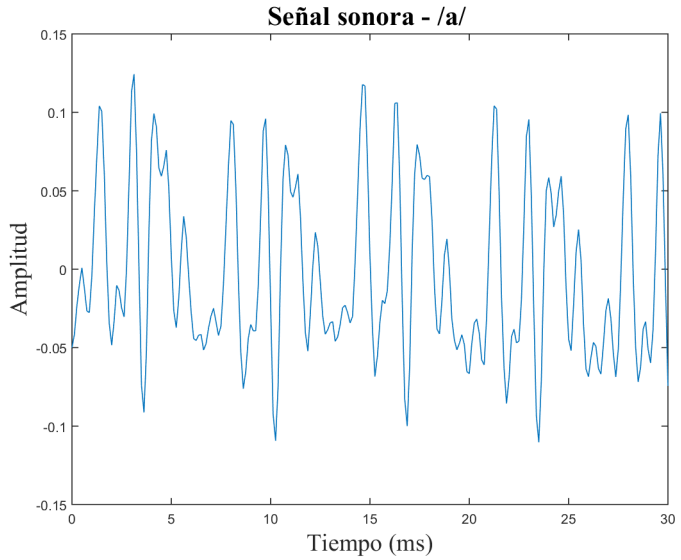

Figure 4.30: Segmento sonoro - Vocal A

El espectro de una señal presenta componentes complejas. Para una visualización más compacta, es muy común representar el espectro en términos de su magnitud y fase. Llegados a este punto se destacan dos factores principales:

- La percepción del oído humano presenta una distribución de frecuencias a escala logarítmica, una señal sonora con el doble de amplitud apenas es percibida con un pequeño incremento. Esto implica que una representación a escala logarítmica sea más natural.

- El ser humano no presenta una percepción aguda de la fase de una señal, por lo que podemos prescindir de la representación de este parámetro.

Es por ello por lo que surge la tendencia de hacer un estudio cuadrático de la magnitud del espectro en términos de la densidad de potencia espectral PSD (Power Spectral Density) buscando siempre un compromiso adecuado entre la resolución temporal y la resolución espectral.

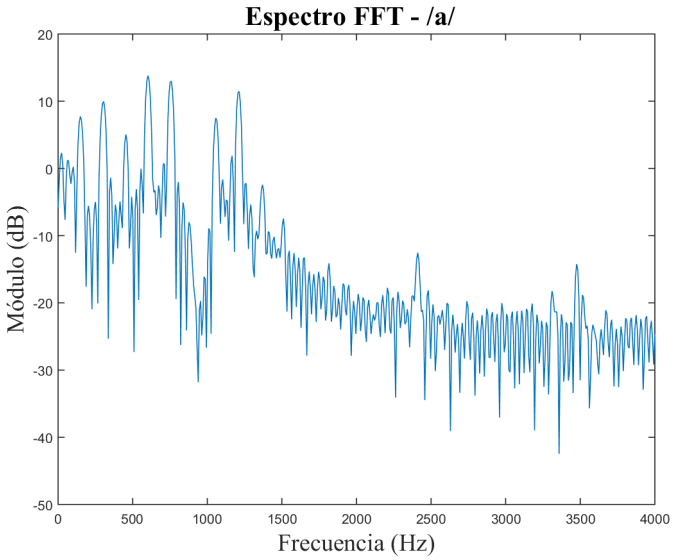

Figure 4.31: Espectro FFT de un segmento sonoro - Vocal A

4.4.1 Coeficientes de Predicción Lineal - LPC

La técnica de predicción lineal es una de las herramientas más utilizadas para el análisis de la señal de voz ya que está fuertemente relacionada con el modelo estimado del aparato fonador humano. En este último, la señal del habla se podía modelar como la salida de un sistema lineal cuya entrada dependía de sí la fonación era sonora o sorda.

La técnica de predicción lineal oferta un método robusto, fiable y seguro para estimar los parámetros que caracterizan el modelo del tracto vocal, $V(z)$, y a partir de los mismos estimar de forma sencilla una buena aproximación de los parámetros fundamentales del mismo: formantes, espectro o la función área del tracto vocal [11, 28].

$$Y(z) = U(z)\,V(z)\,R(z) \text{ siendo} \quad V(z) = \frac{G}{1 - \sum_{k=1}^{P} \alpha_k z^{-k}} \tag{4.36}$$

La idea fundamental de esta técnica es predecir una señal $\hat{s}(n)$ en un instante determinado en función de una combinación lineal de muestras anteriores de la señal original $s(n)$ buscando siempre la minimización del error de estimación [28]. Un predictor lineal ofrece una salida que dependerá tanto del orden de predicción, P, como del conjunto de coeficientes LPC, a_k. Análiticamente la señal predicha se puede obtener según la ecuación (4.37).

$$\hat{s}(n) = \sum_{k=1}^{P} a_k s(n-k) \qquad k = 0, 1, ..., P \tag{4.37}$$

el error cometido vendrá dado entonces como:

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{P} a_k s(n-k) \tag{4.38}$$

Si realizamos una traslado al dominio de la transformada Z podremos derivar la función de transferencia del error de predicción, ecuación (4.40).

$$E(z) = S(z)\left[1 - \sum_{k=1}^{P} a_k z^{-k}\right] \tag{4.39}$$

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^{P} a_k z^{-k} \tag{4.40}$$

Tomando la ecuación genérica simplificada del tracto vocal, ecuación (4.36), se deriva que el filtro de error de predicción puede interpretarse como un filtro inverso para el sistema fonador:

$$V(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{4.41}$$

Por consiguiente, el problema básico para el análisis del tracto vocal se reduce a la busqueda de un conjunto de parámetros a_k que minimicen el error de predicción cuadrático medio.

$$E_m = \sum_m e_n^2(m) = \sum_m \left[s_n(m) - \sum_{i=1}^{P} a_k s_n(m-k)\right]^2 \tag{4.42}$$

Al realizarse el proceso sobre señales acústicas nuevamente se realizará el análisis en tiempo corto. Así pues, el error cuadrático medio en tiempo corto vendrá definido según la ecuación (4.42). Optimizando dicho error, ecuación (4.43) determinamos la expresión de la ecuación (4.44).

$$\frac{\partial E_m}{\partial a_k} = 0 \quad k = 1, 2, ..., p \tag{4.43}$$

$$\sum_m \left[s_n(m-i)\, s_n(m)\right] = \sum_{k=1}^{P} \hat{a}_k \sum_m \left[s_n(m-i)\, s_n(m-k)\right] \qquad 1 \le i \le P \tag{4.44}$$

Teniendo en cuenta que la covarianza localizada de una determinada señal discreta se puede expresar según la ecuación (4.45), se deriva una expresión más compacta de los coeficientes optimizados, ecuación (4.46).

$$\Phi_n(i,k) = \sum_m \left[s_n(m-i)\, s_n(m-k) \right] \tag{4.45}$$

$$\Phi_n(i,0) = \sum_{k=1}^{P} \hat{a}_k \Phi_n(i,k) \qquad 1 \le i \le P \tag{4.46}$$

Asumiendo que la señal de voz ha sido multiplicada por una ventana y que únicamente quedará definida en el intervalo $0 \le n \le N-1$:

$$s_n(m) = s(n+m)\, w(m) \qquad 0 \le m \le N-1$$

el error cuadrático únicamente será distinto en el intervalo $0 \le m \le N-1+P$, de forma que:

$$E_m = \sum_{m=0}^{N-1+P} e_n^2(m) = \sum_{m=0}^{N-1+P} \left[s_n(m) - \sum_{i=1}^{P} a_k s_n(m-k) \right]^2 \tag{4.47}$$

$$\Phi_n(i,k) = r_n(i-k) = \sum_{m=0}^{N-1+P} \left[s_n(m-i)\, s_n(m-k) \right] \tag{4.48}$$

Con los cambios introducidos la ecuación (4.45) quedará reducida a la ecuación (4.48) equivalente al calculo de la autocorrelación, de esta forma:

$$r_n(i) = \sum_{k=1}^{P} \hat{a}_k r_n(i-k) \qquad 1 \le i \le P \tag{4.49}$$

que expresada en forma matricial:

$$\begin{bmatrix} r_n(0) & r_n(1) & r_n(2) & \dots & r_n(P-1) \\ r_n(1) & r_n(0) & r_n(1) & \dots & r_n(P-2) \\ r_n(2) & r_n(1) & r_n(0) & \dots & r_n(P-3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_n(P-1) & r_n(P-2) & r_n(P-3) & \dots & r_n(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(P) \end{bmatrix} \tag{4.50}$$

La matriz planteada puede ser resuelta mediante cualquier técnica de inversión de matrices. El inconveniente de este método es que tiende a presentar costes computacionales elevados y su uso para aplicaciones en tiempo real no es adecuado.

Una solución viable es utilizar técnicas basadas en procesos recursivos como el método de Levinson-Durbin. Este método requiere de $i = P$ iteraciones para construir un conjunto de coeficientes optimizados [5, 11, 14, 28, 38]. El proceso se realiza en varias partes:

1. Previamente, definimos las condiciones iniciales

$$a_0^0 = 0$$

$$E_0^0 = r_n(0)$$

2. Calculamos el coeficiente de reflexión , k_i, de la iteración actual:

$$k_i = \frac{-r_n(i) - \sum_{j=1}^{i-1} a_j^{i-j} r_n(i-j)}{E^{i-1}}$$

3. Estimamos cada uno de los coeficientes de predicción de orden i:

$$a_i^i = k_i$$

$$a_j^i = a_j^{i-1} - k_i a_{i-j}^{i-1} \qquad 1 \leq j \leq i-1$$

4. Actualizamos el error cuadrático medio:

$$E^i = \left(1 - k_i^2\right) E^{i-1}$$

5. Repetimos cada una de las pautas hasta completar todas las iteraciones y construir el conjunto de coeficientes de reflexión así como coeficientes de predicción.

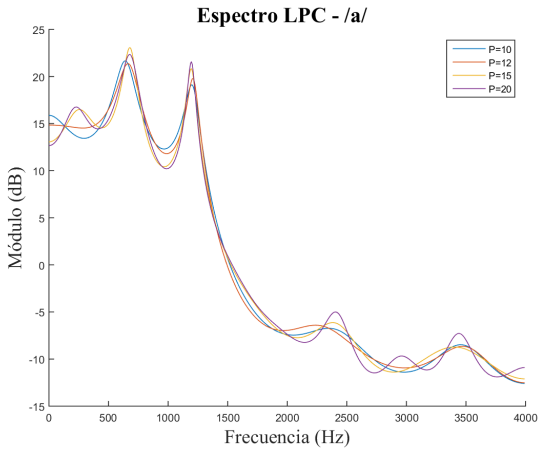

Figure 4.32: Espectro LPC para distintos N - Vocal A

Una vez determinados estos parámetros podremos realizar un análisis más exaustivo espectral de la señal de voz dado que la síntesis LPC aporta una forma automatizada y explicita de separar adecuadamente la fuente de excitación del

propio sistema excitado. La calidad espectral vendrá marcada por el orden de predicción utilizado. Por regla general suele ser inferior al par de decenas. Podemos distinguir el espectro para distintos órdenes del segmento sonoro de la figura 4.30 en la figura 4.32.

Vemos como para órdenes pequeños podemos conseguir una aproximación muy buena de la envolvente del espectro escalado. Dado que la obtención del espectro LPC es sencilla y los tiempos de cómputo muy reducidos es muy común hacer uso de este espectro para la extracción de parámetros. Podemos observar el comportamiento envolvente de este último frente al espectro FFT en la figura 4.33.

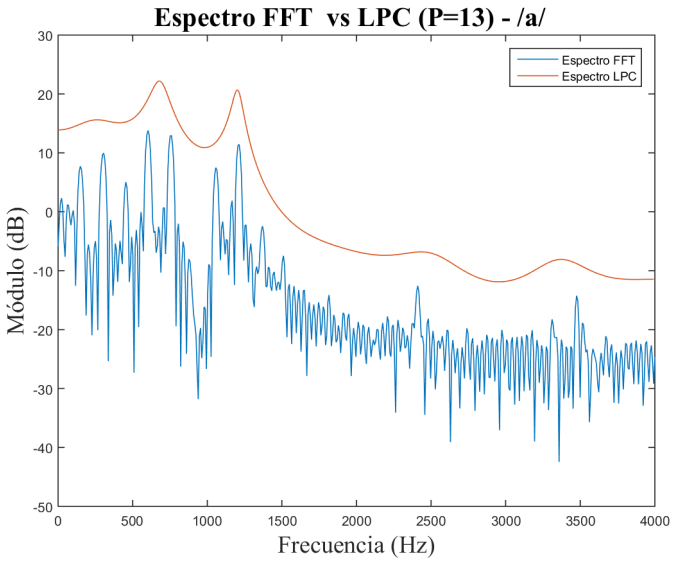

Figure 4.33: Espectro LPC frente al espectro FFT de un segmento sonoro - Vocal A

4.4.2 Cepstrum

En el modelo construido del aparato fonador humano, ecuación (4.36), tenemos el inconveniente de que cada uno de los componentes que lo conforman no son fácilmente separables. El análisis LPC presenta una metodología sencilla para el estudio del tracto vocal pero no presenta información alguna sobre la fuente de excitación.

El cepstrum, ecuación 4.51, de uso extendido en análisis de señales homomórficas, permite la separación espectral de señales combinadas mediante la convolución. Este hecho aplicado al análisis de señales bioacústicas permite la separación de las componentes en frecuencia tanto del tracto vocal como de la excitación. El proceso es realizado mediante un filtrado temporal conocido como *lifterging*.

$$c(n) = \mathcal{F}^{-1}\left\{\ln\left|\mathcal{F}\left\{x(n)\right\}\right|\right\} \tag{4.51}$$

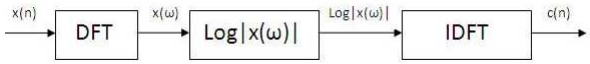

Figure 4.34: Diagrama de bloques para la estimación del cepstrum de una señal

Partiendo de un modelo simplificado en el que los efectos del tracto vocal como de la radiación labial son agrupados en un modelo global $v'(n)$:

$$y(n) = u(n) \otimes v'(n) \quad \underset{\rightarrow}{\mathcal{F}} \quad Y\left(e^{j\omega}\right) = U\left(e^{j\omega}\right) V'\left(e^{j\omega}\right) \tag{4.52}$$

aplicamos el logaritmo de la magnitud de cada una de las componentes, para de esta forma:

$$\ln\left|Y\left(e^{j\omega}\right)\right| = \ln\left|U\left(e^{j\omega}\right) V'\left(e^{j\omega}\right)\right| = \ln\left|U\left(e^{j\omega}\right)\right| + \ln\left|V'\left(e^{j\omega}\right)\right|$$

Logramos entonces una separación tanto de la fuente excitación como del modelo del tracto vocal en el dominio frecuencial. Realizando una conversión al dominio temporal mediante la transformada inversa, obtenemos los coeficientes cepstrales de cada una de las componentes que trabajan en un nuevo dominio, el dominio de la *quefrecuencia*:

$$\mathcal{F}^{-1}\left[\ln\left|Y\left(e^{j\omega}\right)\right|\right] = \mathcal{F}^{-1}\left[\ln\left|U\left(e^{j\omega}\right)\right|\right] + \mathcal{F}^{-1}\left[\ln\left|V'\left(e^{j\omega}\right)\right|\right]$$

de forma más compacta:

$$c_y(n) = c_u(n) + c_v(n)$$

Al usar únicamente la magnitud del espectro "despechando" la información de la fase de nuestra señal obtenemos una nueva distribución temporal conformada por la suma de los efectos del tracto vocal y la excitación y no la convolución de los mismos.

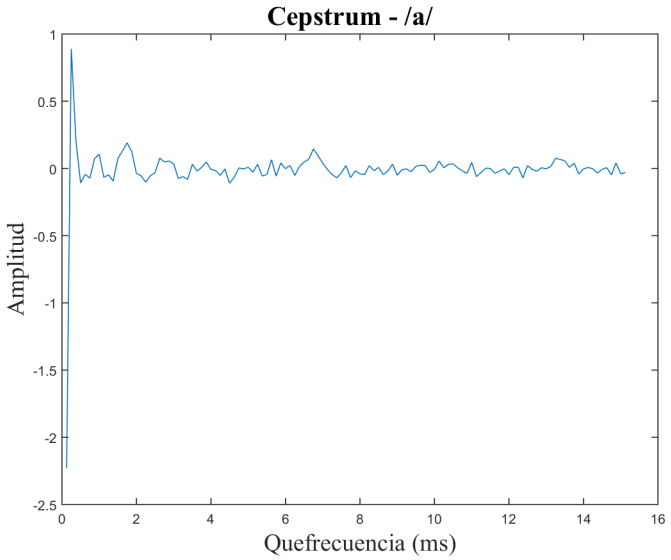

Figure 4.35: Cepstrum de un segmento sonoro - Vocal A

Liftering Paso Baja - Extracción de información del tracto vocal

La extracción de parámetros a partir del cepstrum es un tanto diferente a todo los procesos utilizados hasta ahora ya que utilizan filtros paso baja y paso alta aplicables en el dominio temporal, *quefrecuencia*. Estos filtros son muy sencillos de utilizar y generalmente son implementados mediantes ventanas rectangulares.

Los efectos del tracto vocal conforman las componentes de alta frecuencia del dominio de la quefrecuncia. Albergando únicamente un determinado número de K coeficientes, genéricamente se suele tomar K=30[14], podemos obtener una reconstrucción muy buena del tracto vocal. Para su obtención necesitamos utilizar un filtro paso baja, ecuación (4.53), y realizar las siguientes pautas:

- Una vez calculado el cepstrum, obtenemos la componente cepstral del tracto vocal

$$c_v(n) = c(n)\, w_v(n)$$

- Seguidamente derivamos el espectro del tracto vocal deshaciendo los pasos para obtener el cepstrum:

$$\ln\left|V'\left(e^{j\omega}\right)\right| = \mathcal{F}\left[c_v(n)\right] \quad \Longrightarrow \quad V'\left(e^{j\omega}\right) = e^{\ln\left|V'\left(e^{j\omega}\right)\right|}$$

$$w(n) = \begin{cases} 1 & 0 \leq n \leq K \\ 0 & resto \end{cases} \tag{4.53}$$

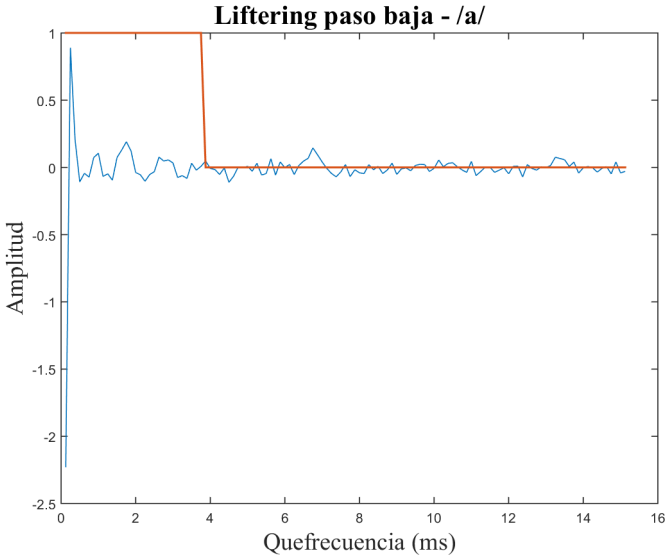

Figure 4.36: Liftering paso baja de un segmento sonoro- Vocal A

Tras realizar un liftering paso baja del cepstrum del segmento sonoro de la figura 4.30 y seguir cada una de las pautas marcadas para obtener el espectro

del tracto vocal obtenemos los resultados de la figura 4.37 así como los de la figura 4.38 frente al espectro original. Podemos observar como la envolvente del tracto vocal se asocia perfectamente al espectro original eliminando todas las componentes ricas en armónicos.

