Universidad de Zaragoza
Departamento de Ingeniería Electrónica y Comunicaciones

Tesis doctoral

Aplicación de las Tecnologías del Habla en la Educación de la Voz Infantil Alterada

Application of Speech Technology in the Education of Children's Altered Voice

William Ricardo Rodríguez Dueñas

Director de Tesis
Prof. Eduardo Lleida Solano

October 12, 2010

Esta tesis esta dedicada a mi familia,
sin olvidar a mis amigos, a mi gente,
y por supuesto, a mi bello país.

..que las pequeñas contribuciones fruto
de esta investigación, germinen y se
conviertan en reales mejoras de la
calidad de vida de personas con
capacidades diferentes.

Agradecimientos

En el transcurso de estos años de investigación son innumerables los hechos y situaciones que generan sentimientos de gratitud, tenerlos en cuenta a todos ellos en unas pocas líneas es una labor de por si complicada. Lo que siempre he tenido claro en todo momento, es que esta tesis doctoral no hubiese sido posible sin la ayuda de Eduardo, él, mi tutor, mi mentor, y mi amigo, quien un buen día decidió abrirle la puerta a un ingeniero biomédico colombiano para hacer parte de su equipo de trabajo. Hoy cuatro años después y con esta tesis finalizada, aun me siento corto al decirle gracias, gracias por la oportunidad, por el tiempo, la paciencia, y por su siempre buena disposición para conmigo y la investigación.

También quisiera agradecer a todo el grupo de tecnologías del habla de la Universidad de Zaragoza, por su colaboración, su apoyo, y constante orientación, a Oscar Saz por permitirme trabajar a su lado y aprender tantas cosas de él al responder mis infinitas preguntas, y por supuesto, a mis compañeros de laboratorio por su eterna paciencia con mis vocales fuera de tono.

Agradezco también al Banco Santander por brindarme el apoyo económico tan fundamental para cualquier persona que quiera realizar una tesis doctoral, a las instituciones de educación especial y sus profesionales de *la Alborada* en España y *CEDESNID* en Colombia, por su orientación y por permitirme trabajar con ellos en la búsqueda de soluciones para las personas con capacidades diferentes.

Por supuesto, no dejare de agradecer siempre a mi familia por creer en mi, por darme ánimos, y recordarme siempre lo valioso de la unidad familiar. Gracias también a todos aquellos en España y Colombia que, de un modo u otro, contribuyeron a que en este momento éstas páginas puedan ser leídas.

Resumen

El trabajar con voz infantil alterada es una tarea difícil para los profesionales de terapia de voz, y más aun para quienes trabajan con población infantil con discapacidad. Los profesionales de instituciones de ayuda a la discapacidad y colegios de educación especial, a pesar de conocer las ventajas de trabajar de la mano con la tecnología, experimentan grandes necesidades y limitaciones a la hora de tratar la voz alterada de este tipo de población. Por una parte, por las pocas herramientas disponibles las cuales tienen un alto costo de adquisición, y por otro lado, porque la mayoría vienen en idiomas diferentes al español además de ofrecer limitadas opciones para trabajar con sonidos vocálicos. Los terapeutas se ven obligados en la mayoría de los casos a trabajar la voz con herramientas manuales fruto de su iniciativa, que les demandan mucho tiempo en su preparación reduciendo el tiempo disponible para trabajar con cada niño.

La tesis doctoral: *Aplicación de las Tecnologías del Habla en la Educación de la Voz Infantil Alterada*, afronta este problema estudiando las características acústicas de la voz infantil sin alteraciones aplicando Tecnologías del Habla (TH), para crear herramientas libres en español y para español que permitan educar la voz alterada de un niño con o sin discapacidad. El punto de partida para lograrlo fue entrar en contacto con instituciones especializadas para conocer su entorno de trabajo, sus necesidades y herramientas disponibles, y también fue necesaria la adquisición de un corpus de voz infantil no alterada para la estimación de sus parámetros acústicos. Se estimaron parámetros como la energía, la frecuencia fundamental o pitch, y con mucha dificultad y poca fiabilidad los formantes vocálicos, debido a la alta tonalidad presente en este tipo de voz. Este problema se abordó utilizando técnicas tradicionales en procesado de voz, y proponiendo un método que basado en dichas técnicas, permite estimar de manera robusta los formantes vocálicos en voz infantil, y al mismo tiempo, reducir su alta variabilidad por medio de una normalización en la que se utiliza una estimación de la longitud del tracto vocal del locutor.