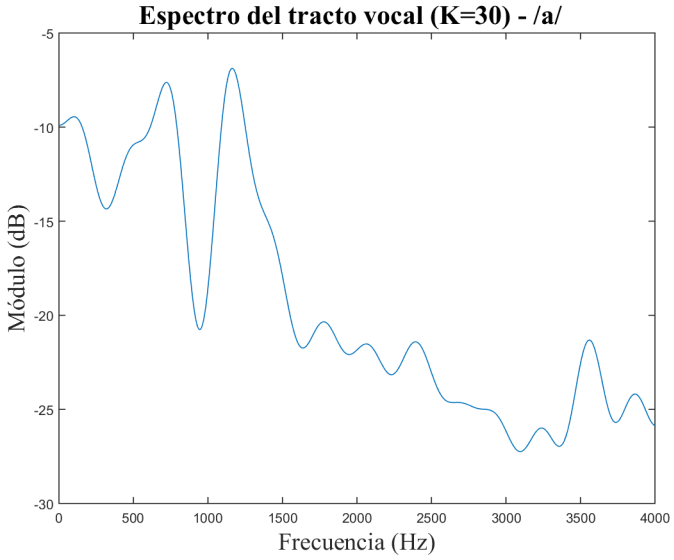

Figure 4.37: Espectro del tracto vocal de un segmento sonoro - Vocal A

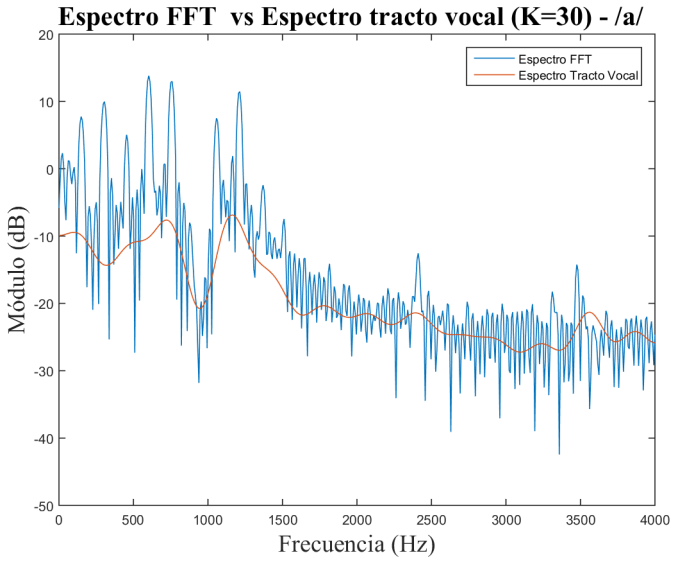

Figure 4.38: Espectro del tracto vocal frente al espectro FFT de un segmento sonoro - Vocal A

Liftering Paso Alta - Extracción de información de la fuente de excitación

Si en el proceso anterior mediante un proceso de liftering paso baja se obtuvo información sobre el tracto vocal, es lógico pensar que en un proceso inverso,extraia la información sobre las componentes de baja frecuencia que conforman la fuente de excitación glotal. Para su obtención necesitamos utilizar un filtro paso alta, ecuación (4.54), y así poder discriminar las componentes de interes:

$$c_u(n) = c(n)\, w_u(n)$$

$$w(n) = \begin{cases} 1 & K \geq n \\ 0 & n < K \end{cases} \tag{4.54}$$

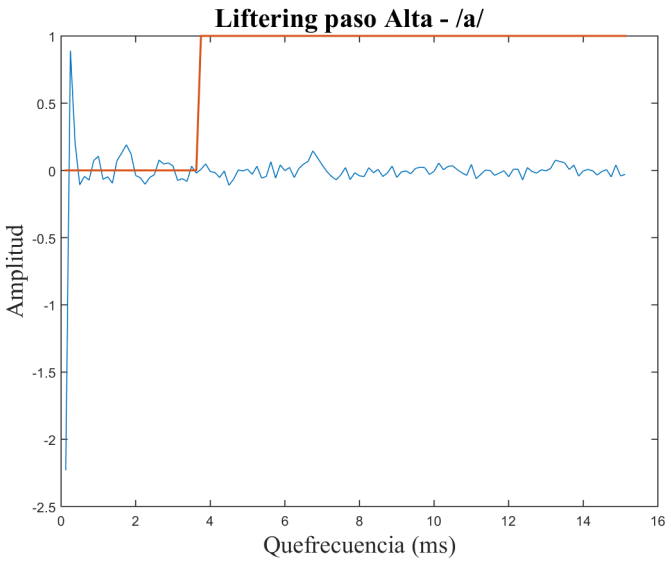

Figure 4.39: Liftering paso alta de un segmento sonoro- Vocal A

La información cepstral extraída contiene información de extrema relevancia como la frecuencia fundamental de la señal de excitación. El proceso se reduce a la estimación del máximo cepstral dentro de un rango determinado. La quefrecuencia se puede interpretar como la representación del retardo del tono fundamental de la señal sonora, luego la región donde se encuentre el mayor promedio energético corresponderá con la frecuencia fundamental de la señal [36].

Nuevamente, nos encontramos con el inconveniente de que no siempre el tono se corresponde con la frecuencia fundamental, más aún si la señal es rica en armónicos, por lo que podríamos encontrar el máximo en frecuencias múltiplo de la frecuencia fundamental. Por otro lado tenemos el problema de que en segmentos sordos el máximo puede anidarse en cualquier posición por lo que

es necesario el uso de un detector de actividad sonora para su correcta implementación [36, 37, 6].

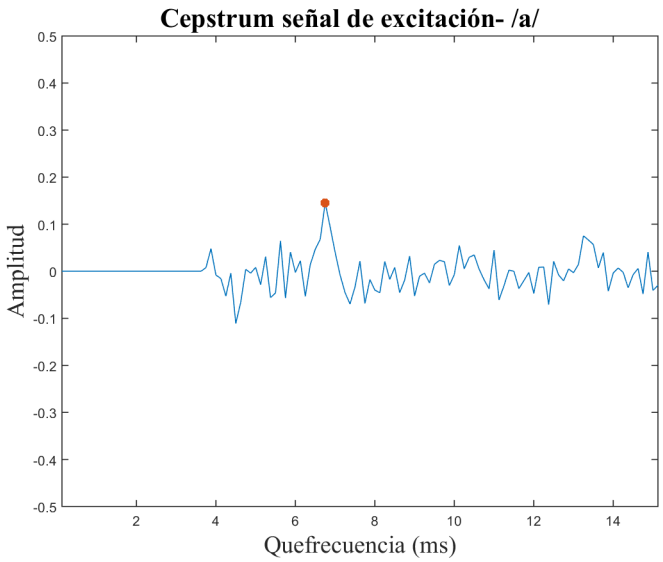

Figure 4.40: Maximo local tras liftering paso alta de un segmento sonoro - Vocal A

Si observamos la figura 4.40, determinamos que el maximo local se encuentra entorno a los 6.75 ms. Esto se traduce en que la frecuencia fundamental del locutor en cuestión se encuentra entorno 148Hz, que se corresponde a un rango común para un locutor varón.

Coeficientes de predicción cepstrales

El proceso de estimación cepstral requiere de varias etapas con operaciones complejas que ralentizan su implementación en sistema de tiempo real. Cuando lo que se busca es únicamente la información relativa al tracto vocal, podemos obtener una aproximación del comportamiento cepstral de una señal a partir de los coeficientes de predicción lineal. La estimación cepstral se construye de forma recursiva por lo que los tiempos de computo son reducidos.

Partiendo de un conjunto de coeficientes LPC, derivamos coeficientes de predicción cepstrales basándonos en la ecuación (4.55) [7, 6].

$$c(n)\begin{cases} \log(G) & n=0 \\ a_n + \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k} & 1 \leq n \leq P \\ \sum_{k=n-p}^{n-1} \frac{k}{n} c_k a_{n-k} & n > P \end{cases} \tag{4.55}$$

Es una forma sencilla de obtener la información más relevante ya que se ha demostrado que órdenes de estimación pequeños son más que suficiente para la

estimación del espectro, además de presentar una mayor robustez que el uso del espectro LPC.

4.4.3 Banco de filtros

Los bancos de filtros permiten separar una señal en un conjunto de señales sub-banda que ocupan una determinada porción del espectro de frecuencias de la señal original. El espectro es reconstruido mediante un conjunto de M filtros paso-banda, $H_k(z)$, con una entrada común o salida sumada, figura 4.41. La implementación de los mismos se basa en la construcción de cada uno de los filtros centrados en torno a una frecuencia concreta y con un determinado ancho de banda. Mediante el filtrado del espectro se deriva una representación energética promediada del espectro de frecuencias en diferentes bandas.

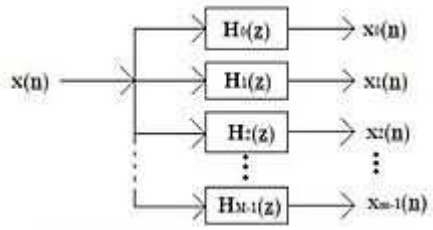

Figure 4.41: Esquema de un banco de filtros

La distribución de las frecuencias centrales de cada uno de los filtros permite un análisis espectral a diferentes escalas. Como la sensibilidad del oído humano a las distintas frecuencias no sigue una distribución lineal sino logarítmica el uso de estos filtros es muy extendido. Numerosos estudios han desglosado un gran número de escalas auditivas que definen un determinado rango de frecuencias que se adaptan a la sensibilidad humana [14, 38]. Una de las escalas más extendidas es la escala Mel, figura 4.42, que plantea intervalos equiespaciados basados en observaciones experimentales con numerosos individuos [26], siendo el punto de partida la equiparación de un tono de 1000Hz con un tono de 1000 Mels. La relación entre escalas viene dada como:

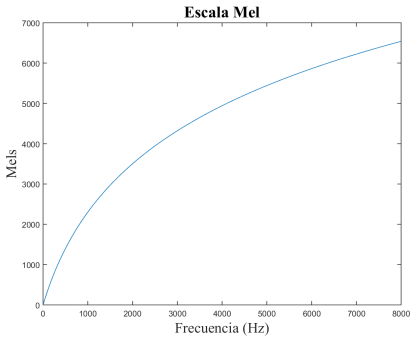

Figure 4.42: Escala de Mel

$$m = 2595 \log\left(1 + \frac{f}{700}\right) \tag{4.56}$$

Estudios recientes han demostrado que una escala más ajustada al oido humano sigue una distribución lineal para las frecuencias bajas y una distribución logaritmica para las altas frecuencias [14, 28]. La escala que mejor se ajusta a esta distribución es la escala Bark[41], ecuación (4.43), representada en la figura 4.43. Esta escala tiene un rango del 1-24 que se corresponde con las primeras 24 bandas críticas del oido humano [35].

$$B = 13 \arctan\left(0.76 \frac{f}{1000}\right) + 3.5 \arctan\left(\left(\frac{f}{7500}\right)^2\right) \tag{4.57}$$

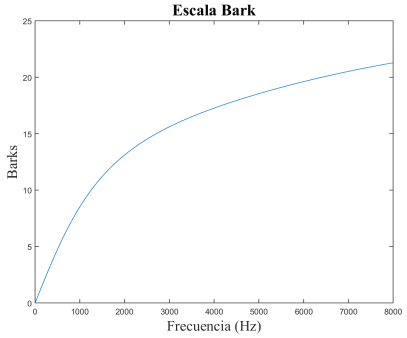

Figure 4.43: Escala de Bark

En la práctica podemos utilizar cualquiera de las escalas definidas para obtener unos resultados adecuados como frecuencias centrales del conjunto de filtros, aunque por su facilidad es común utilizar la escala de Mel. Los filtros pueden ser implementados como filtros triangulares [37, 4, 38], y la salida de los mismos, media ponderada energética, se plantea en la ecuación (4.58).

$$X_k(k) = \sum_{i=f_{ini}}^{f_{ini}+B} |X(i)|^2 H_k(i) \tag{4.58}$$

Construcción de un banco de filtros

Existen diversos tipos de bancos de filtros pero el uso más extendido lo conforman los bancos de filtros triangulares. La construcción de estos últimos es sencilla a nivel computacional y no requiere de costes muy altos por lo que uso es altamente recomendable para aplicaciones en tiempo real. La construcción del banco de filtros se implementa de forma sencilla en varios pasos:

1. Supongamos que deseamos construir N filtros distribuidos en el rango de frecuencias $[0, F_s/2]$. Partiendo de la equivalencia entre Hz y Mels, distribuimos la escala Mel en un conjunto de puntos M_i equiespaciados, siendo i=N+2.

2. Seguidamente utilizando una operación inversa convertimos cada una de las frecuencias dadas en Mels, a una escala natural dada en Hz, obteniendo el conjunto de frecuencias F_i . Nótese que en este caso $F_1 = 0$ y $F_{N+2} = F_s/2$.

3. Llegados a este punto necesitamos conocer la resolución espectral de cada una de las tramas, es decir el numero de muestras, N_{fft}, que conforman la distribución de frecuencias entre $[0, F_s/2]$. Realizamos entonces la conversión para obtener la posición de la frecuencia central de cada filtro dentro del conjunto de muestras, conjunto P_i:

$$P_i = \left\lfloor (N_{fft} + 1) \frac{F_i}{F_s} \right\rfloor$$

4. Por último solo nos quedará construir el conjunto de banco de filtros, H_k, utilizando la ecuación (4.59).

$$H_k(n) \begin{cases} 0 & n < P_{i-1} \\ \dfrac{n - P_{i-1}}{P_i - P_{i-1}} & P_{i-1} \leq n \leq P_i \\ \dfrac{P_{m+1} - n}{P_{i+1} - P_i} & P_i \leq n \leq P_{i+1} \\ 0 & n > P_{i+1} \end{cases} \qquad k = [1, N] \tag{4.59}$$

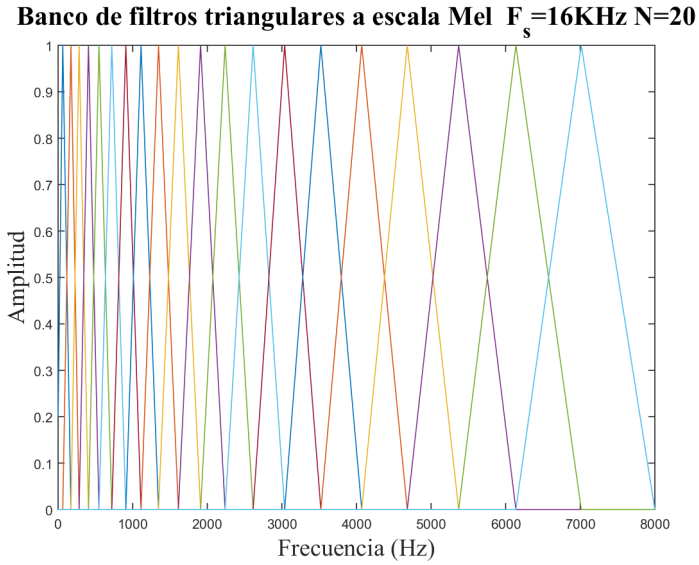

Figure 4.44: Banco de filtros triangulares a escala Mel

Coeficientes cepstrales a escala Mel

Uno de los parámetros de uso más extendido en la síntesis de voz son los coeficientes cepstrales a escala Mel (MFCC mel-frequency cepstral coefficient).

Se basan fundamentalmente en el uso de banco de filtros a escala Mel y una transformación al dominio cepstral pero haciendo uso de la transformada DCT (Discrete Cosine Transform). El proceso es sencillo:

1. Previamente se contruye el espectro de frecuencias de un fragmento de señal y obtenemos la magnitud del msmo.

$$\left|X_n\left(e^{j\omega}\right)\right|^2 = \left|\mathcal{F}\left[x\left(n\right)\right]\right|^2$$

2. Se realiza la estimación energética de la señal para cada una de las frecuencias de la señal utilizando el banco de filtros, generalmente 20-30 filtros con una distribución de frecuencias centrales a escala Mel.

$$X_k\left(k\right) = \sum_{i=f_{ini}}^{f_{ini}+B} \left|X_n\left(i\right)\right|^2 H_k\left(i\right)$$

3. Finalmente, realizamos la transformación al dominio cepstral haciendo uso de la trasnformada discreta coseno (DCT) que requierá únicamente la parte real para ser calculada.

$$c_k = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi k}{N}\left(j - 0.5\right)\right) \tag{4.60}$$

Comúnmente se ignoran los coeficientes cepstrales de órdenes elevados extendiéndose el uso de 12 coeficientes MFFC. Estos últimos presenta una característica ampliamente usada en el reconocimiento automático del discurso o el locutor.

4.4.4 Espectrogramas

Un espectrograma consiste en la representación gráfica del espectro de frecuencias. Se puede interpretar como una gráfica tridimensional que representa la energía del contenido en frecuencia de una señal temporalmente. Hasta ahora realizábamos un análisis espectral para visualizar el contenido en frecuencia de un determinado fragmento de una señal acústica en un determinado periodo temporal. La señal era segmentada en tramas que podrían o no estar solapadas de forma que coexistía cierta correlación entre tramas sucesivas. Si aplicásemos la FFT a cada una de las tramas obtendríamos el contenido en frecuencia de la señal para distintos periodos temporales, por lo que para construir el espectrograma únicamente tendríamos que concatenar las distintas tramas, figura 4.45.

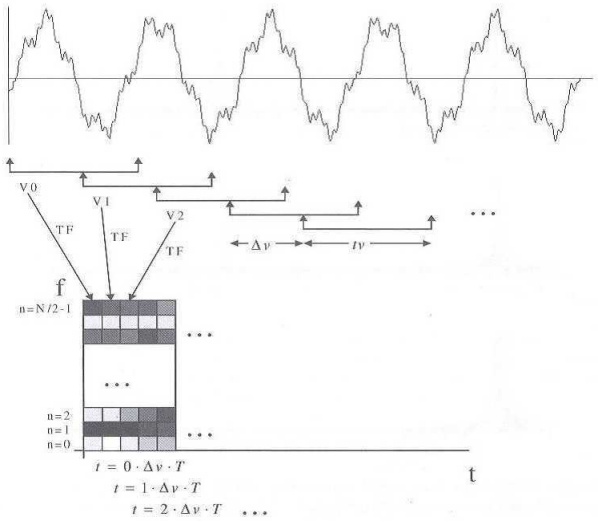

Figure 4.45: Contrucción de espectrogramas. Espectro variante temporalmente.

La representación de los resultados puede presentarse de dos formas:

- Gráficos tridimensionales: se utiliza una representación tridimensional para representar los valores energéticos de las componentes en frecuencia de cada una de las tramas.
- Gráficos unidimensionales utilizando un escalado de colores (colormaps) para la representación de las componentes en frecuencia de cada una de las tramas.

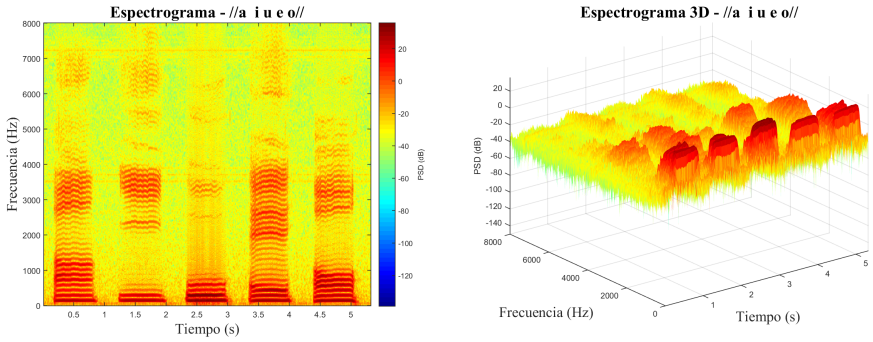

Figure 4.46: Representación unidemensional frente a representación multidimensional de un espectrograma.

Para agilizar los procesos de representación de resultados se ha optado por realizar un escalado de los resultados. La idea es distribuir todos los valores del espectrograma desde su mínimo y máximo local hacia una escala [0,1]. El escalado no afecta a los resultados, aplicando la ecuación (4.61) [31, 11] obtenemos los mismos resultados que con la escala inicial, figura 4.47 y 4.48.

$$S(i,j) = \frac{S(i,j) - S_{min}}{S_{max} - S_{min}} \tag{4.61}$$

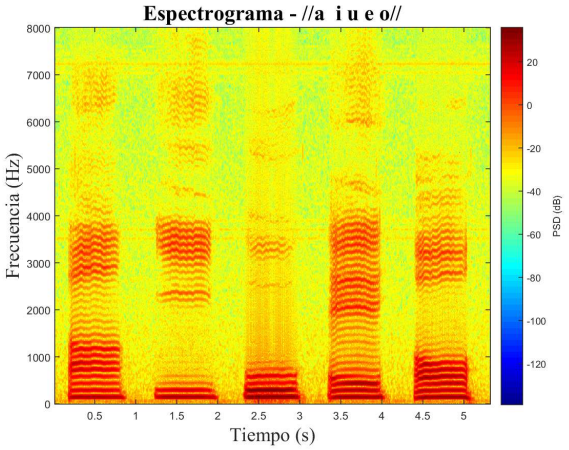

Figure 4.47: Espectrogram sin escalar.

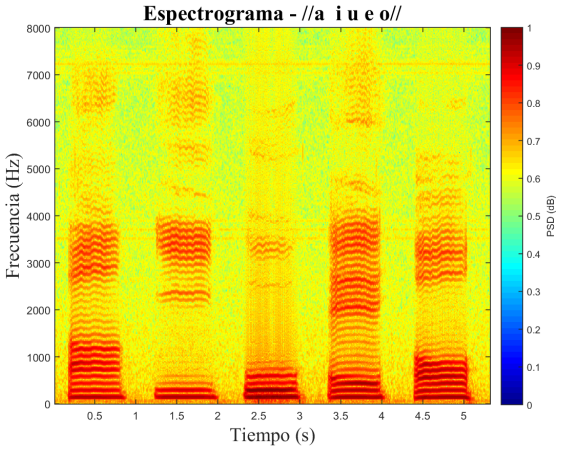

Figure 4.48: Espectrograma escalada a una escala [0-1]

Si observamos la imagen de la figura 4.48 observamos como el mínimo local se encuentra muy alejado de la media de los valores. Esto hace que en el proceso de escalado, componentes en frecuencia energéticamente insignificantes tomen valores relativamente elevados y compliquen en proceso de percepción visual. Para estos casos la solución recomendable es el uso de técnicas de comprensión/extensión de valores.

El problema radica en que las componentes en frecuencia de la señal del habla humana no presentan una distribución uniforme, por su propia naturaleza las componentes de baja frecuencia presentan una mayor importancia ya que son las más importantes para la percepción del habla [31].

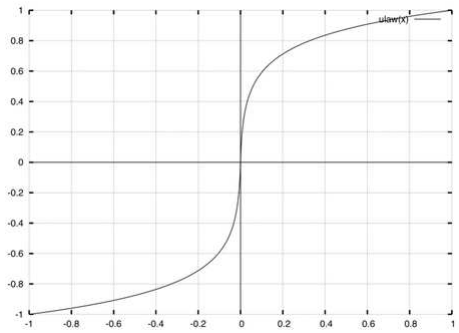

Figure 4.49: Escala utilizada para expansión/compresión en la Ley-μ [41, 31]

Lo ideal sería buscar técnicas que mejor se ajusten a la escala de audición humana por lo que deben de seguir una distribución logarítmica. Una de las técnicas más extendidas es la Ley Mu o Ley μ ($\mu - Law$) que explota el factor de que las componentes de alta frecuencia no necesitan tanta resolución como los bajos por lo que su uso es altamente recomendable.

$$S(i,j) = \frac{\log\left(1+\mu\left|S(i,j)\right|\right)}{\log\left(1+\mu\right)} \tag{4.62}$$

$$S(i,j) = \frac{1}{\mu}\left(e^{|S(i,j)|\log(1+\mu)} - 1\right) \tag{4.63}$$

La técnica es muy sencilla de aplicar y funciona tamto para compresión, ecuación(4.62), como expansión del escalado, ecuación (4.63) [31]. Podemos observar los efectos de cada uno de ellos en las figura 4.63 e 4.51 respectivamente.

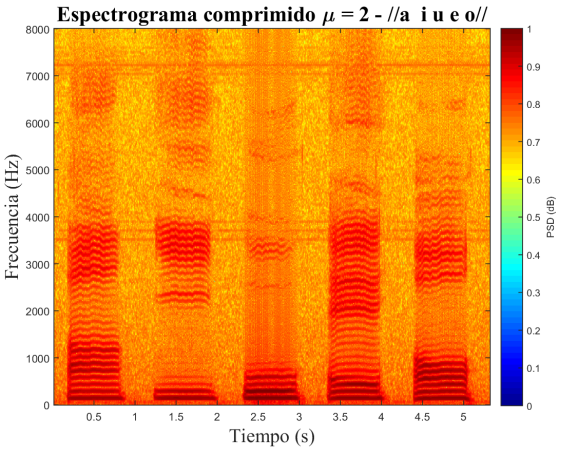

Figure 4.50: Espectro comprimido con $\mu = 2$

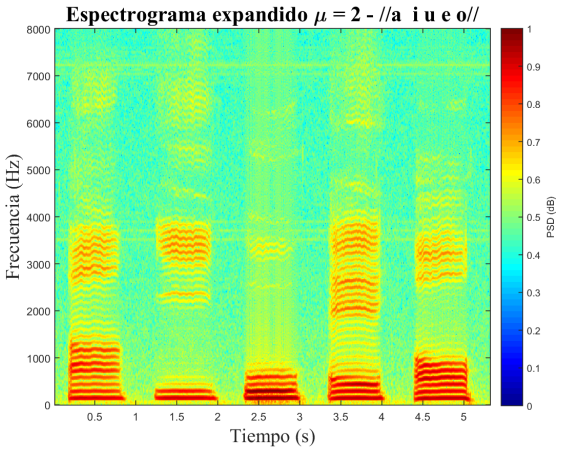

Figure 4.51: Espectro expandido con $\mu = 2$

La escala de colores es trasladada en función de la técnica que se le aplique. Para este ejemplo en concreto vemos como la compresión de datos satura la representación de la información por lo que es más trascendental utilizar técnicas de expansión. Con estas técnicas las componentes de interés de los segmentos de voz son realzados y a la vez las componentes del ruido son comprimidas. Si utilizamos un factor de expansión muy elevado podemos observar como las diferencias entre cada una de las tramas sonoras están mucho mejor diferenciadas, figura 4.52.

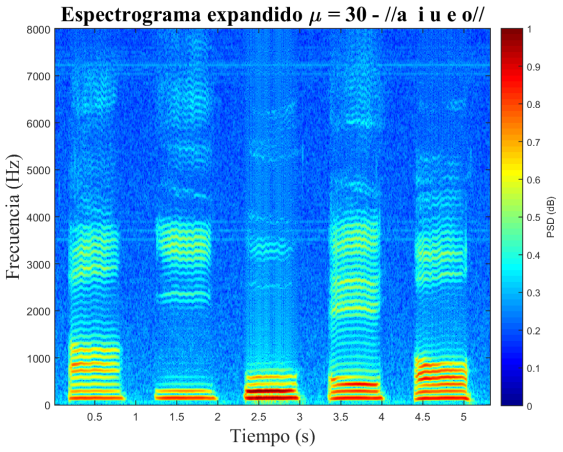

Figure 4.52: Espectro expandido con $\mu = 30$

Capítulo 5

Desarrollo de la herramienta

5.1 Planificación de la herramienta

La labor más importante para un desarrollo adecuado de un proyecto de software es la obtención y análisis de los requisitos y objetivos que se deben de alcanzar para de esta forma elegir la herramienta más adecuada para la elaboración del mismo. Se deben de definir las pautas a seguir recompilando, examinando y formulando todas las ventajas y desventajas para así determinar cualquier restricción previa. Con anterioridad se definieron cada uno de los objetivos que debía de cumplir la aplicación entre los que podemos destacar el hecho de que esta debía de ser libre, multiplataforma y multilenguaje. Uno de los factores principales de la aplicación era el soporte para el análisis fonético en tiempo real. La clave del procesamiento de señales en tiempo real sostenible radica en la rapidez y eficiencia del programa.

En el mercado existe muy amplia variedad de lenguajes de programación así como entornos de desarrollo, cada uno de ellos con distintos rendimientos en función del uso que se le dé. Cuando un desarrollador busca rapidez, eficiencia y transportabilidad es lógico pensar en un lenguaje de programación de bajo nivel o al menos próximo al mismo, y el lenguaje por excelencia que se presenta como primera alternativa es el lenguaje C.

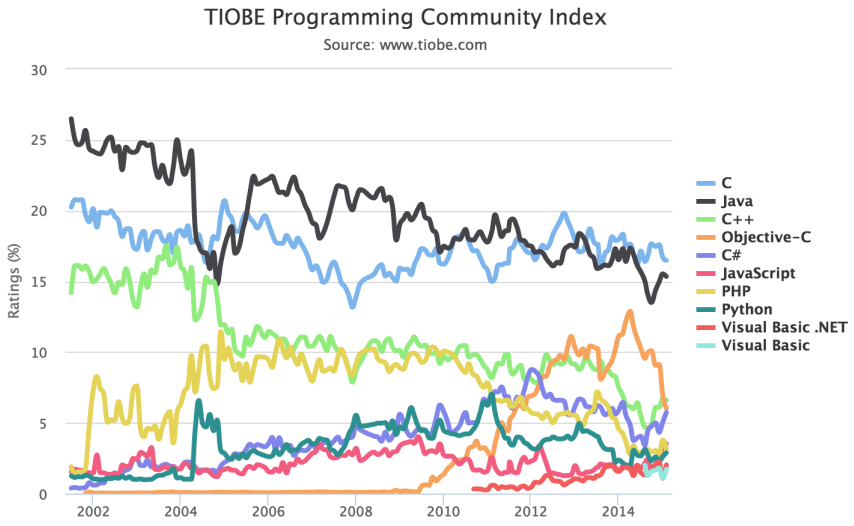

Figure 5.1: Evolución del uso de los lenguajes de programación en los últimos años

El lenguaje C fue diseñado para ser altamente transportable por lo que nos facilitaría la labor del desarrollo multiplataforma y presume de una eficiencia que pocos lenguajes pueden alcanzar, especialmente en procesos de análisis matemático. Es el lenguaje más extendido y coexisten una infinidad de librerías gratuitas que hacen más amena su integración en cualquier proyecto. El inconveniente principal radica en que este no permite la integración del paradigma de la programación orientada a objetos de forma nativa así como su ineficiencia para desarrollos con interfaz de usuario. Con el fin de construir una herramienta con interfaz de usuario (GUI) se opta por una combinación del mismo con su heredero, el lenguaje C++. Esta combinación acapara un gran nicho de mer-

cado en la comunidad de desarrolladores, figura 5.1, por lo que coexisten muchas comunidades de desarrolladores que dan soporte y actualizaciones.

Figure 5.2: Logo - Librería Qt.

La biblioteca gráfica por excelencia para el desarrollo en C/C++ es la librería Qt, por lo que fue la primera elección como entorno desarrollo. Antes de confirmar su uso, previamente se debe de realizar un análisis para analizar la integridad de la misma con los principales requisitos establecidos:

- *Multiplataforma:* Qt es una librería multiplataforma que permite su integración nativa con la mayoría de los sistemas operativos, incluidos sistemas embedidos y móviles.

- *Multilenguaje:* Qt da soporte para la integración de diferentes lenguajes mediante el entorno Qt Linguist.

- *Software Libre:* Qt dispone de distintas distribuciones tanto para uso comercial como libre. Dispone de un entorno de desarrollo para proyectos de software libre totalmente gratuito.

- *Uso en tiempo real:* Qt la herramienta permite la integración de proyectos en tiempo real cumpliendo con la mayoría de los estándares POSIX.

- *Recursos Reducidos:* Qt hace uso de las librerías nativas de cada sistema operativo para reconstruir su entorno gráfico por lo que el consumo de recursos es mínimo.

- *Exportación de resultados:* Qt incluye librerías nativas para exportar información en los formatos más comunes.

La herramienta cumple con cada uno de los requisitos y además incluye librerías nativas especialmente útiles para la elaboración del proyecto, algunas como:

- La librería Qt Multimedia: proporciona un amplio abanico de clases para el manejo de contenido multimedia, proporcionando APIs necesarias tanto para la reproducción como captura tanto de señales de audio como de video. Es una libreria gratuita soportada en las distintas distribuciones de Linux como en cada una de las versiones modernas de Windows y Mac.

- La librería Qt Chart: proporciona un conjunto de clases para la integridad de gráficos de análisis matemático. Es una librería sencilla y eficiente que presenta el inconveniente de que solo es distribuida en la versión comercial de Qt.

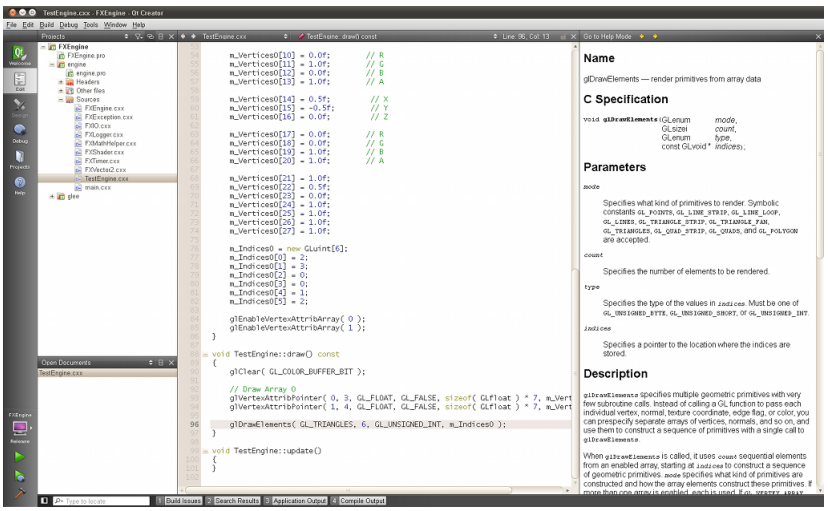

Figure 5.3: Entorno de desarrollo QtCreator

Este último factor implica que debemos de buscar una alternativa para la representación de gráficos de uso libre. La primera opción sería implementar nuestra propia librería basándonos en el propio framework gráfico de Qt, Qt Graphics View Framework. Esta opción era muy poco viable y alargaría demasiado el desarrollo del proyecto. Como solución alternativa se planteó el uso de la librería QCustomPlot, una librería de código abierto que permite una fácil integridad con la librería Qt. La librería presenta un rendimiento adecuado para una representación de gráficos 2D y sienta las bases para la implementación de gráficos 3D.

Figure 5.4: Logo - Libreria QCustomPlot

En lo que al procesamiento respecta para las operaciones elementales se utilizarían las propias librerías nativas del lenguaje. En el caso de la realización de las operaciones más complejas como la FFT se ha optado por la librería FFTW, también usada por el propio entorno de simulación utilizado, MATLAB (MATrix LABoratory). Esta librería está compuesta por un conjunto de subrutinas en C que permiten el cálculo de la Transformada de Fourier con un tamaño arbitrario y el manejo de datos tanto reales como complejos. Incluye además, rutinas para el cálculo de otras transformadas así como la de sus inversas..

Figure 5.5: Logo - Libreria FFTW

5.2 Implementación de la herramienta

Una vez elegido el entorno de desarrollo así como las librerías a utilizar se comienza con la planificación de la implementación de la herramienta. Durante este periodo se comienza a programar cada una de las líneas de código que conformaran nuestro programa. Se estructurará en diferentes bloques siendo cada uno de ellos testeado de forma independiente para detectar de forma prematura los diferentes errores. Las fases a seguir fueron:

- La integración del proceso de captura de datos.
- La integración del proceso de lectura de datos.
- La integración del proceso de representación visual de los datos.
- La integración de las librerías y métodos para el procesamiento de datos.
- Enfoque global y optimización de resultados.

Cada una de las etapas se realizaría en cascada apoyándonos en el sistema de control de versiones, GIT. La falta de personal humano lleva a la optimización del proyecto para un sistema operativo en cuestión para posteriormente adaptarlo para el resto de sistemas. La elección del sistema no es trivial y aunque es preferible la programación bajo entornos Unix, al final se optó por dar soporte al sistema operativo más utilizado en la actualidad, Windows 7, figura5.6.

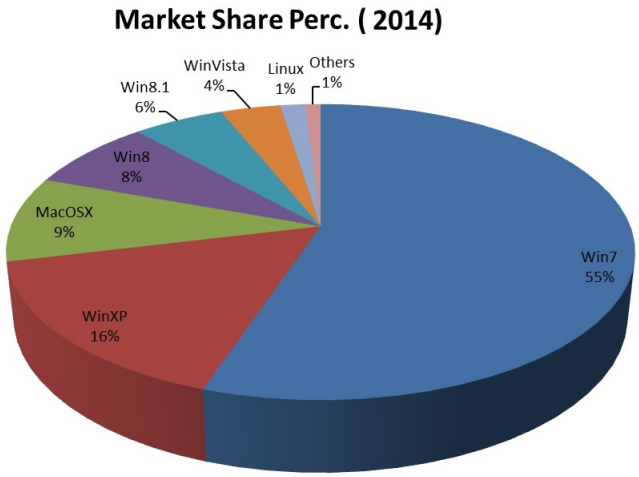

Figure 5.6: Sistemas operativos más usados

La fase más importante del proyecto implica el diseño e implementación de cada uno de los algoritmos utilizados para el proceso de extracción de parámetros de señales acústicas. Para dicha labor se ha implementado un conjunto de subrutinas en C, en una librería propia nacida con el nombre de LogoLibrary. Esta librería ha sido optimizada para la implementación de los siguientes procesos:

- Preprocesado de la señal: permite de forma sencilla el fragmentado, ventanado y preprocesado de un conjunto de datos. Los parámetros configurables son:
  - El tamaño de la ventana de anáñlisis así como el solapamiento entre ventanas sucesivas.
  - La selección de la ventana utilizada: Hamming, Hanning, Gauss...
  - Configuración del filtro de compensación DC
  - Configuración del filtro de realce de frecuencias
- Análisis temporal de la señal: permite la estimación de parámetros como la energía o la tasa de cruces por cero.
- Análisis frecuencial de la señal: permite el calculo de diferentes espectrogramas de la señal. Permite la selección del tamaño de la FFT y la selección de los diferentes métodos:
  - Espectro LPC: indicando el orden del mismo.
  - Espectro Cepstral: indicando la cantidad de muestras a utilizar para el filtrado paso baja.
  - Banco de filtros a escala Mel: indicando el número de filtros a utilizar.
- Detector de actividad sonora (VAD) configurable a distintos niveles de energía y tolerancia.
- Escalado y comprensión/expansión de espectrogramas.

La librería utiliza datos de doble precisión y se fundamenta en el uso de la librería FFTW así como de las propias nativas del estándar C/C++ y Qt para su correcto funcionamiento. Para garantizar la validez de los resultados se le añadió la capacidad de exportar los resultados y así poder realizar comparaciones con el entorno de simulación de MATLAB.

5.3 Despliegue y mantenimiento

Tras terminar una primera versión de la aplicación entramos en una fase de testeo y mejora de rendimiento. El software debía de ser testeado por diferentes usuarios para decidir su correcto funcionamiento y posibles mejoras. Para ello se diseñó un sencillo plan de entrenamiento y soporte que desglosaba distintas pruebas piloto, entre las que se pueden destacar:

- Representación de información correcta en tiempo real.
- Carga de archivos y representación de resultados en tiempo real.
- Funcionamiento de las distintas opciones configurables de la aplicación.

El software fue testeado en diferentes sistemas operativos nativos de Windows, y presentaba un correcto funcionamiento en las versiones:

- Windows XP, SP3

- Windows 7, versión 32 y 64 bits.
- Windows 8, versión 32 y 64 bits.
- Windows 10, versión 64 bits.

El programa respondía fluidamente salvo con cargas excesivas en el procesamiento por usuarios no experimentados. Esto implica que era necesario limitar el nivel de configuración ya que no tenía lógica alguna utilizar una resolución espectral de varios exponentes.

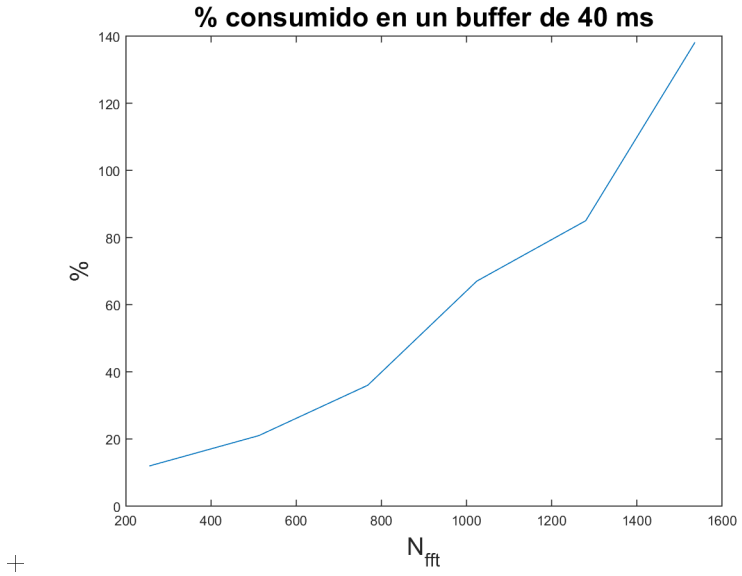

Figure 5.7: Tiempos de computo requerido para estimación de parametros

El principal cuello de botella en los tiempos de computo se encontraba en la representación del espectrograma. En este proceso los tiempos crecen exponencialmente con la resolución espectral utilizada, figura 5.7. Era necesario una adaptación dinámica de la representación del mismo y una modificación en la librería utilizada. Podemos observar la mejora de rendimiento lograda en la figura 5.8.

Optimizado el código se procede a intentar la compilación de la aplicación en distintas plataformas:

- Plataformas MAC: se requiere de la instalación de XCode para la compilación del programa. Es totalmente compatible con versiones de Mac OS X Lion y superiores. La aplicación presenta rendimientos limitados en plataformas MAC dado que no se disponía de ordenadores que soportasen esa plataforma para su optimización.
- Plaformas Linux: se requiere que la aplicación soporte la integración de la librería Qt Multimedia para su correcto funcionamiento. Es totalmente compatible con todas las distribuciones derivadas de Debian.