Con los parámetros acústicos estimados de manera robusta, se inició la etapa de desarrollo de herramientas libres para terapia de voz, las cuales están disponibles en www.vocaliza.es para la comunidad en

general, y cuyos resultados y aportes recibidos por parte de quienes las están utilizando, han permitido la mejora y continua evolución de las herramientas. Con el objetivo de evaluar la herramienta principal fruto de esta investigación denominada *PreLingua*, se diseño un estudio para aplicar la herramienta en casos reales de población con discapacidad y voz alterada en dos instituciones de educación especial, y cuyos resultados cuantitativos y cualitativos muestran los beneficios y limitantes de esta herramienta para tratar la voz, así como otros beneficios derivados de aplicar la tecnología propuesta en la comunicación pre-lingüística y demás áreas de la discapacidad y la educación especial.

La discusión científica y las conclusiones muestran que, aunque trabajar con voz infantil alterada es una tarea difícil, la aplicación de las TH en la educación de la voz infantil alterada es posible y viable, los resultados cualitativos obtenidos demuestran que se está trabajando por buen camino, y que estas tecnologías tienen un gran campo de aplicación y especialmente, un alto potencial para intentar mejorar la calidad de vida de estas personas.

Abstract

Working with altered child's voice is a difficult task by speech therapist, even more fot those who work with impaired children. Therapist of institutions which offer aid to impaired people and special education schools, despite knowing the benefits of working hand in hand with technology, experience considerable needs and constraints in dealing with altered voice of this kind of population. On the one hand, by the few tools available that have a high acquisition cost, and on the other hand, due most of them are available in other languages than Spanish and offer limited options for dealing with vocalic sounds. Therapists are forced in most cases to work with hand tools that demand a long time for being build, reducing the time available to work with each child.

The PhD thesis: *Application of Speech Technology in the Education of Children's Altered Voice*, addresses this problem by studying the acoustic parameters in voices without alterations using speech technologies in order to create free tools in Spanish to allow the education of altered voices in children with or without disabilities. The starting point to achieve this, was to know the specialized institutions, their needs, and available tools, and it was also necessary the acquisition of speech corpus in order to estimate its acoustic parameters. Parameters as energy, fundamental frequency or pitch, and formant frequencies with difficulty were estimated, due to the typical high pitch present in this type of voice. This problem was addressed by using traditional techniques in speech processing, and propose a method based on these techniques that allow to estimated formant frequencies in children's speech robustly, and at the same time reducing the high variability by means of a normalization that uses an estimation of the vocal tract length.

With the acoustic parameters estimated robustly, the development of free tools for voice therapy started and now are available in www.vocaliza.es to the whole community. The results and contributions received from those who are using the tools have allowed the improvement of them. In order to evaluate the main tool of this thesis called *PreLingua*, it was designed a study to apply this tool in real cases of people with disabilities and altered voices in two special education

institutes, the quantitative and qualitative results show the benefits and limitations of this tool for training voice as well as other benefits of implementing the proposed technology in pre-linguistic communication and other areas related to special education.

The scientific discussion and conclusions shows that while working with voice impaired children is a difficult task, the application of speech technologies in the education of altered voice is possible, the qualitative results obtained show this work goes on the right direction and these technologies have a wide range of applications, especially its high potential to try to improve the quality of life of these people.