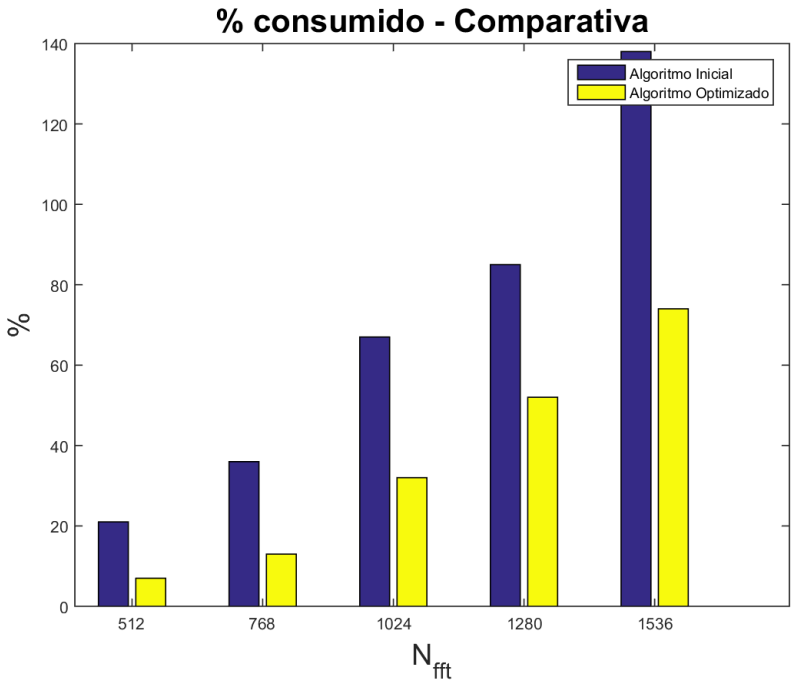

Figure 5.8: Mejora en el rendimiento del proceso de representación del espectrograma

Cumplidas todas estas pautas se compiló la primera versión release interna de la aplicación que nació con el nombre de LogoSpeech Studio, lista para su uso en la fase de experimentación del proyecto.

Capítulo 6

Manual de usuario

LogoSpeech Studio es una herramienta diseñada especialmente para la investigación fonética o fonológica de señales bioacústicas. Desarrollada para ser una herramienta multiplataforma, nace con la idea de dar apoyo en la docencia logopeda en aspectos como el reconocimiento del habla o la instrucción fonadora. Por ello se diseñó, íntegramente, buscando un funcionamiento simple y ameno para su correcto uso por usuarios menos experimentados. Esta sección pretende servir de guía para la iniciación en las tareas básicas a desarrollar con la misa.

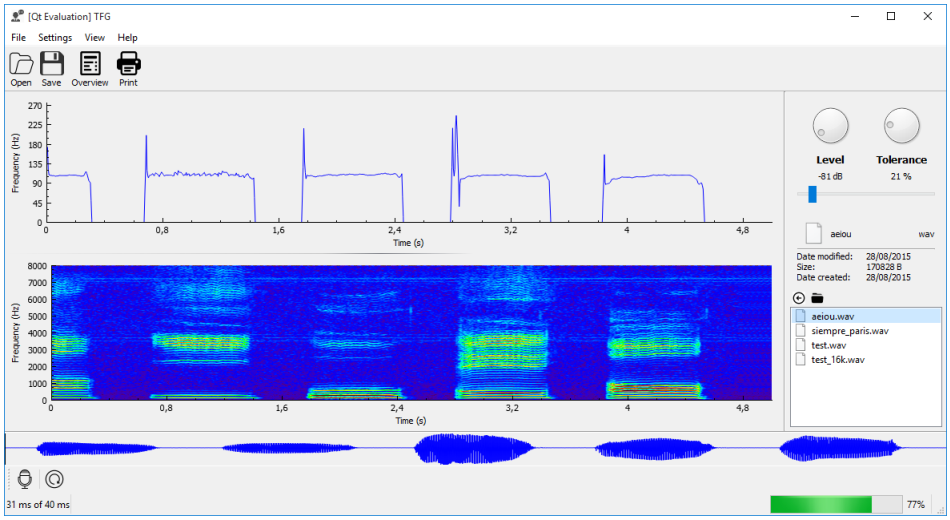

Figure 6.1: LogoSpeech Studio, entorno de usuario

En la Figura 6.1 podemos observar la interfaz principal de la aplicación. Se ha elegido una interfaz clásica y sencilla, con elementos de gran tamaño y fácilmente reconocibles para su rápida integración con el usuario. De esta forma se pueden distinguir principalmente 4 secciones de control, figura 6.2:

- Barra de control y conjuración: conforma el menú principal de la aplicación. A través del mismo tendremos un control total de la aplicación. Permite la realización de distintas operaciones y así como la configuración del proceso de extracción de parámetros tanto de la señal capturada como de la señal cargada desde un archivo.

- Representación de resultados: conforma el conjunto de gráficas que visualizan los resultados obtenidos una vez procesados los datos. Es totalmente redimensionable y se puede elegir que gráficas se desean mostrar a través de los paneles superiores. Así mismo permite la exportación de resultados a través de menús contextuales.

- Panel lateral: permite un rápido acceso al directorio actual para facilitar la carga de archivos. Así mismo incorpora un panel de control superior a través del cual se define el factor de comprensión/expansión del espectrograma así como los niveles de limitación y la tolerancia del detector de actividad sonora utilizado en el proceso de estimación de la frecuencia fundamental.

- Barra de captura de datos o reproducción de archivos: incluye unos controles para el inicio/pausa tanto del proceso de captura como de reproducción. Así mismo incluye una barra de progreso con el porcentaje de tiempo utilizado para el cálculo y visualización de parámetros en relación al tamaño del propio buffer de captura del dispositivo. Representa, además, de forma global el archivo de audio cargado, si lo hubiera.

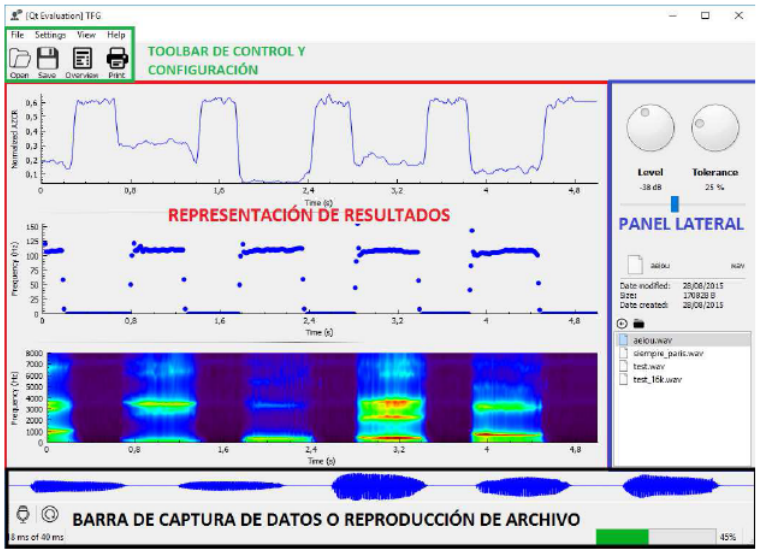

Figure 6.2: LogoSpeech Studio, secciones principales

6.1 Captura de datos

La captura de datos en esta aplicación se ha implementado para que sea fácilmente configurable. El usuario puede iniciar la captura al lanzar la aplicación a través del propio botón habilitado para esta labor, o atravesó de la barra principal. El proceso de captura es totalmente personalizable, el usuario puede decidir el dispositivo de captura deseado, el formato de salida del archivo capturado así como la frecuencia de muestreo o el codec utilizado. Estos parámetros son introducidos a través del dialogo de configuración del proceso de captura de datos, figura 6.3.

El usuario debe de guardar la conguración para reiniciar la captura y ser utilizada en futuras ocasiones. Así mismo siempre podrá volver a la conguraci ón por defecto de la propia aplicación.

6.2 Conguración de procesado

Esta es sin duda la sección más importante de la aplicación. Es la que controla todo el proceso de tratamiento previo y extracción de parámetros de la señal de entrada. Viene estructura en tres secciones bien diferenciadas:

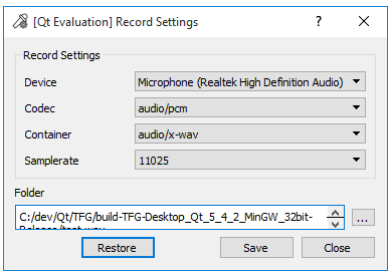

Figure 6.3: Dialogo para la configuración de captura de sonido

- Preprocesado: incluye la configuración de las técnicas de preprocesado desarrolladas. El usuario puede determinar si desea utilizar un filtro de compensación de tensión DC así como de realce de componentes de alta frecuencia. Por último y no por ello menos importante puede elegir la ventana que desea utilizar en el proceso de fragmentación de señal.

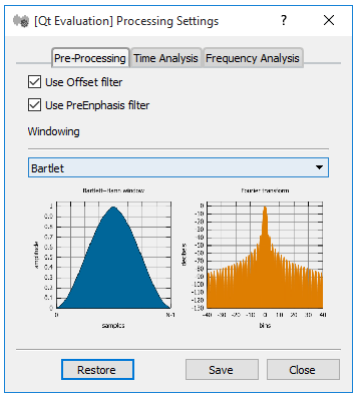

Figure 6.4: Dialogo para la configuración del preprocesado

- Análisis temporal: • incluye el dimensionado de las ventanas a utilizar así como el tamaño buffer global temporal que será visualizado para el usuario. Desde esta sección podremos configurar el tamaño de la ventana así como el tiempo de salto. Por otro lado permite la selección del parámetro a calcular en la gráfica de representación temporal (Energía o tasa de cruces por cero) así como la conjuración del método a utilizar para la estimación de la frecuencia fundamental así como el margen dinámico de la misma.
- Análisis de frecuencia: es el campo más importante. Permite la conguración de los distintos espectrogramas que pueden visualizarse. Permite definir el tamaño de las tranformadas así como la propia técnica a utilizar para su visualización. Actualmente permite la estimación de distintos campos: :
  - Espectro real FFT
  - Espectro LPC, indicando el orden del mismo.

– Espectro Cepstral, indicando el número de muestra a utilizar.

– Espectro por banco de ltros, indicando el número de ltros que serán distribuidos en el rango de frecuencias a escala Mel.

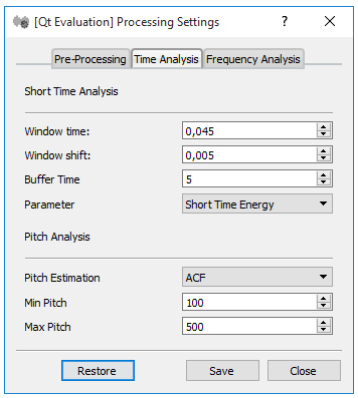

Figure 6.5: Dialogo para la conguración del análisis temporal

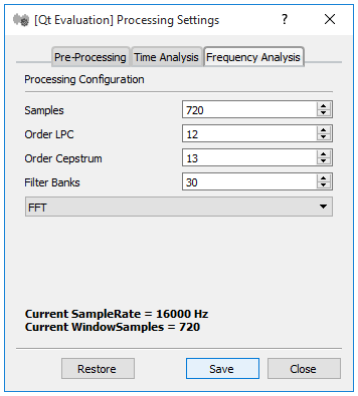

Figure 6.6: Dialogo para la conguración del análisis frecuencial

6.3 Conguración de gráficas

LogoSpeech Studio permite la configuración de los datos así como la forma en que estos se visualizarán. Para ello incorpora una sección dedicada a la personalización de las propias grácas. A través de la misma se pueden elegir:

- El uso de ejes verticales/horizontales para la representación de la escala en que se visualizan los datos.
- El uso de un grids..
- La selección del color así como del grosor de la línea que visualiza los datos.

- La elección de la escala de colores utilizada para la representación del espectrograma.
- La elección del factor de expansión/comprensión de los datos visualizados en el espectrograma.

Todo ello se integra en subsecciones fácilmente configurables. Cada una de ellas dedicada a cada una de las gráficas representadas. La capacidad de personalización es tal que nos permite seleccionar que gráficas deseamos visualizar así como un redimensionado arbitrario a gusto del propio usuario, figura 6.8.

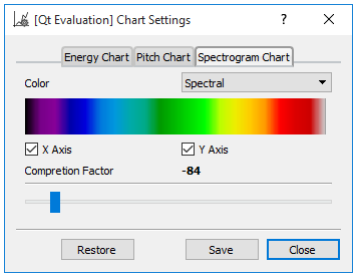

Figure 6.7: Dialogo para la conguración de la visualización del espectrograma

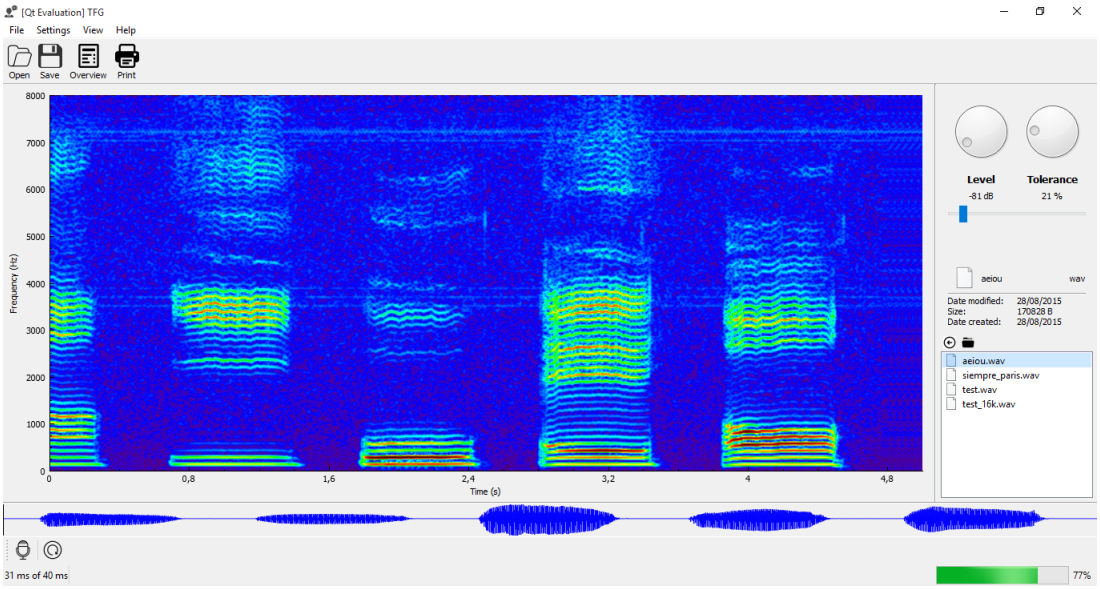

Figure 6.8: LogoSpeech Studio, visualización personalizada

Part III

Evaluación

Capítulo 7

Diseño Experimental

Tras el diseño, implementación y testeo de la herramienta se hace necesario una planificación de la estrategia con la que se desglosaran todos las pruebas empíricos sobre la misma. Dado que el proceso de aprendizaje es mucho más ágil en edades tempranas [10] se ha decidido diseñar un sistema de familiarización con la aplicación previa con el fin de agilizar la captación de la información mostrada para individuos de temprana edad.

La mejor forma de acceder y percibir el lenguaje oral es mediante el sistema auditivo. Cuando se carece de un correcto funcionamiento del mismo se deben de buscar alternativas que satisfagan esa necesidad. La idea que desglosa este proyecto es trasladar parte o la totalidad de estas percepciones al campo visual a través de espectrogramas y operaciones derivadas. La asimilación de estas capacidades puede ser lenta y debe de planificarse adecuadamente. Entre los objetivos que se deben satisfacer podemos destacar:

- Facilitar el proceso de rehabilitación auditiva para casos con implantes cocleares o similares.
- Incrementar las capacidades durante el proceso de aprendizaje lingüístico de un idioma.
- Agilizar el aprendizaje fonético de individuos con discapacidades auditivas.

Finalmente se ha optado por diseñar el programa fundamentándose en las estructuras de rehabilitación de la doctrina logopeda. De esta manera se ha estructura en cuatro fases bien diferenciadas:

- Detección
- Discriminación
- Identificación
- Reconocimiento y comprensión

7.1 Detección

En esta fase se pretende probar las capacidades auditivas del usuario factor fundamental para las fases posteriores ya que demuestra si el individuo indica presencia o ausencia parcial o total del sonido. Para el desarrollo de la fase se plantean diferentes ejercicios en los que se emitirán distintos sonidos (voz, instrumentos, ambiente...). El usuario deberá de indicar mediante gestos si percibe el sonido o es capaz de reconocerlo. Entre los ejercicios propuestos podemos destacar:

- Reconocimiento de vocales aisladas.
- Reconocimiento de sonidos del propio cuerpo
- Sonidos del entorno
- Sonidos musicales sencillos

7.2 Discriminación

Una vez determinadas las capacidades auditivas del individuo es necesario hacer un análisis del comportamiento de estas últimas ante diferentes fragmentos sonoros. La idea es buscar si el usuario tiende a confundir o a no distinguir sonidos diferenciados. Esta fase es relativamente importante ya que acentúa las capacidades para reconocer aspectos como la duración, melodía, intensidad o timbre de señales acústicas. Antes de iniciar el programa es necesario hacer entender al individuo el concepto de diferencia así como el gesto a realizar para identificarlo[17, 10]. Una vez comprendido, algunas de las actividades a realizar serán:

- Discriminación de instrumentos musicales simples
- Discriminación de cualidades del sonido: sonidos con elevada potencia o viceversa.
- Discriminación de vocales aisladas
- Discriminación de vocales en silabas
- Discriminación de palabras cortas.

7.3 Identificación

Esta etapa identifica las capacidades que presenta el individuo para el reconocimiento del habla. Se trata de introducir a diferentes escalas sonidos cada vez más complejos y observar las capacidades de reconocimiento. Comenzaremos con sonidos aislados y derivaremos paulatinamente hacia niveles más complejos (silabas, palabras y frases finalmente). Se debe de haber adquirido los conocimientos de la fase anterior correctamente, por lo que si fuera necesario se debe de realizar una repetición de la misma con el fin de posibilitar una posterior discriminación eficiente de logotomas[17]. Entre los ejercicios propuestos encontramos:

- Identificación de intrumentos musicales simples.
- Identificación de sonidos del entorno
- Identificación de fonemas vocálicos aislados.
- Identificación de silabas aisladas
- Identificación de palabras y de sonidos propios de la voz humana.
- Identificación de frases: etapa más compleja de todo el proceso.

7.4 Reconocimiento y comprensión

En esta etapa el individuo debe de asimilar la información del sonido una vez adquirido las capacidades previas. Para su reconocimiento, se debe de haber aprendido a extraer la información subyacente en el sonido. Los resultados de esta prueba marcan los logros del experimento, por lo que debe de realizarse con rigurosidad.

La naturaleza del programa es cíclica, se deben de realizar todas las pruebas con niveles de dificultad añadidos (ruidos de ambiente, múltiples conversaciones...). El individuo debe de tender a un proceso de mejora de vocalización y por tanto en una derivación contextual más natural del lenguaje hablado .

Capítulo 8

Pruebas de campo

La evolución de esta herramienta va ligada al desarrollo de estas pruebas y a los posibles requisitos demandados por los usuarios que las realicen. Se ha habilitado un foro de sugerencias y quejas con el fin de mejorar sus funcionalidades en próximas versiones. A lo largo de estas pruebas se han ido interviniendo en cada una de sus posibles configuraciones para poder implementar un uso más sencillo de la misma y que no requiera de conocimientos expertos.

Las pruebas realizadas se han desarrollado sobre dos plataformas, Windows 8 así como con el propio sistema operativo desarrollado por la Junta de Andalucía para ámbitos educativos, Guadalinex, 8.1. Se eligió este como segunda opción ya que es liberado y distribuido por todos los centros educativos de la comunidad Andaluza. La realización del programa se ha llevado a cabo en cada una de las fases preestablecidas en el programa a diferentes escalas de complejidad. Previo al inicio se realiza una pequeña introducción sobre la usabilidad de la aplicación con el fin de que cada usuario intente utilizar y configurar de manera personalizada cada una de las opciones integradas.

Figure 8.1: Logo: Guadalinex, distribución de Linux promovida por la Junta de Andalucía

Las pruebas se realizaron con distintos voluntarios de diferentes sexos y edades (algunas incluso con diagnóstico de deficiencia auditiva detectado). La mayoría de la población encuestada eran jóvenes, hecho que beneficiaba el rendimiento de las pruebas ya que es a esa edad cuando las capacidades de aprendizaje son más elevadas. En total fueron encuestados 47 personas de las cuales 32 fueron hombres y 15 mujeres

8.1 Pruebas de detección

Con esta fase se pretende demostrar si el individuo percibe presencia o ausencia parcial o total de sonido. La herramienta permite la reproducción de sonido así como el control de la potencia emitida. Para todos los casos se ha utilizado una biblioteca común de archivos de sonido de 16KHz de frecuencia de muestreo con diferentes sonidos y se han clasificado en varios grupos:

1. Individuos sin problemas auditivos: cumplieron la totalidad de la prueba a diferentes niveles. Estos fueron seleccionados para el resto de las pruebas limitando sus capacidades auditivas.

2. Individuos con hipoacusia leve: aquellos que han completado al menos el 75% de la prueba.

3. Individuos con hipoacusia media: aquellos que han completado al menos el 50% de la prueba.

4. Individuos con hipoacusia severa: aquellos que no alcanzaron ninguna de las pruebas anteriores.

En un ambiente aislado y con poco ruido ambiental cada uno de los individuos es sometido a la percepción de 10 sonidos concretos a diferentes escalas de potencia, figura 8.2. Como el oído humano percibe con mayor facilidad las componentes de baja frecuencia se ha optado por el uso de notas musicales a diferentes escalas para reconocer la capacidad auditiva a diferentes márgenes frecuenciales.