Índice

Indice de figuras

Indice de tablas

Lista de Acronimos

ECM Error Cuadrático Medio

FFT Transformada Rápida de Fourier

GRBAS Grade, Rough, Breathy, Asthenic

GMMs Modelos de Mezclas de Gaussianas

GTC Grupo de Tecnologías de las Comunicaciones

HMM Hidden Markov Model

I3A Instituto de Investigación en Ingeniería de Aragón (Aragon Institute for Engineering Research)

IES Institución de Educación Secundaria

IPA International Phonetic Alphabet

LPC Linear Prediction Coefficients

LTV Longitud del Tracto Vocal

MAP Maximum A Posteriori

MFCC Mel Frequency Cepstral Coefficients

ML Maximum Likelihood

MRI Magnetic Resonance Image

RAH Reconocimiento Automático del Habla

SAMPA Speech Assessment Methods Phonetic Alphabet

TH Tecnologías del Habla

TIC Tecnologías de la Información y la Comunicación

TME Tiempo Máximo de Espiración

TMF Tiempo Máximo de Fonación

VAD Voice Activity Detector

VTLN Vocal Tract Length Normalization

WSS Wide Sense Stationary

Capítulo 1

Introducción

1.1 Introducción

Desde la antigüedad la población con discapacidad ha sido vulnerable y tratada de formas diferentes y muchas veces injustas hasta nuestros días, en la edad media, se consideraba la discapacidad como un castigo de Dios o posesión demoníaca, la sociedad no tenía la más mínima responsabilidad con las personas discapacitadas. La revolución industrial permitió que las personas discapacitadas fueran vistas como responsabilidad publica, y empezaban a verse diferentes, en el último siglo la situación cambia de forma positiva gracias a diversos factores como los avances de la medicina, una mejor educación de la comunidad frente al problema de las personas con discapacidad, avances en la ciencia y la creación de nuevas ramas de la salud más acordes a sus necesidades. Sin embargo, pese a los progresos logrados en el siglo XX, la sociedad en general seguía considerando a las personas con limitaciones como un problema, hoy día los niños discapacitados aun tienen que luchar contra una marginación educativa y de acceso a la tecnología, los niños discapacitados constituyen la minoría más desfavorecida del mundo, ya que se estima que el 20 por ciento de la población más pobre del mundo está formada por discapacitados, y que en los países en desarrollo más del 90 por ciento de los niños discapacitados no asisten a la escuela y son víctimas de exclusión.

Ahora vivimos en la sociedad de la información, en la era del conocimiento, las Tecnologías de la Información y la Comunicación (TIC) nos permiten el acceso a este conocimiento y por el mismo motivo no deben convertirse en un elemento más de marginación y discriminación a nivel educativo y social, por el contrario, las TIC deben permitir y potenciar en un niño con discapacidad su desarrollo integral y su inclusión social con dignidad. Las Tecnologías del Habla (TH) al formar parte de las TIC pueden ayudar a lograr este cometido, su evolución en las últimas décadas han permitido el avance en el estudio de la voz, el habla, y el desarrollo de sistemas especializados como el Reconocimiento Automático del Habla (RAH) o de síntesis de voz. Entonces, porque no aprovechar también estos avances científicos para que ésta población con necesidades quizá más básicas, tenga acceso a la tecnología y de alguna manera se contribuya a mejorar su calidad de vida?.

Los profesionales de logopedia y educación especial conscientes de la ventaja de trabajar de la mano con la tecnología, experimentan grandes necesidades y limitaciones a la hora de trabajar con población infantil con discapacidad y voz alterada, no solo por las pocas

herramientas disponibles y por su alto costo de adquisición, sino porque la mayoría vienen en idiomas diferentes al español y por las limitadas prestaciones de éstas para trabajar con sonidos vocálicos entre otras necesidades. Los terapeutas se ven obligados en la mayoría de los casos a trabajar la voz con herramientas manuales fruto de su iniciativa, como láminas, inflar globos, trabajar frente a espejos con los niños, imitar los sonidos de instrumentos musicales, y un sin número de otros imaginativos recursos para poder trabajar con los niños sus problemas de voz. De manera que con apoyo de las TH, ésta tesis se propone estudiar las características acústicas de la voz infantil sin alteraciones, para crear herramientas libres en español y para español que permitan educar la voz alterada de un niño con o sin discapacidad.