	Bajo (20-40dB)	Medio (40-70dB)	Alto (70-90dB)
Vocal /a/ (150 Hz)	S/N	S/N	S/N
Vocal /e/ (150 Hz)	S/N	S/N	S/N
Vocal /o/ (150 Hz)	S/N	S/N	S/N
Consonante /s/	S/N	S/N	S/N
Timbre (433 Hz)	S/N	S/N	S/N
Bocina (600Hz)	S/N	S/N	S/N
Do (130 Hz)	S/N	S/N	S/N
Do (16.47 KHz)	S/N	S/N	S/N
Mi (164 Hz)	S/N	S/N	S/N
Mi (10.5KHz)	S/N	S/N	S/N

Figure 8.2: Estructura de la fase de detección

Los resultados obtenidos fueron los mostrados en las Tablas 8.1 y 8.2, en conjunto son mostrados en la figura 8.5. La población general encuestada presenta unas capacidades auditivas del margen de la normalidad. Estos resultados coinciden gratamente con la estadística nacional ya que en España apenas conviven dos millones de personas con alguna hipoacusia. La detección de algún nivel de hipoacusia (son meros niveles clasificatorios sin ningún diagnostico medico) es más propensa a edades más avanzadas, esencialmente en la población masculina. Para dar validez a los resultados los archivos utilizados serían reproducidos en ausencia o limitación de sonido para de estar forma simular una hipoacusia severa.

	Encuestados	Nada	H. Leve	H. Media	H. Severa
Niños (5-16 años)	4	3	1	0	0
Jovenes (17-30)	20	17	2	1	0
Mediana Edad (30-65)	5	3	1	0	1
Avanzada Edad (65 o más)	3	0	1	2	0
Global - Masculino	32	72%	16%	9%	3%

Table 8.1: Resultados de la fase de detección. - Sexo Masculino

8.2 Pruebas de discriminación

En esta fase se comenzó a utilizar la naturaleza propia para la que fue diseñada LogoSpeech Studio, el análisis de señales acústicas. La idea de esta fase es la observación del comportamiento de estas personas ante diferentes sonidos anulando sus capacidades auditivas. Durante este proceso se cargaron diferentes

	Encuestados	Nada	H. Leve	H. Media	H. Severa
Niños (5-16 años)	2	2	0	0	0
Jovenes (17-30)	11	10	1	0	0
Mediana Edad (30-65)	1	0	1	0	0
Avanzada Edad (65 o más)	1	0	0	0	1
Global - Femenino	15	80%	13%	0%	7%

Table 8.2: Resultados de la fase de detección. - Sexo Femenino

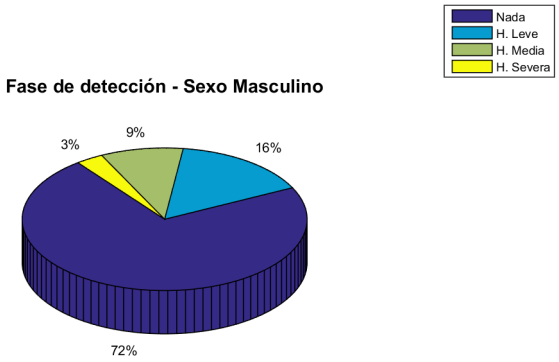

Figure 8.3: Resultados de la fase de detección - Sexo Masculino

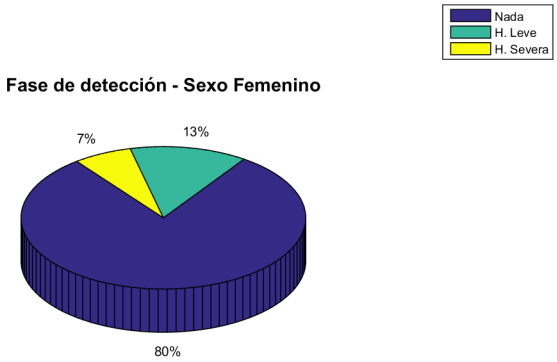

Figure 8.4: Resultados de la fase de detección - Sexo Masculino

archivos con fragmentos sonoros indicando mediante imágenes la correspondencia de los mismos.

Se les explicó de manera simple y compacta la información importante sobre cada sonido: timbre, frecuencia fundamental, componentes en frecuencia así como la distinción de niveles de potencia sonora. La idea es que con un entrenamiento base el individuo comience a ser capaz de distinguir entre sonidos sordos/sonoros o entre sonidos emitidos por la voz humana o cualquier instrumento. Para ello se tomó como punto de partida la base de datos de la sección anterior ya que el individuo se encontraba familiarizado con la misma y se le planteo la posibilidad de discriminación haciendo uso de diferentes niveles:

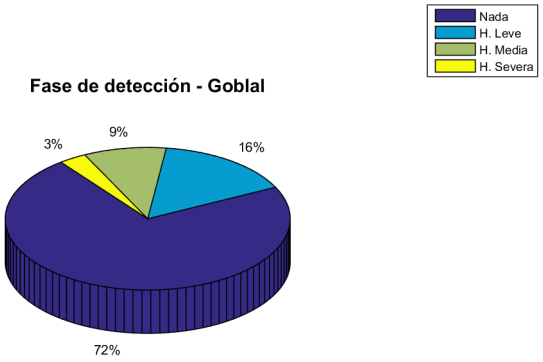

Figure 8.5: Resultados globales de la fase de detección

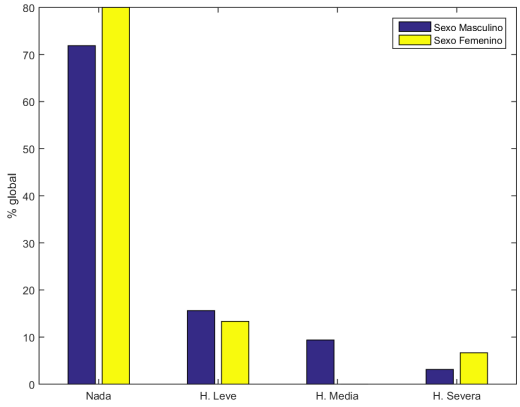

Figure 8.6: Resultados globales de la fase de detección- Sexo Masculino vs Femenino

- Usando la representación energética del sonido.
- Usando la representación de la frecuencia fundamental del sonido.
- Usando la representación del espectrograma del sonido con diferentes configuraciones.
- Usando una combinación de cualquiera de los anteriores.

En cada una de estas, el usuario evaluaría el nivel de complejidad que encontró para distinguir los parámetros. Así mismo se presentó la opción de destacar la información relevante que era capaz de interpretar a través de los resultados mostrados en la aplicación. La mayoría de los voluntarios carecían de conocimiento alguno sobre procesado de señales de audio o sus características fundamentales así que la información extraída se limitaba esencialmente:

- Distinguir el nivel energético del fragmento para saber le tono en que se emite.

- Distinguir entre las componentes de bajas y altas frecuencias.
- Distinguir, en el mejor de los casos, la frecuencia fundamental.
- Distinguir entre sonidos sordos o sonoros.

A continuación se presentará un ejemplo al lector de una de las pruebas a las que fueron sometidos algunos de los voluntarios, el fragmento sonoro en cuestión fue la vocal /a/. Aunque el software permite el cálculo de un mayor número de parámetros solo se adjuntan alguno de los más relevantes.

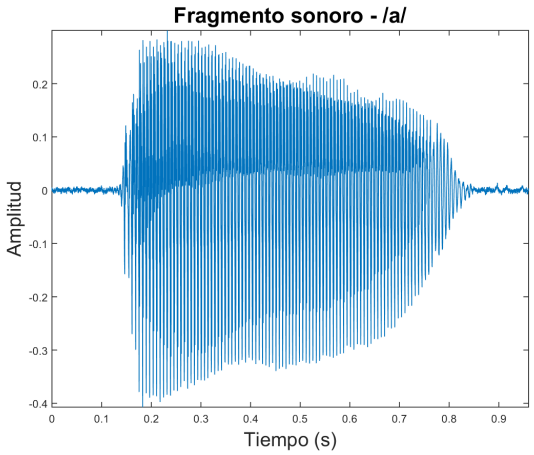

Figure 8.7: Fragmento sonoro usado en la fase de discriminación

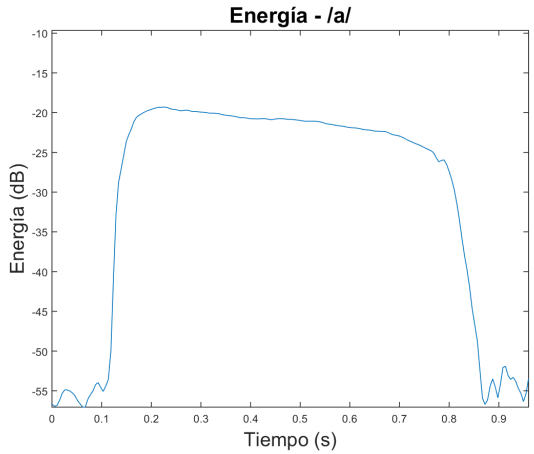

Figure 8.8: Energía en dB. Vocal /a/

Visualmente podemos distinguir unas características concretas del fragmento. En la figura 8.9 observamos como la frecuencia fundamental, f 0 , permanece constante durante el fragmento, por lo que podemos distinguir que el fragmento es sonoro. Por otro lado las características energéticas implican un nivel normalizado adecuado, entorno al nivel en que sitúa una conversación normal en

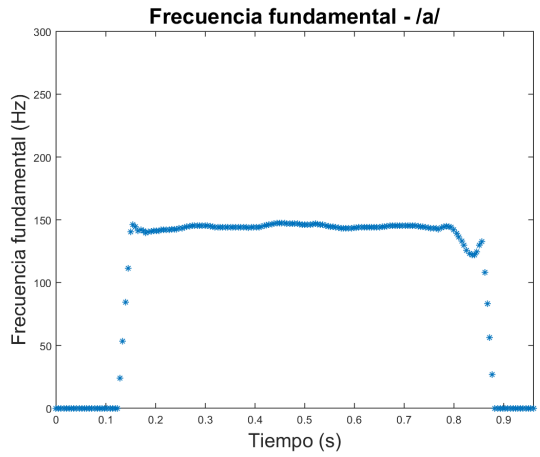

Figure 8.9: Frecuencia Fundamental (150 Hz) - Vocal /a/

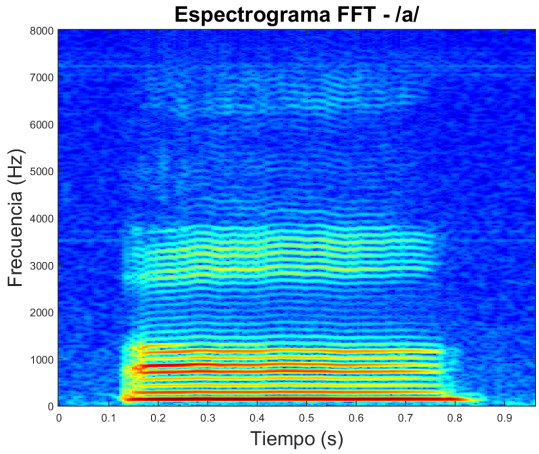

Figure 8.10: Espectrograma FFT - Vocal /a/

un ambiente poco ruidoso. El espectrograma es quizás el campo que mayor información transfiere por ello se le aplica un tratamiento previo de normalización y compresión/expansión. A través del mismo podemos distinguir las diferentes componentes en frecuencia:

- Bajas frecuencias: el habla humana tiende a transmitir toda la información a bajas frecuencias ya que esta es perceptualmente más sensitiva. Si analizamos la gráfica de la figura 8.10 podemos observar cámo se atisba una linea constante con un alto contenido en frecuencia entorno a los 150Hz, esta componente se corresponde con la frecuencia fundamental del hablante. Seguidamente podemos distinguir un conjunto de N armónicos entorno a la frecuencia nf_0siendo $1 < n \leq N + 1$.

- Altas frecuencias: los sonidos sordos tienden a presentar componentes de alta frecuencia muy marcados y a penas presentan energía a bajas frecuen-

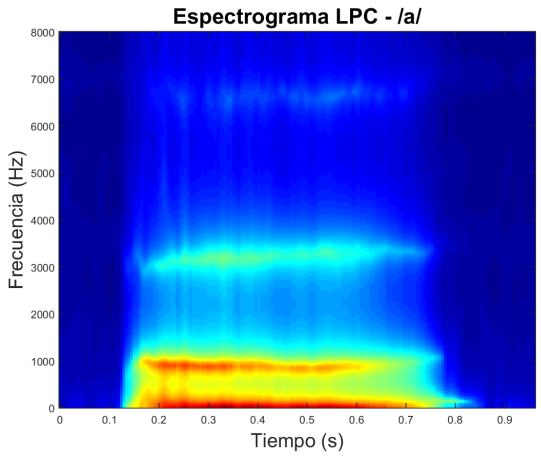

Figure 8.11: Espectrograma LPC - Vocal /a/

cias, véase uno caso de los casos expuestos en esta fase en la figura 8.12. La consonante /s/ presenta un espectro totalmente: para altas frecuencias el espectro presenta componentes energéticas elevadas bien diferenciadas del resto, frente al caso vocálico, figura 8.10, donde podemos observar un espectro a altas frecuencias prácticamente plano y que presenta niveles de energía despreciables.

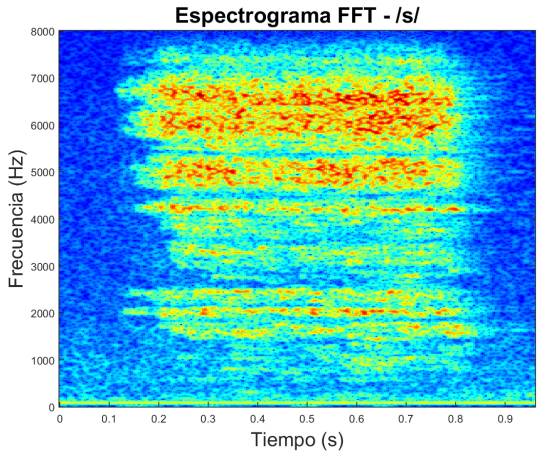

Figure 8.12: Espectrograma FFT - Consonante /s/

Un factor interesante es el uso de diferentes espectrogramas. El espectro LPC de la figura 8.11 atenúa las componentes de alta frecuencia pero realza las de baja. En algunos casos se ha demostrado este tipo de espectrogramas junto al cepstrograma presentan muy buenos resultados. Ha habido cierta tendencia a utilizarlos como elementos de cálculo por defecto, ya que aunque el espectro FFT transfiere más información en su mayoría es inasumible y mucho menos en tiempo real. Los resultados encontrados son los mostrados en las Tablas 8.3 y .

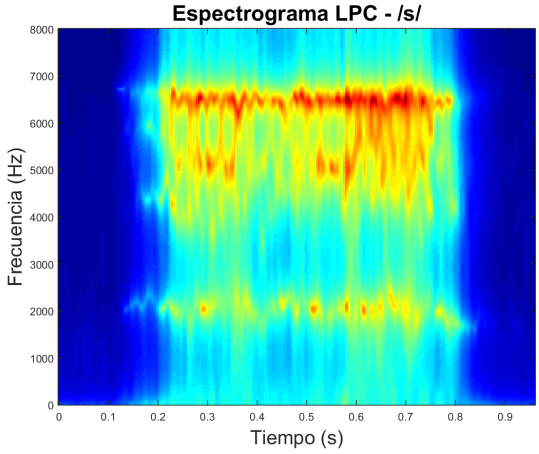

Figure 8.13: Espectrograma LPC - Consonante /s/

Distinción\Herramienta	Energía	F. Fundamental	Espectro	Combinación
Tono	56%	1%	12%	31%
Baja/Alta Frecuencia	2%	13%	79%	6%
F. Fundamental	1%	91%	3%	5%
Sonoro/Sordo	2%	13%	19%	66%
Global	15.25%	29.5%	28.25%	27%

Table 8.3: Resultados de la fase de discriminación - Sexo Masculino

Distinción\Herramienta	Energía	F. Fundamental	Espectro	Combinación
Tono	62%	1%	16%	21%
Baja/Alta Frecuencia	3%	7%	86%	4%
F. Fundamental	1%	87%	6%	6%
Sonoro/Sordo	1%	17%	14%	68%
Global	16.75%	28%	30.75%	24.75%

Table 8.4: Resultados de la fase de discriminación - Sexo Femenino

A nivel global, figura 8.16, podemos determinar que los algoritmos elegidos por la mayoría de los usuarios son el uso del espectrograma o la estimación de la frecuencia fundamental. Una cosa interesante es que gran parte preferían una combinación de los algoritmos pero acentuaban el hecho de que la estimación de la energía solo aportaba información sobre el tono de voz. La idea entonces es permitir que el usuario elija que combinación de algoritmos desea utilizar, así como la importancia que le da a cada uno de ellos asignándoles una mayor proporción de pantalla.

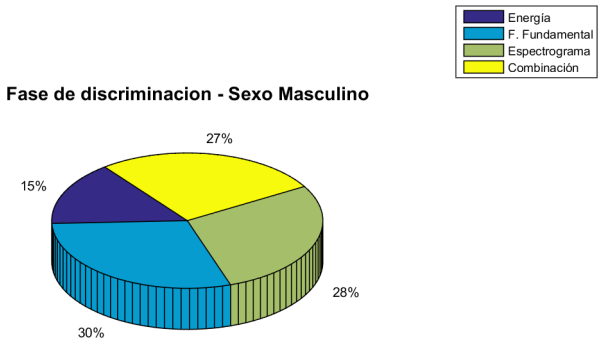

Figure 8.14: Resultados de la fase de discriminación, métodos más utilizados - Sexo Masculino

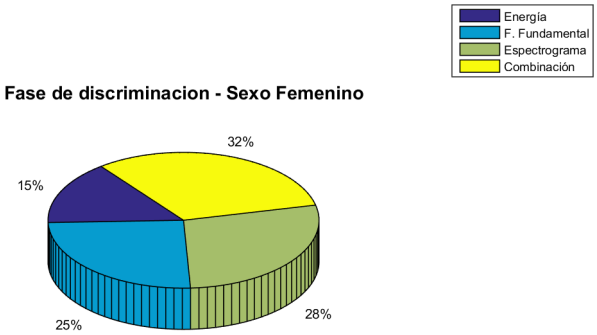

Figure 8.15: Resultados de la fase de discriminación, método utilizados - Sexo Masculino

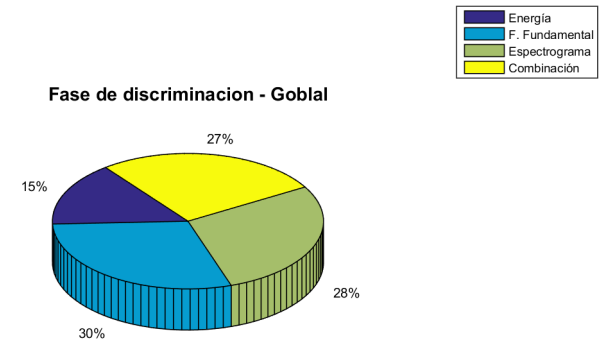

Figure 8.16: Resultados globales de la fase de discriminación

8.3 Pruebas de identificación, reconocimiento y compresión.

Se ha optado por agrupar los tres últimos niveles del programa diseñado para agilizar el proceso de desarrollo de la prueba debido a la fuerte correlación

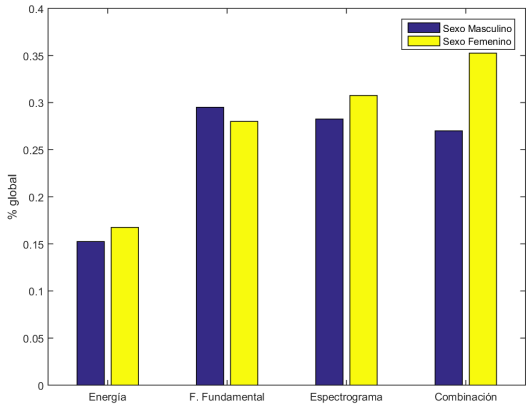

Figure 8.17: Resultados globales de la fase de discriminación- Sexo Masculino vs Femenino

existente entre cada una de las fases. Esta etapa marca las capacidades que presenta el individuo para el reconocimiento del habla una vez adquiridas las competencias de las fases anteriores. Es la etapa más compleja y de más lento absorción, la cual determina si se cumplen todos los objetivos para los que se diseñó la herramienta. La prueba se ha dividido en diferentes niveles de complejidad con el fin de realizar un análisis de cada uno de ellos de forma independiente.

- Identificación de sonidos aislados: vocales, consonantes, notas musicales...
- Identificación de palabras y silabas aisladas.
- Identificación de frases simples.

8.3.1 Identificación de sonidos aislados

Para esta fase se presentaron dos combinaciones de sonidos vocálicos de la base de datos utilizada anteriormente:

- Tren de vocales: /a e o/
- Tren de vocales y consonantes / a s e /

La representación espectral de algunas de las cadenas están disponibles en las figuras 8.18, 8.19, 8.20 y 8.21. Si observamos los espectros podemos enfrentrarnos a la principal complejidad del problema planteado y es la tendencia a confundir sonidos. Los espectros de las vocales /a/ y /o/ presentan comportamientos muy similares, son difícilmente diferenciables por lo que existía una tendencia a confundirlos.

Lo vocal /e/, sin embargo, contiene componentes a altas frecuencias con mayor distribución energética, hecho que facilita su identificación. La única componente que no presenta duda alguna es la componente /s/. La consonante,

al ser sorda, presente una distribución energética muy acentuada a altas frecuencias y un espectro plano a bajas frecuencias por lo que su identificación es muy sencilla.

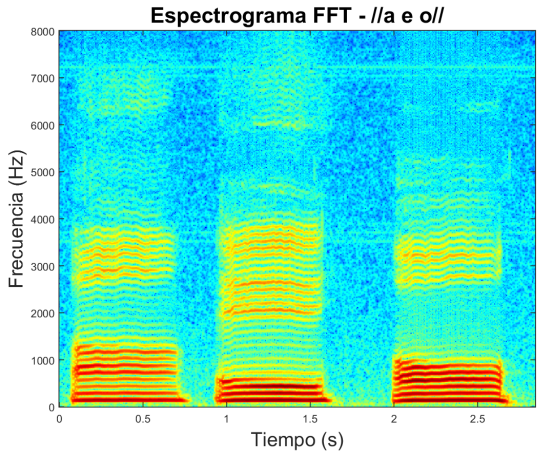

Figure 8.18: Espectrograma FFT - Cadena /a e o/

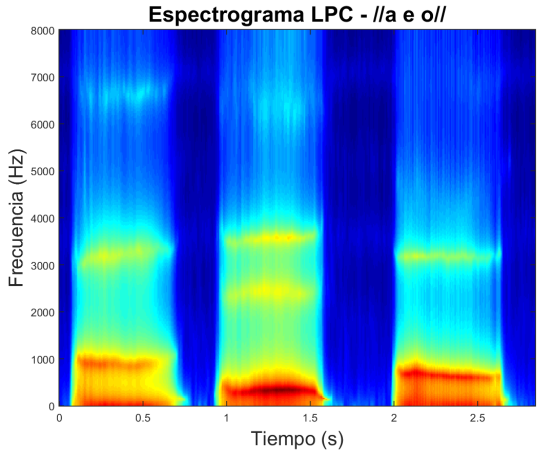

Figure 8.19: Espectrograma LPC - Cadena /a e o/

Los resultados obtenidos, figuras 8.22 y 8.23, son complejos de analizar pero la mayoría de los casos presentaban porcentajes de acierto elevados. El aspecto más importante es que los sujetos con apenas práctica han logrado identificar con éxito algunos sonidos, especialmente los sonidos musicales que presentaban componentes en frecuencia constantes o sonidos sordos aislados en un entorno vocálico como la consonante /s/. El compromiso entre horas requeridas para el aprendizaje y resultados obtenidos es plenamente satisfactorio en esta etapa.

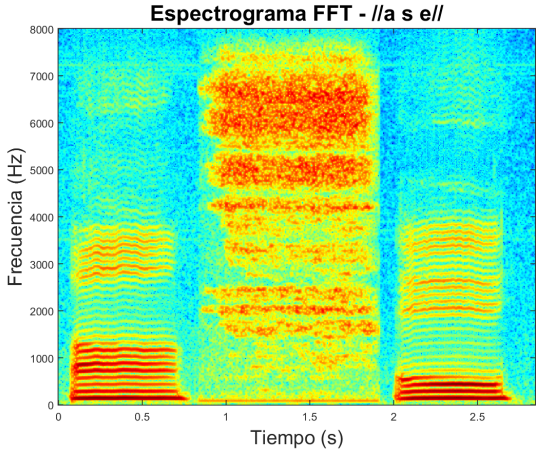

Figure 8.20: Espectrograma FFT - Cadena /a e o/

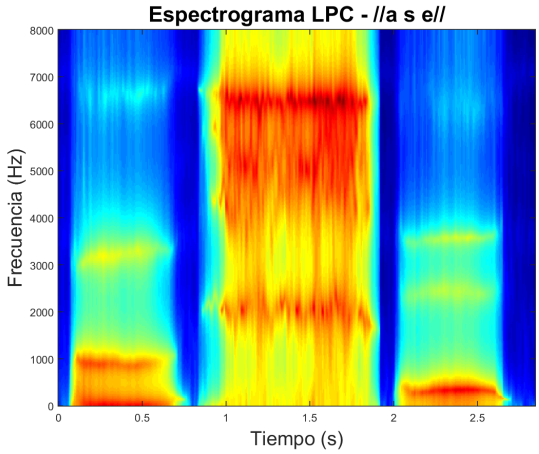

Figure 8.21: Espectrograma LPC - Cadena /a e o/

Sexo\Sonido	/a/	/e/	/o/	/s/
Masculino	62.63%	84.38%	53.13%	90.63%
Femenino	53.33%	86.66%	33.33%	100%

Table 8.5: Resultados de la fase de identificación de sonidos aislados

8.3.2 Identificación de palabras y silabras aisladas

El desarrollo del programa en la mayoría de los casos se hizo muy extenso por lo que se tuvo que prescindir de algunos fragmentos de la base de datos. Se distribuyó una fuente impresa con el espectro de la mayoría de las componentes en frecuencia de las letras del abecedario para así agilizar el proceso. En esta fase se optó por mostrar combinaciones de fragmentos vocálicos que resultaran sencillos de comprender ademas de temder al uso de palabras monosílabas o con una fuerte influencia vocálica. Finalmente se eligieron las siguientes: /uno/

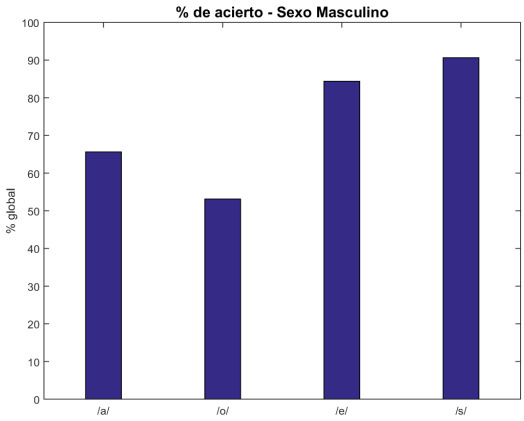

Figure 8.22: Procentaje de acierto en sonidos aislados - Sexo masculino

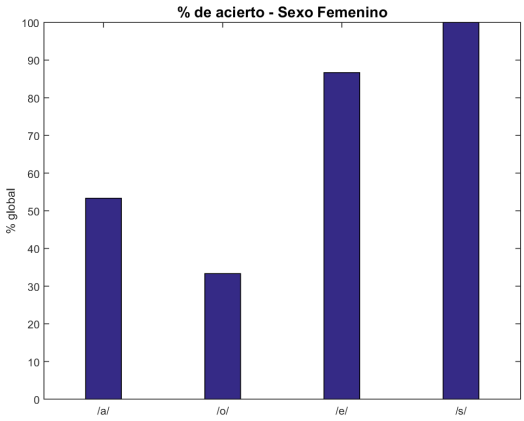

Figure 8.23: Procentaje de acierto en sonidos aislados - Sexo femenino

(figura 8.25), /dos/ (figura 8.26), /tres/ (figura 8.27) y /patata/ (figura 8.28).

La dificultad de esta prueba requiere de un tiempo de entrenamiento más avanzado por lo que los resultados en esta fase no fueron del todo complacientes. Aun así algunos sujetos presentaron aptitudes muy avanzadas siendo la mayoría capaces de reconocer las vocales, esencialmente la /e/ y la /a/ (esta última seguía siendo confundida con la /o/) o consonantes como /s/. Los casos más comunes fueron:

- En las palabras /dos/ y /tres/, los encuestados eran capaces de reconocer con facilidad las dos últimas letras, pero no eran capaces de reconocer la primera de las consonantes. La mayoría planteaba que las palabras terminarían en los fonemas /as/, /os/ u /es/ y propusieron aleatoriamente palabras monosílabas.

- En el caso de la palabra /uno/, la calidad del audio en la base de datos

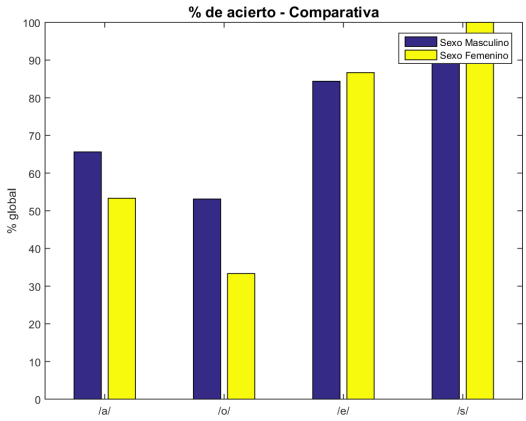

Figure 8.24: Procentaje de acierto en sonidos aislados - Comparativa

parecía difuminar y acomplejar el reconocimiento espectral. Propusieron vocablos aleatorios que no presentaban lógica alguna, pero siempre con la tendencia de resaltar que coexistía la vocal /a/ u /o/ en la palabra en cuestión.

- Con la palabra /patata/ la situación era compleja. La mayoría solo localizo la presencia de la vocal /a/ y lo interpreto como un tren de pulsos de la misma. Algunos reconocieron la presencia de una o varias consonantes pero no lograron reconocerlas ya que esta presentaba también energía a bajas frecuencias. En un periodo tan corto difícilmente podrían memorizar toda la base de datos del abecedario y las consonantes /p/ y /t/ no entraban dentro de sus conocimientos asimilados.

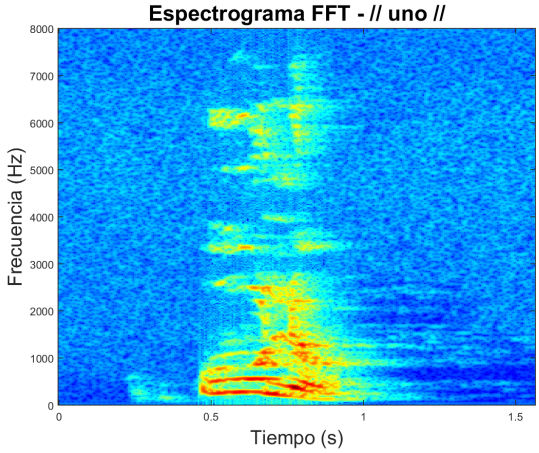

Figure 8.25: Espectro FFT - / uno /

Este nivel de la prueba concluye que existe una tendencia a mejorar las

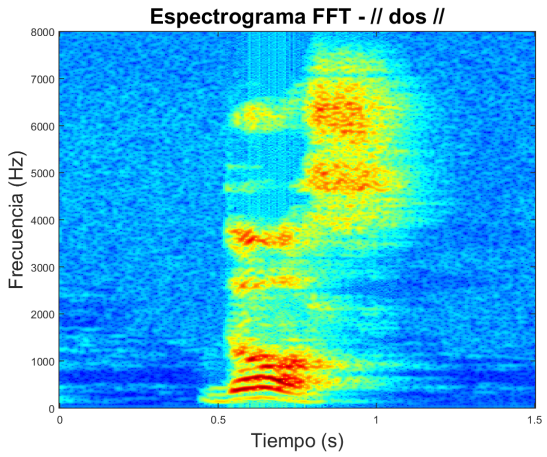

Figure 8.26: Espectro FFT - / dos /

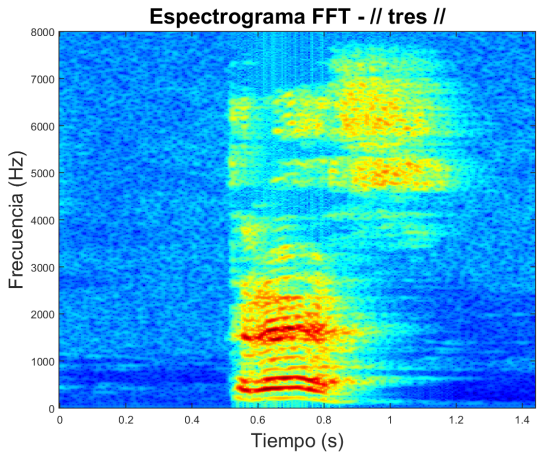

Figure 8.27: Espectro FFT - / tres /

capacidades de percepción y extracción de información conforme el sujeto es entrenado. La idea será desarrollar un programa completo que estructure cada una de las fases a diferentes niveles y que implique varias horas de entrenamiento semanales, cuestión que no incumbe en el proyecto actual. La complejidad incrementará no solo por las palabras usadas sino además de si se trata de un proceso en movimiento en tiempo real o un proceso estático la forma de visualizar los datos. Los sujetos tienden a presentar un mejor reconocimiento cuando la imagen es estática y realizan un análisis más profundo. Será la práctica la que mejore sus capacidades de reconocimiento y les permita alcanzar una aptitudes más agiles para la introducción en el análisis en tiempo real.

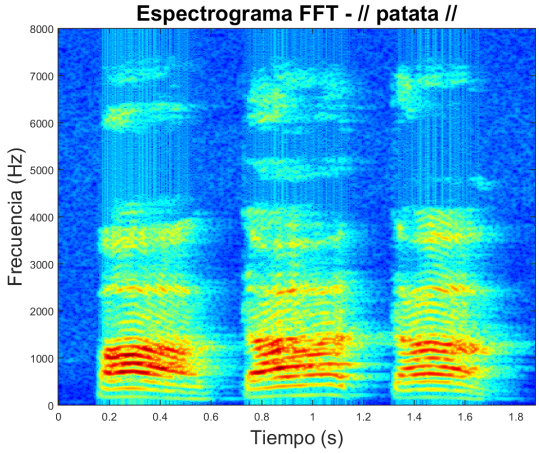

Figure 8.28: Espectro FFT / patata /

8.3.3 Identificación de frases cortas

Llegados a este punto, se determinó que no tenía sentido alguno continuar con esta fase proponiendo el problema en tiempo real. Si los sujetos no eran capaces de reconocer monosílabos difícilmente reconocerían palabras completas o frases. La idea fue imprimir varios espectrogramas escalados y mejorados, cortesía del profesor Ángel de la Torre Vega, para su análisis posterior. Se les planteaba que reconocieran las consonantes y vocales y finalmente las palabras que componían cada una de las frases expuestas.

Muchos sujetos se veían incapaces de reconocer palabra alguna pero si lograban distinguir distintas letras. Entre las tendencias más comunes destacamos:

- Los sujetos presentaban una tendencia a reconocer la existencia de una vocal correctamente. Coexistía el problema de que se solía confundir las vocales /a/ y /u/ con las vocales /o/ e /i/ respectivamente.
- Los sujetos reconocían sin problemas la presencia de vocales nasales (/n/, /m/ y /ñ/) pero tendía a no distinguir diferencias entre la /m/ y /n/.
- Existía una incertidumbre en el reconocimiento de las consonantes sordas. Se tendía a asignar el valor con el que estaban más familiarizados, la /s/, a todas las situaciones posibles. Un ejemplo muy llamativo es que la mayoría fue capaz de reconocer erróneamente una palabra genérica común, y es que confundían /siesas/ con /fiestas/.

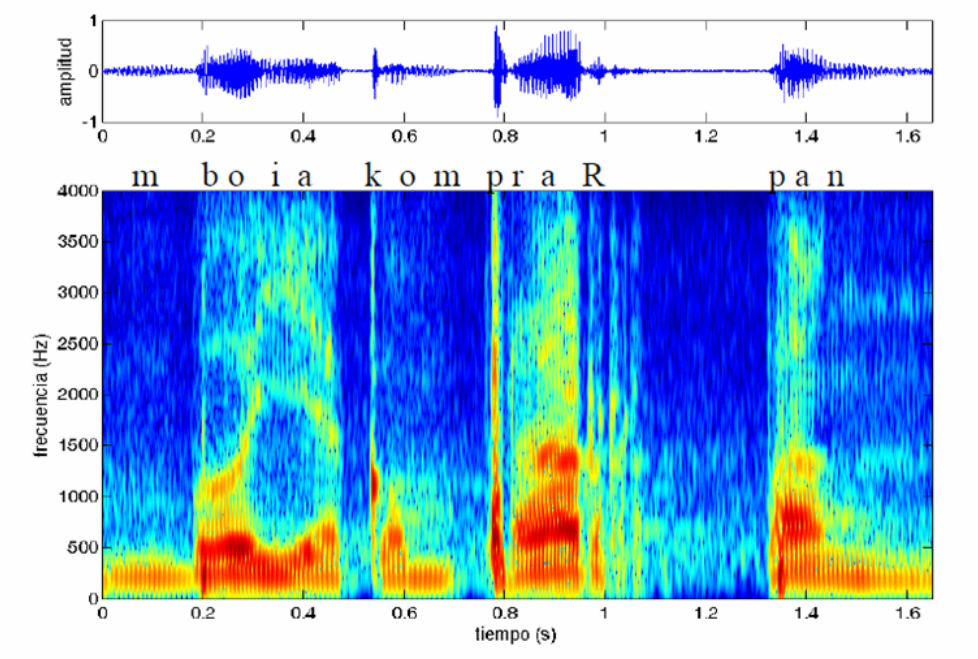

Figure 8.29: Espectrograma LPC - Frase: Me voy a comprar pán

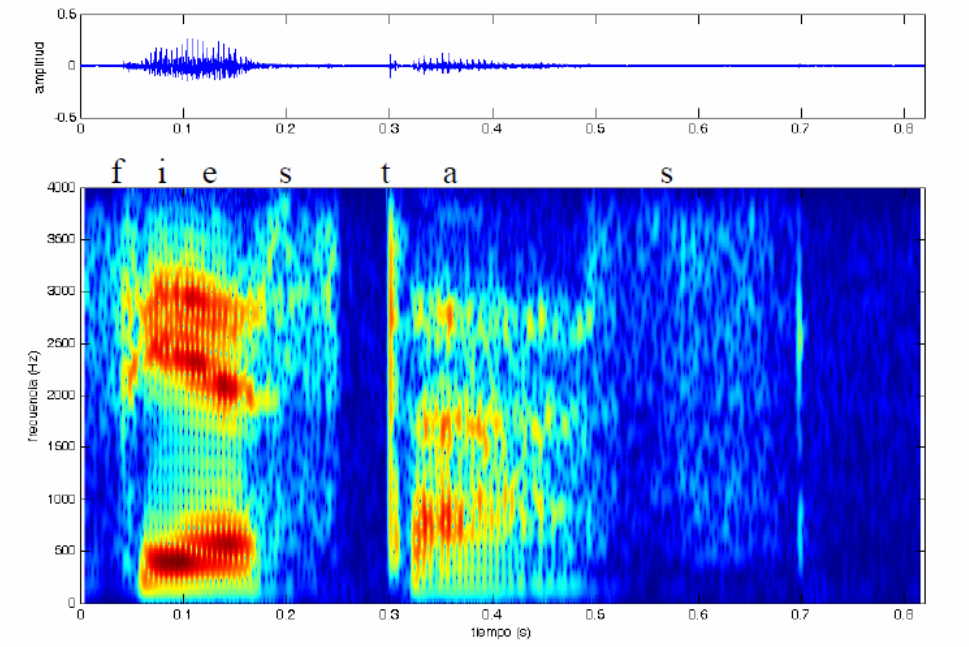

Figure 8.30: Espectrograma LPC - Frase: Fiestas

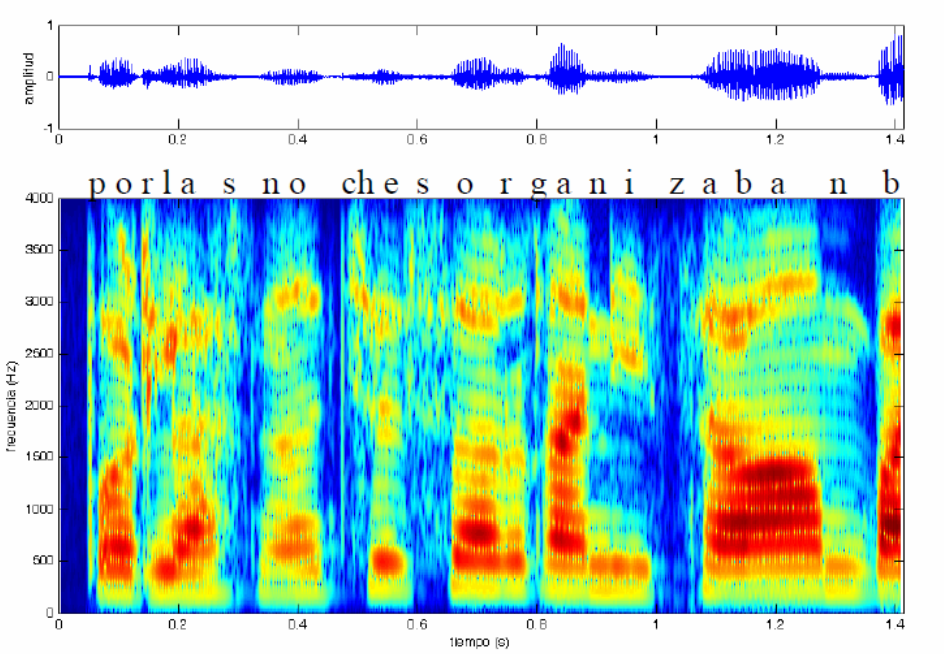

Figure 8.31: Espectrograma LPC - Frase: Por las noches organizaban

Capítulo 9

Evaluación de resultados

La herramienta ha presentado un comportamiento correcto en las distintas plataformas en las que se testeó durante cada una de las pruebas. Funcionaba fluidamente incluso en ordenadores de bajas prestaciones, y su fácil uso y personalización ha sido de agrado de todos los encuestados. LogoSpeech Studio finalmente cumple con todos los requisitos preestablecidos, incorporándosele, además, las sugerencias propuestas por los usuarios ha presentado retardos mínimos y rendimiento óptimos sobre ambas plataformas.

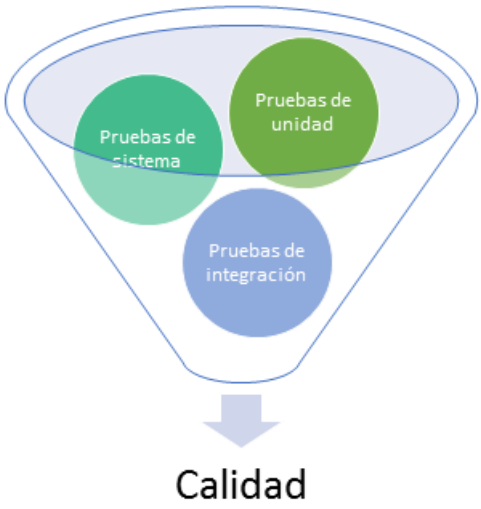

Figure 9.1: Control de calidad de la herramienta

La realización de cada una de las pruebas concluye unos resultados experimentales satisfactorios. El análisis temporal y espectral de las señales acústicas ha permitido la distinción de fonemas de forma aislada o dentro de un conjunto concreto de un determinado sector poblacional con pocas horas de entrenamiento. La falta de experiencia y la carencia de conocimiento básicos sobre el procesado de la señal de voz han dificultado y emborronado los resultados de las pruebas con niveles de complejidad más elevados.

Los candidatos han tenido dificultades especialmente en la distinción de sonidos que comparten similitudes, esencialmente entre oclusivas tanto sonoras como sordas. Los resultados están justificados por la carencia de esperanza, la falta de un análisis del contexto previo para extraer información prosódica de la señal y un programa de familiarización limitado.

En cualquier caso LogoSpeech Studio nació para servir como herramienta de entrenamiento para agudizar las capacidades de audición y fonación del usuario. La herramienta buscaba ser una alternativa para una posible percepción sonora aislada de forma visual y las pruebas han dejado evidencias empíricas claras de sus logros. Finalmente podemos concretar que se han alcanzado todos los objetivos y requisitos propuesto y que los resultados de las pruebas abren las puertas a la liberación de una primer versión reléase, LogoSpeech Studio 0.1, disponible para descargar desde el repositorio oficial en GitHub:

REPOSITORIO DE GIT

9.1 Conclusiones

El proyecto desarrollado presenta una perspectiva más cercana al desarrollo y la innovación que al de investigación en lo que respecta a I+D+i. Se basa fundamentalmente en la uso de conocimientos científicos sobre procesado de señal bioacústica para ofertar una solución de ingeniería para un determinado problema. Es por ello que las conclusiones que se pueden extraer se aferran más al éxito de la consecución de los objetivos así como en la experiencia adquirida al abordar el problema para posibles proyectos futuros.

El proyecto se centra en buscar solución a las necesidades que presentan tanto pacientes como instructores durante el proceso de aprendizaje y/o de rehabilitación lingüística. Se desarrolló bajo la premisa de construir una herramienta de entrenamiento para la mejora de las capacidades auditivas y fonadoras del usuario final partiendo con la idea de introducir una percepción sonora total o parcial de forma visual.

La idea esencial es hacer un uso de conceptos técnicos sobre procesado de señal de una forma más práctica, sencilla y compacta, para de esta forma, facilitar el proceso de asimilación y extracción de información. Algunos de los usos destacables de la aplicación:

- Uso como herramienta para el reconocimiento del habla: las pruebas realizadas han permitido a los usuarios reconocer determinados fonemas anulando por completo sus capacidades auditivas en muy pocas horas de entrenamiento. Un uso avanzado de la misma y una preparación adecuada podría permitir a un usuario el reconocimiento sonoro en tiempo real de forma eficiente y sencilla.

- Uso como herramienta para agilizar el proceso de rehabilitación auditiva para personas con audífonos o implantes cocleares: adquiridos los conocimientos básicos sobre las características de cada uno de los fonemas del lenguaje esta herramienta puede auxiliar durante el proceso de rehabilitación auditiva. Se trata de agilizar y agudizar las aptitudes alcanzadas, añadiendo a las capacidades auditivas limitadas del paciente la posibilidad de recibir información añadida visualmente.

- Uso como herramienta para mejorar la fonazión de personas con problemas dislálicos o similares: los pacientes con dislalia tienden a confundir o invertir sonidos en el proceso de fonación. Esta herramienta puede ayudar a distinguir los sonidos emitidos y agilizar el proceso de aprendizaje lingüístico.

- Uso como herramienta para la rehabilitación fonética de personas con carencias lingüísticas avanzadas o deficiencias auditivas: esta es quizás la utilidad más avanzada y útil. Las capacidades lingüísticas están marcadas tanto por las aptitudes fonadoras como auditivas, estando estas fuertemente relacionadas.

  Una persona con problemas auditivos con un tratamiento adecuado no tendría por qué limitar sus capacidades fonéticas. LogoSpeech Studio puede servir como herramienta de entrenamiento para un correcto aprendizaje fonético, el hecho de que el usuario pueda visualizar el sonido emitido agilizaría el proceso de integración lingüística y mejora de la vocalización.

- Uso como herramienta de apoyo para estudios fonéticos y fonológicos
- Uso como herramienta para el entrenamiento fonético durante el proceso de aprendizaje de terceros idiomas: esta opción está fuertemente relacionada con las anteriores. El ser humano tiende a adaptar sus capacidades articulatorias a su lenguaje materno por lo que presenta dificultades para aprender la fonación correcta de otros idiomas.

  LogoSpeech Studio sirve de herramienta para una correcta vocalización ya que el sujeto puede comparar los resultados que obtiene con los aislados de una determinada base de datos, partiendo del alfabeto fonético internacional.

Alfabeto Fonético Internacional - Consonantes											
	Bilabial	Labio dental	Dental	Alveolar	Post Alveolar	Retrof.	Palatal	Velar	Uvular	Faring.	Glotal
Oclusiva	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Vibrante múltiple	ʙ			r					ʀ		
Vibrante simple				ɾ		ɽ					
Fricativa	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Fricativa lateral				ɬ ɮ							
Aproximante		ʋ		ɹ		ɻ	j	ɰ			
Aproximante lateral				l		ɭ	ʎ	ʟ			
Oclusiva eyectiva	pʼ			tʼ		ʈʼ	cʼ	kʼ	qʼ		
Implosiva	ɓ ɗ			ʄ ɗ			ʗ ʄ	ɠ ɠ	ʛ ɢ		

Figure 9.2: Alfabeto Fonético Internacional - Consonantes

Determinamos entonces que el desarrollo del proyecto ha sido todo un éxito. La aplicación ha recibido una fuerte acogida tanto por los usuarios como por la comunidad de desarrolladores que está dispuesta a contribuir y hacer crecer la aplicación.

9.2 Trabajos futuros

La herramienta presenta unas grandes facultades evolutivas ya que permite la integración de plugins de terceros. Algunas de las posibles mejoras son la integración de nuevos algoritmos más robustos para la estimación de la frecuencia fundamental o nuevas técnicas de procesamiento para extracción de parámetros. Además requiere de un soporte técnico que garantice la evolución de la misma y garantice una mejora del rendimiento en implementaciones futuras.

9.2.1 Integración en dispositivos móviles

Hasta hace unos años era impensable la posibilidad de utilizar un dispositivo movil para el procesamiento digital. En la última década la tecnología ha sufrido

un avance explosivo que ha situado la potencia de muchos de los móviles inteligentes actuales en la órbita del rendimiento de muchos ordenadores. El auge de estos dispositivos es tal, que actualmente superan en ventas a los ordenadores, figura 9.3.

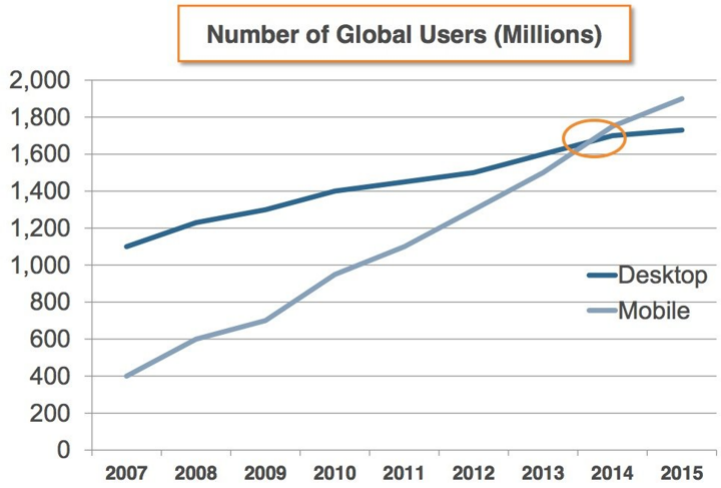

Figure 9.3: Ventas de dispositivos móviles frente a ordenadores - Fuente Morgan Stanley Research

El hecho de que sean dispositivos portátiles, manejables y asequibles para la mayoría de la población hace que una de las prioridades sea la implementación de una versión de LogoSpeech Studio para dispositivos móviles. La cuestión es que estos dispositivos al igual que ocurría con los ordenadores convencionales disponen de distintos sistemas operativos.

Se ha comenzado a desarrollar una versión para el sistema operativo más extendido, Android, figura 9.4. La versión recibe el nombre de LogoSpeech Studio Lite, y es una versión limitada del software con las características principales del mismo totalmente compatible con la mayoría de los dispositivos Android del mercado.

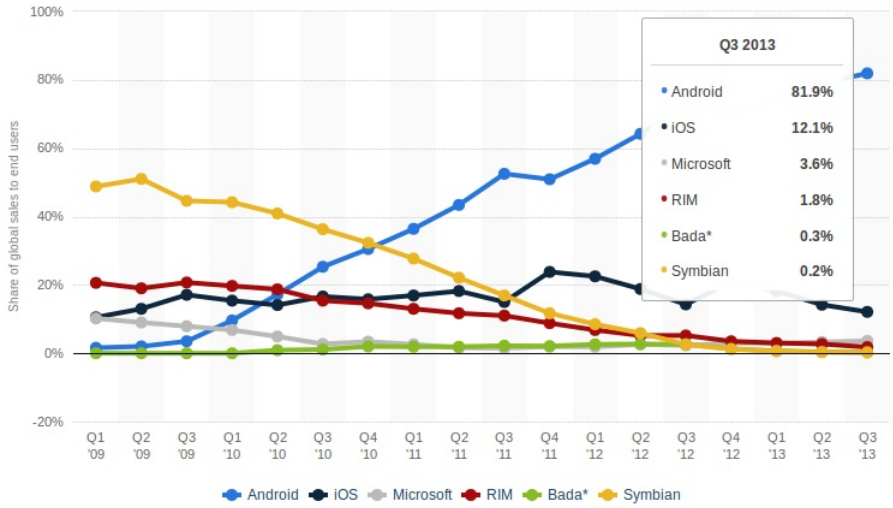

Figure 9.4: Evolución de los sistemas operativos en dispositivos móviles

Part IV

Anexos

Capítulo 10

Presupuesto

En este apartado se presentan las justificaciones de todos los costes globales que han supuesto la investigación, desarrollo y evaluación del presente proyecto. Incluye los costes imputables tanto a personal humano como a material informático requerido. Se toma como referencia el salario personal de un Ingeniero Junior en España de 1894.39€ para 131.25 horas mensuales. Análogamente siendo:

- A: dedicación en número de meses o fracción.
- B: periodo de depreciación en los que se amortiza el equipo
- C: Coste del software sin IVA.

El coste imputable para el material informático se puede obtener como:

$$P_i(€) = \frac{A \cdot C}{B}$$

Fase	Dedicación (h)	Coste I. Junior (Sueldo/mes)	C. I. (€)
Investigación	120	1894,39	1690,51
Desarrollo	400	1894,39	5635,04
Evaluación	130	1894,39	1831,88
Redacción de informa	50	1894,39	704,38
Horas totales	700	1894,39	9861.13

Table 10.1: Fases de elaboración de proyecto y horas invertidas

	C. Licencia (€)	Dedicación (h)	Depreciación (Meses)	C. I. (€)
M. Office 2013	94,00	10	24	0,05
MVS 2013	387,00	50	60	0,44
Matlab R2014a	2000,00	75	60	3,47
QtCreator	0,00	350	60	0
TexMaker	0,00	40	24	0
Coste Total	4387,00	525	-	3,96

Table 10.2: Material informático utilizado

Concepto	Coste Imputable
Desarrollo	9861,13 €
Software	3,96 €
Manutención	700,00 €
Local de investigación	1840,00 €
Total	12767,13 €

Table 10.3: Costes totales para el desarrollo de LogoSpeech Studio

Capítulo 11

Licencia GNU GPL

The GNU General Public License is a free, copyleft license for software and other kinds of works.

The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program–to make sure it remains free software for all its users. We, the Free Software Foundation, use the GNU General Public License for most of our software; it applies also to any other work released this way by its authors. You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for them if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs, and that you know you can do these things.

To protect your rights, we need to prevent others from denying you these rights or asking you to surrender the rights. Therefore, you have certain responsibilities if you distribute copies of the software, or if you modify it: responsibilities to respect the freedom of others.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must pass on to the recipients the same freedoms that you received. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

Developers that use the GNU GPL protect your rights with two steps: (1) assert copyright on the software, and (2) offer you this License giving you legal permission to copy, distribute and/or modify it.

For the developers' and authors' protection, the GPL clearly explains that there is no warranty for this free software. For both users' and authors' sake, the GPL requires that modified versions be marked as changed, so that their problems will not be attributed erroneously to authors of previous versions.

Some devices are designed to deny users access to install or run modified versions of the software inside them, although the manufacturer can do so. This is fundamentally incompatible with the aim of protecting users' freedom to change the software. The systematic pattern of such abuse occurs in the area of products for individuals to use, which is precisely where it is most unacceptable. Therefore, we have designed this version of the GPL to prohibit the practice for those products. If such problems arise substantially in other domains, we stand ready to extend this provision to those domains in future versions of the GPL, as needed to protect the freedom of users.

Finally, every program is threatened constantly by software patents. States should not allow patents to restrict development and use of software on general-purpose computers, but in those that do, we wish to avoid the special danger that patents applied to a free program could make it effectively proprietary. To prevent this, the GPL assures that patents cannot be used to render the program non-free.

The precise terms and conditions for copying, distribution and modification follow.

Terms and Conditions

0. Definitions.

"This License" refers to version 3 of the GNU General Public License.

"Copyright" also means copyright-like laws that apply to other kinds of works, such as semiconductor masks.

"The Program" refers to any copyrightable work licensed under this License. Each licensee is addressed as "you". "Licensees" and "recipients" may be individuals or organizations.

To "modify" a work means to copy from or adapt all or part of the work in a fashion requiring copyright permission, other than the making of an exact copy. The resulting work is called a "modified version" of the earlier work or a work "based on" the earlier work.

A "covered work" means either the unmodified Program or a work based on the Program.

To "propagate" a work means to do anything with it that, without permission, would make you directly or secondarily liable for infringement under applicable copyright law, except executing it on a computer or modifying a private copy. Propagation includes copying, distribution (with or without modification), making available to the public, and in some countries other activities as well.

To "convey" a work means any kind of propagation that enables other parties to make or receive copies. Mere interaction with a user through a computer network, with no transfer of a copy, is not conveying.

An interactive user interface displays "Appropriate Legal Notices" to the extent that it includes a convenient and prominently visible feature that (1) displays an appropriate copyright notice, and (2) tells the user that there is no warranty for the work (except to the extent that warranties are provided), that licensees may convey the work under this License, and how to view a copy of this License. If the interface presents a list of user commands or options, such as a menu, a prominent item in the list meets this criterion.

1. Source Code.

The "source code" for a work means the preferred form of the work for making modifications to it. "Object code" means any non-source form of a work.

A "Standard Interface" means an interface that either is an official standard defined by a recognized standards body, or, in the case of interfaces specified for a particular programming language, one that is widely used among developers working in that language.

The "System Libraries" of an executable work include anything, other than the work as a whole, that (a) is included in the normal form of packaging a Major Component, but which is not part of that Major Component, and

(b) serves only to enable use of the work with that Major Component, or to implement a Standard Interface for which an implementation is available to the public in source code form. A "Major Component", in this context, means a major essential component (kernel, window system, and so on) of the specific operating system (if any) on which the executable work runs, or a compiler used to produce the work, or an object code interpreter used to run it.

The "Corresponding Source" for a work in object code form means all the source code needed to generate, install, and (for an executable work) run the object code and to modify the work, including scripts to control those activities. However, it does not include the work's System Libraries, or general-purpose tools or generally available free programs which are used unmodified in performing those activities but which are not part of the work. For example, Corresponding Source includes interface definition files associated with source files for the work, and the source code for shared libraries and dynamically linked subprograms that the work is specifically designed to require, such as by intimate data communication or control flow between those subprograms and other parts of the work.

The Corresponding Source need not include anything that users can regenerate automatically from other parts of the Corresponding Source.

The Corresponding Source for a work in source code form is that same work.

2. Basic Permissions.

All rights granted under this License are granted for the term of copyright on the Program, and are irrevocable provided the stated conditions are met. This License explicitly affirms your unlimited permission to run the unmodified Program. The output from running a covered work is covered by this License only if the output, given its content, constitutes a covered work. This License acknowledges your rights of fair use or other equivalent, as provided by copyright law.

You may make, run and propagate covered works that you do not convey, without conditions so long as your license otherwise remains in force. You may convey covered works to others for the sole purpose of having them make modifications exclusively for you, or provide you with facilities for running those works, provided that you comply with the terms of this License in conveying all material for which you do not control copyright. Those thus making or running the covered works for you must do so exclusively on your behalf, under your direction and control, on terms that prohibit them from making any copies of your copyrighted material outside their relationship with you.

Conveying under any other circumstances is permitted solely under the conditions stated below. Sublicensing is not allowed; section 10 makes it unnecessary.

3. Protecting Users' Legal Rights From Anti-Circumvention Law.

No covered work shall be deemed part of an effective technological measure under any applicable law fulfilling obligations under article 11 of the WIPO

copyright treaty adopted on 20 December 1996, or similar laws prohibiting or restricting circumvention of such measures.

When you convey a covered work, you waive any legal power to forbid circumvention of technological measures to the extent such circumvention is effected by exercising rights under this License with respect to the covered work, and you disclaim any intention to limit operation or modification of the work as a means of enforcing, against the work's users, your or third parties' legal rights to forbid circumvention of technological measures.

4. Conveying Verbatim Copies.

   You may convey verbatim copies of the Program's source code as you receive it, in any medium, provided that you conspicuously and appropriately publish on each copy an appropriate copyright notice; keep intact all notices stating that this License and any non-permissive terms added in accord with section 7 apply to the code; keep intact all notices of the absence of any warranty; and give all recipients a copy of this License along with the Program.

   You may charge any price or no price for each copy that you convey, and you may offer support or warranty protection for a fee.

5. Conveying Modified Source Versions.

   You may convey a work based on the Program, or the modifications to produce it from the Program, in the form of source code under the terms of section 4, provided that you also meet all of these conditions:

   (a) The work must carry prominent notices stating that you modified it, and giving a relevant date.

   (b) The work must carry prominent notices stating that it is released under this License and any conditions added under section 7. This requirement modifies the requirement in section 4 to "keep intact all notices".

   (c) You must license the entire work, as a whole, under this License to anyone who comes into possession of a copy. This License will therefore apply, along with any applicable section 7 additional terms, to the whole of the work, and all its parts, regardless of how they are packaged. This License gives no permission to license the work in any other way, but it does not invalidate such permission if you have separately received it.

   (d) If the work has interactive user interfaces, each must display Appropriate Legal Notices; however, if the Program has interactive interfaces that do not display Appropriate Legal Notices, your work need not make them do so.

   A compilation of a covered work with other separate and independent works, which are not by their nature extensions of the covered work, and which are not combined with it such as to form a larger program, in or on a volume of a storage or distribution medium, is called an "aggregate" if the compilation and its resulting copyright are not used to limit the access or legal rights of the compilation's users beyond what the individual works

permit. Inclusion of a covered work in an aggregate does not cause this License to apply to the other parts of the aggregate.

6. Conveying Non-Source Forms.

   You may convey a covered work in object code form under the terms of sections 4 and 5, provided that you also convey the machine-readable Corresponding Source under the terms of this License, in one of these ways:

   (a) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by the Corresponding Source fixed on a durable physical medium customarily used for software interchange.

   (b) Convey the object code in, or embodied in, a physical product (including a physical distribution medium), accompanied by a written offer, valid for at least three years and valid for as long as you offer spare parts or customer support for that product model, to give anyone who possesses the object code either (1) a copy of the Corresponding Source for all the software in the product that is covered by this License, on a durable physical medium customarily used for software interchange, for a price no more than your reasonable cost of physically performing this conveying of source, or (2) access to copy the Corresponding Source from a network server at no charge.

   (c) Convey individual copies of the object code with a copy of the written offer to provide the Corresponding Source. This alternative is allowed only occasionally and noncommercially, and only if you received the object code with such an offer, in accord with subsection 6b.

   (d) Convey the object code by offering access from a designated place (gratis or for a charge), and offer equivalent access to the Corresponding Source in the same way through the same place at no further charge. You need not require recipients to copy the Corresponding Source along with the object code. If the place to copy the object code is a network server, the Corresponding Source may be on a different server (operated by you or a third party) that supports equivalent copying facilities, provided you maintain clear directions next to the object code saying where to find the Corresponding Source. Regardless of what server hosts the Corresponding Source, you remain obligated to ensure that it is available for as long as needed to satisfy these requirements.

   (e) Convey the object code using peer-to-peer transmission, provided you inform other peers where the object code and Corresponding Source of the work are being offered to the general public at no charge under subsection 6d.

   A separable portion of the object code, whose source code is excluded from the Corresponding Source as a System Library, need not be included in conveying the object code work.

   A "User Product" is either (1) a "consumer product", which means any tangible personal property which is normally used for personal, family,

or household purposes, or (2) anything designed or sold for incorporation into a dwelling. In determining whether a product is a consumer product, doubtful cases shall be resolved in favor of coverage. For a particular product received by a particular user, "normally used" refers to a typical or common use of that class of product, regardless of the status of the particular user or of the way in which the particular user actually uses, or expects or is expected to use, the product. A product is a consumer product regardless of whether the product has substantial commercial, industrial or non-consumer uses, unless such uses represent the only significant mode of use of the product.

"Installation Information" for a User Product means any methods, procedures, authorization keys, or other information required to install and execute modified versions of a covered work in that User Product from a modified version of its Corresponding Source. The information must suffice to ensure that the continued functioning of the modified object code is in no case prevented or interfered with solely because modification has been made.

If you convey an object code work under this section in, or with, or specifically for use in, a User Product, and the conveying occurs as part of a transaction in which the right of possession and use of the User Product is transferred to the recipient in perpetuity or for a fixed term (regardless of how the transaction is characterized), the Corresponding Source conveyed under this section must be accompanied by the Installation Information. But this requirement does not apply if neither you nor any third party retains the ability to install modified object code on the User Product (for example, the work has been installed in ROM).

The requirement to provide Installation Information does not include a requirement to continue to provide support service, warranty, or updates for a work that has been modified or installed by the recipient, or for the User Product in which it has been modified or installed. Access to a network may be denied when the modification itself materially and adversely affects the operation of the network or violates the rules and protocols for communication across the network.

Corresponding Source conveyed, and Installation Information provided, in accord with this section must be in a format that is publicly documented (and with an implementation available to the public in source code form), and must require no special password or key for unpacking, reading or copying.

7. Additional Terms.

"Additional permissions" are terms that supplement the terms of this License by making exceptions from one or more of its conditions. Additional permissions that are applicable to the entire Program shall be treated as though they were included in this License, to the extent that they are valid under applicable law. If additional permissions apply only to part of the Program, that part may be used separately under those permissions, but the entire Program remains governed by this License without regard to the additional permissions.

When you convey a copy of a covered work, you may at your option remove any additional permissions from that copy, or from any part of it. (Additional permissions may be written to require their own removal in certain cases when you modify the work.) You may place additional permissions on material, added by you to a covered work, for which you have or can give appropriate copyright permission.

Notwithstanding any other provision of this License, for material you add to a covered work, you may (if authorized by the copyright holders of that material) supplement the terms of this License with terms:

(a) Disclaiming warranty or limiting liability differently from the terms of sections 15 and 16 of this License; or

(b) Requiring preservation of specified reasonable legal notices or author attributions in that material or in the Appropriate Legal Notices displayed by works containing it; or

(c) Prohibiting misrepresentation of the origin of that material, or requiring that modified versions of such material be marked in reasonable ways as different from the original version; or

(d) Limiting the use for publicity purposes of names of licensors or authors of the material; or

(e) Declining to grant rights under trademark law for use of some trade names, trademarks, or service marks; or

(f) Requiring indemnification of licensors and authors of that material by anyone who conveys the material (or modified versions of it) with contractual assumptions of liability to the recipient, for any liability that these contractual assumptions directly impose on those licensors and authors.

All other non-permissive additional terms are considered "further restrictions" within the meaning of section 10. If the Program as you received it, or any part of it, contains a notice stating that it is governed by this License along with a term that is a further restriction, you may remove that term. If a license document contains a further restriction but permits relicensing or conveying under this License, you may add to a covered work material governed by the terms of that license document, provided that the further restriction does not survive such relicensing or conveying.

If you add terms to a covered work in accord with this section, you must place, in the relevant source files, a statement of the additional terms that apply to those files, or a notice indicating where to find the applicable terms.

Additional terms, permissive or non-permissive, may be stated in the form of a separately written license, or stated as exceptions; the above requirements apply either way.

8. Termination.

You may not propagate or modify a covered work except as expressly provided under this License. Any attempt otherwise to propagate or modify it is void, and will automatically terminate your rights under this License

(including any patent licenses granted under the third paragraph of section 11).

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, you do not qualify to receive new licenses for the same material under section 10.

9. Acceptance Not Required for Having Copies.

   You are not required to accept this License in order to receive or run a copy of the Program. Ancillary propagation of a covered work occurring solely as a consequence of using peer-to-peer transmission to receive a copy likewise does not require acceptance. However, nothing other than this License grants you permission to propagate or modify any covered work. These actions infringe copyright if you do not accept this License. Therefore, by modifying or propagating a covered work, you indicate your acceptance of this License to do so.

10. Automatic Licensing of Downstream Recipients.

    Each time you convey a covered work, the recipient automatically receives a license from the original licensors, to run, modify and propagate that work, subject to this License. You are not responsible for enforcing compliance by third parties with this License.

    An "entity transaction" is a transaction transferring control of an organization, or substantially all assets of one, or subdividing an organization, or merging organizations. If propagation of a covered work results from an entity transaction, each party to that transaction who receives a copy of the work also receives whatever licenses to the work the party's predecessor in interest had or could give under the previous paragraph, plus a right to possession of the Corresponding Source of the work from the predecessor in interest, if the predecessor has it or can get it with reasonable efforts.

    You may not impose any further restrictions on the exercise of the rights granted or affirmed under this License. For example, you may not impose a license fee, royalty, or other charge for exercise of rights granted under this License, and you may not initiate litigation (including a cross-claim or counterclaim in a lawsuit) alleging that any patent claim is infringed by making, using, selling, offering for sale, or importing the Program or any portion of it.

11. Patents.

A "contributor" is a copyright holder who authorizes use under this License of the Program or a work on which the Program is based. The work thus licensed is called the contributor's "contributor version".

A contributor's "essential patent claims" are all patent claims owned or controlled by the contributor, whether already acquired or hereafter acquired, that would be infringed by some manner, permitted by this License, of making, using, or selling its contributor version, but do not include claims that would be infringed only as a consequence of further modification of the contributor version. For purposes of this definition, "control" includes the right to grant patent sublicenses in a manner consistent with the requirements of this License.

Each contributor grants you a non-exclusive, worldwide, royalty-free patent license under the contributor's essential patent claims, to make, use, sell, offer for sale, import and otherwise run, modify and propagate the contents of its contributor version.

In the following three paragraphs, a "patent license" is any express agreement or commitment, however denominated, not to enforce a patent (such as an express permission to practice a patent or covenant not to sue for patent infringement). To "grant" such a patent license to a party means to make such an agreement or commitment not to enforce a patent against the party.

If you convey a covered work, knowingly relying on a patent license, and the Corresponding Source of the work is not available for anyone to copy, free of charge and under the terms of this License, through a publicly available network server or other readily accessible means, then you must either (1) cause the Corresponding Source to be so available, or (2) arrange to deprive yourself of the benefit of the patent license for this particular work, or (3) arrange, in a manner consistent with the requirements of this License, to extend the patent license to downstream recipients. "Knowingly relying" means you have actual knowledge that, but for the patent license, your conveying the covered work in a country, or your recipient's use of the covered work in a country, would infringe one or more identifiable patents in that country that you have reason to believe are valid.

If, pursuant to or in connection with a single transaction or arrangement, you convey, or propagate by procuring conveyance of, a covered work, and grant a patent license to some of the parties receiving the covered work authorizing them to use, propagate, modify or convey a specific copy of the covered work, then the patent license you grant is automatically extended to all recipients of the covered work and works based on it.

A patent license is "discriminatory" if it does not include within the scope of its coverage, prohibits the exercise of, or is conditioned on the non-exercise of one or more of the rights that are specifically granted under this License. You may not convey a covered work if you are a party to an arrangement with a third party that is in the business of distributing software, under which you make payment to the third party based on the extent of your activity of conveying the work, and under which the third party grants, to any of the parties who would receive the covered work

from you, a discriminatory patent license (a) in connection with copies of the covered work conveyed by you (or copies made from those copies), or (b) primarily for and in connection with specific products or compilations that contain the covered work, unless you entered into that arrangement, or that patent license was granted, prior to 28 March 2007.

Nothing in this License shall be construed as excluding or limiting any implied license or other defenses to infringement that may otherwise be available to you under applicable patent law.

12. No Surrender of Others' Freedom.

If conditions are imposed on you (whether by court order, agreement or otherwise) that contradict the conditions of this License, they do not excuse you from the conditions of this License. If you cannot convey a covered work so as to satisfy simultaneously your obligations under this License and any other pertinent obligations, then as a consequence you may not convey it at all. For example, if you agree to terms that obligate you to collect a royalty for further conveying from those to whom you convey the Program, the only way you could satisfy both those terms and this License would be to refrain entirely from conveying the Program.

13. Use with the GNU Affero General Public License.

Notwithstanding any other provision of this License, you have permission to link or combine any covered work with a work licensed under version 3 of the GNU Affero General Public License into a single combined work, and to convey the resulting work. The terms of this License will continue to apply to the part which is the covered work, but the special requirements of the GNU Affero General Public License, section 13, concerning interaction through a network will apply to the combination as such.

14. Revised Versions of this License.

The Free Software Foundation may publish revised and/or new versions of the GNU General Public License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns.

Each version is given a distinguishing version number. If the Program specifies that a certain numbered version of the GNU General Public License "or any later version" applies to it, you have the option of following the terms and conditions either of that numbered version or of any later version published by the Free Software Foundation. If the Program does not specify a version number of the GNU General Public License, you may choose any version ever published by the Free Software Foundation.

If the Program specifies that a proxy can decide which future versions of the GNU General Public License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Program.

Later license versions may give you additional or different permissions. However, no additional obligations are imposed on any author or copyright holder as a result of your choosing to follow a later version.

15. Disclaimer of Warranty.

    THERE IS NO WARRANTY FOR THE PROGRAM, TO THE EXTENT PERMITTED BY APPLICABLE LAW. EXCEPT WHEN OTHERWISE STATED IN WRITING THE COPYRIGHT HOLDERS AND/OR OTHER PARTIES PROVIDE THE PROGRAM "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. THE ENTIRE RISK AS TO THE QUALITY AND PERFORMANCE OF THE PROGRAM IS WITH YOU. SHOULD THE PROGRAM PROVE DEFECTIVE, YOU ASSUME THE COST OF ALL NECESSARY SERVICING, REPAIR OR CORRECTION.

16. Limitation of Liability.

    IN NO EVENT UNLESS REQUIRED BY APPLICABLE LAW OR AGREED TO IN WRITING WILL ANY COPYRIGHT HOLDER, OR ANY OTHER PARTY WHO MODIFIES AND/OR CONVEYS THE PROGRAM AS PERMITTED ABOVE, BE LIABLE TO YOU FOR DAMAGES, INCLUDING ANY GENERAL, SPECIAL, INCIDENTAL OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE OR INABILITY TO USE THE PROGRAM (INCLUDING BUT NOT LIMITED TO LOSS OF DATA OR DATA BEING RENDERED INACCURATE OR LOSSES SUSTAINED BY YOU OR THIRD PARTIES OR A FAILURE OF THE PROGRAM TO OPERATE WITH ANY OTHER PROGRAMS), EVEN IF SUCH HOLDER OR OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

17. Interpretation of Sections 15 and 16.

    If the disclaimer of warranty and limitation of liability provided above cannot be given local legal effect according to their terms, reviewing courts shall apply local law that most closely approximates an absolute waiver of all civil liability in connection with the Program, unless a warranty or assumption of liability accompanies a copy of the Program in return for a fee.

End of Terms and Conditions

How to Apply These Terms to Your New Programs

If you develop a new program, and you want it to be of the greatest possible use to the public, the best way to achieve this is to make it free software which everyone can redistribute and change under these terms.

To do so, attach the following notices to the program. It is safest to attach them to the start of each source file to most effectively state the exclusion of warranty; and each file should have at least the "copyright" line and a pointer to where the full notice is found.

```
<one line to give the program's name and a brief idea of what it does.>
```

```
Copyright (C) <textyear>  <name of author>

This program is free software: you can redistribute it and/or modify
it under the terms of the GNU General Public License as published by
the Free Software Foundation, either version 3 of the License, or
(at your option) any later version.

This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
GNU General Public License for more details.

You should have received a copy of the GNU General Public License
along with this program.  If not, see <http://www.gnu.org/licenses/>.
```

Also add information on how to contact you by electronic and paper mail.

If the program does terminal interaction, make it output a short notice like this when it starts in an interactive mode:

```
<program>  Copyright (C) <year>  <name of author>

This program comes with ABSOLUTELY NO WARRANTY; for details type ‘show w’.
This is free software, and you are welcome to redistribute it
under certain conditions; type ‘show c’ for details.
```

The hypothetical commands `show w` and `show c` should show the appropriate parts of the General Public License. Of course, your program's commands might be different; for a GUI interface, you would use an "about box".

You should also get your employer (if you work as a programmer) or school, if any, to sign a "copyright disclaimer" for the program, if necessary. For more information on this, and how to apply and follow the GNU GPL, see `http://www.gnu.org/licenses/`.

The GNU General Public License does not permit incorporating your program into proprietary programs. If your program is a subroutine library, you may consider it more useful to permit linking proprietary applications with the library. If this is what you want to do, use the GNU Lesser General Public License instead of this License. But first, please read `http://www.gnu.org/philosophy/why-not-lgpl.html`.

Bibliography

[1] Claudio Becchetti and Lucio Prina Ricotti. *Speech Recognition. Theory and C++ implementation.* John Wiley and Sons, Ltd, 1999.

[2] Francisco CasaCuberta and Enrique Vidal. *Reconocimiento Automático del habla.* Marcombo Boixareu Editores, 1987.

[3] Valentín Cardeñoso Payo César Llamas Bello. *Reconocimiento automático del habla. Técnicas y aplicación.* Universidad de Valladolid, 1997.

[4] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing*, 1980.

[5] Gustavo Santos García. *Inteligencia artificial y matemática aplicada.* Secretariado de Publicaciones e Intercambio Científico, Universidad de Valladolid, 2001.

[6] Vincenzo Fabio Rollo y Gabriele Venturi Giuliano Antoniol. Linear predictive coding and cepstrum coefficients for mining time variant information from software repositories. *RCOST- University Of Sannio*, 2000.

[7] Mark Hasegawa-Johnson. Lecture notes in speech production, speech coding, and speech recognition. *University of Illinois at Urbana-Champaign*, 2000.

[8] Jyh-Shing Roger Jang. Pitch tracking. *National Taiwan University*, 2012.

[9] Jeús Bobadilla Sancho y Pedro Gomez Vilda Jesús Bernal Bermúdez. *Reconocimiento de voz y fonética acústica.* RA-MA Editorial, 2000.

[10] M Jiménez Torres, M y López Sánchez. *Deficiencia auditiva.* Editorial CEPE., 2003.

[11] Jhon G. Proakis y John H. L. Hansen John R. Deller JR. *Discrete-Time Processing of Speech Signals.* Prentice Hall, Upper Saddle River, 1993.

[12] Hae Young Kim ; Jae Sung Lee ; Myung-Whun Sung ; Kwang Hyun Kim. Pitch detection with average magnitude difference function using adaptive threshold algorithm for estimating shimmer and jitter. *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 20(6), 1998.

[13] A. M. Kondoz. *Digital Speech. Coding for low bit rate communications systems.* John Wiley and Sons, Ltd, 1994.

[14] A. M. Kondoz. *Digital Speech. Coding for low bit rate communications systems.* John Wiley & Sons Ltd., 1999.

[15] P. Lieberman. *The biology and evolution of languaje.* Harvard University Press. Cambridge, 1984.

[16] A.E. Rosenberg L.R. Rabiner, M.J. Cheng and C.A. McGonegal. A comparative performance study of seceral pitch detection algorithms. *IEEE Trans. on ASSP*, October, 1796.

[17] A Manrique, M.J. ; Huarte. *Implantes cocleares.* 2001.

[18] Ghulam Muhammad. Extended average magnitude difference function based pitch detection. *The International Arab Journal of Information Technology,*, 8(2), 2011.

[19] P.Preethi N. Raju, S. Mathini T. Lakshmi and M. Chandrasekar. Identifying the population of animals through pitch, formant, short time energy - a sound analysis. *ICCEET*, 2012.

[20] V.E. Negus. *The comparative anatomy and physiology of the larynx.* Hafner, 1949.

[21] Antonio M Peinado and José C. Segura. *Speech Recognition over digital channels, Robustness and Standar.* John Wiley and Sons, Ltd, 2006.

[22] Lawrence Rabiner. Time domain methods in speech processing. 2012.

[23] Pedro L. Galindo Riaño. *Introducción al reconocimiento de la voz.* Dpto. de Lenguajes y Sistemas Informáticos, Universidad de Cádiz, 1996.

[24] A.E. Rosengerg. Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.*, Febrero 1971.

[25] H. ; Cohen A. ; Freudberg R. Ross, M. ; Shaffer. Average magnitude difference function pitch extractor. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 22, 1974.

[26] J. Volkman y NewMA S. Stevens. E.a. scale for the mesarurement of the psychological magnitude of pitch. *Journal of the Acoustical Society of America*, 1937.

[27] Gustavo Santos-García. *Inteligencia artificial y matemática aplicada.* Universidad de Valladolid, 2001.

[28] L. Rabiner R. Schafer. *Digital Processing of Speech Signals.* Prentice Hall, 1978.

[29] Ronald W. Schafer and Lawrence R. Rabiner. Digital representations of speech signals. *Proceedings of the IEEE*, 63(4), 1975.

[30] Kang Guangyu ; Guo Shize. Improving amdf for pitch period detection. *The Ninth International Conference on Electronic Measurement & Intruments*, 2009.

[31] Bernard. Sklar. *Digital Communications: Fundamentals and Applications.* Englewood Cliffs, N.J.:Prentice-Hall,, 1988.

[32] ETSI Standard. Etsi es 202 211 v1.1.1 (2003-11). *ETSI Standard*, 2003.

[33] Li Tan and Montri Karnjanadecha. Pitch detection algorithm: Autocorrelation method and amdf.

[34] Li Ru-Wei ; Cao Long tao ; Li Yang. Pitch detection method for noisy speech signals based on wavelet transform and autocorrelation function. *Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2013.

[35] Hartmut Traunmüller. Analytical expressions for the tonotopic sensory scale. *ournal of the Acoustical Society of America*, 1990.

[36] Hsiao-wuen hon Xuedong Huang, Alex Acero. *Spoken Language Processing: A guide to Theory, Algorithm, and System Development.* Prentice Hall, 2001.

[37] Lawrence Rabiner y Biing-Hwang Juang. *Fundamentals of speech recognition.* Prentice-Hall, Inc. Upper Saddle River, 1993.

[38] Antonio M. Peinado y José C. Segura. *Speech recognition over digital channels. Robustness and standards.* John Wiley & Sons, Ltd, 2006.

[39] Marcos Faúndez Zanuy. *Tratamiento digital de voz e imagen y aplicación a la multimedia.* Marcombo Boixareu Editores, 2000.

[40] Huan Zao and Wenjie Gan. A new pitch estimation method based on amdf. *Journal of multimedia*, 8(5), Octubre 2013.

[41] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models.* Springer-Verlag, 1990.