1.2 Motivación de la Tesis

La aplicación de las TH en la integración de niños con discapacidad se encuentra todavía en un periodo inicial del proceso, los esfuerzos actuales se orientan principalmente al desarrollo de sistemas de ayuda a la logopedia en la adquisición del habla y el lenguaje, y en menor medida, a herramientas para trabajar directamente la voz infantil. Ésta tesis está orientada no sólo a realizar aportaciones al conocimiento científico-técnico de la voz infantil, sino también a la creación de herramientas con aplicación real sobre usuarios con voz alterada con o sin discapacidad, en caso contrario, todo el conocimiento, recursos, y esfuerzos invertidos, no serán realmente tales mientras no brinden una funcionalidad a usuarios con necesidades reales.

1.3 Objetivos y Metodología

Esta investigación se plantea estudiar y proponer una alternativa a la problemática de trabajar con voz infantil alterada, que brinde apoyo a los profesionales de instituciones de ayuda a la discapacidad y colegios de educación especial. Para lograrlo, se adquirirá un corpus de voz infantil no alterada para su análisis y conocimiento a fondo, y así poder proponer un método que basado en las técnicas existentes de procesado de voz, permitan trabajar de manera robusta este tipo de voz, y posibiliten la creación de herramientas para ser aplicadas en población infantil con discapacidad.

Teniendo en cuenta lo anterior, la presente tesis se plantea por una parte cumplir unos objetivos científicos para llegar a la tecnología que permita trabajar con robustez la voz infantil, y por otro lado, unos objetivos de desarrollo que permitan aplicar dicha tecnología creando herramientas libres para terapia de voz, y así beneficiar directamente tanto a terapeutas como a población infantil con discapacidad.

1.3.1 Objetivos Científicos

Hay fundamentalmente tres objetivos científicos a cumplir en ésta tesis:

- El primer objetivo de la tesis es lograr un *acercamiento a instituciones especializadas en logopedia y educación especial, y la adquisición de un corpus de voz no alterada*

para investigación; Un adecuado acercamiento al mundo de la terapia en logopedia y educación especial, permitirá ubicar la investigación en el contexto adecuado y conocer mejor las necesidades reales y herramientas disponibles para los profesionales de este medio. De igual manera se buscará establecer alianzas con instituciones educativas con población infantil y adolescente, para la adquisición de un corpus de voz no alterada y disponer así de material para la investigación.

- En posesión del corpus de voz infantil no alterada de la fase anterior, el siguiente objetivo a completar será la *investigación sobre técnicas de procesado de voz que permitan la estimación robusta de parámetros acústicos en la voz infantil, y establecer como cambian estos en función del crecimiento y sexo*; Una vez establecido el método que basado en las técnicas existentes de procesado de voz estimen de mejor manera sus parámetros acústicos, en especial los formantes de manera robusta sin que la alta tonalidad los afecte, la siguiente tarea será establecer como cambian estos formantes en función de la talla y sexo del locutor.

- El siguiente objetivo por cumplir consistirá en *reducir la alta variabilidad formántica entre diferentes locutores por medio de alguna técnica de normalización*; Debido a que la información formántica depende en gran medida de características geométricas del tracto vocal como su longitud, existe una alta variabilidad entre los formantes de diferentes locutores, de manera que se debe trabajar en la reducción de ésta variabilidad llevando los formantes estimados a un espacio más homogéneo de trabajo por medio de una normalización.

1.3.2 Objetivos de Desarrollo

Como fruto de la presente investigación, se espera cumplir con dos objetivos relacionados con el desarrollo de herramientas, y que tengan aplicación en casos reales de niños con discapacidad y con voz alterada.

- El primero de ellos es *la creación de herramientas para terapia de voz en español y de libre distribución*; Se busca el desarrollo de herramientas que permitan trabajar con voz infantil alterada, diseñadas en español y para español, que sean de libre uso y que estén disponibles para toda la comunidad hispano-hablante.

- Por otra parte, *que las herramientas desarrolladas representen para el terapeuta una ayuda real en su trabajo, y que su diseño de cara al usuario final sea el más adecuado posible.*

1.3.3 Metodología

Para poder alcanzar los objetivos propuestos, la metodología seguida en la investigación siguió el diagrama de bloques de la Figura 1.1.

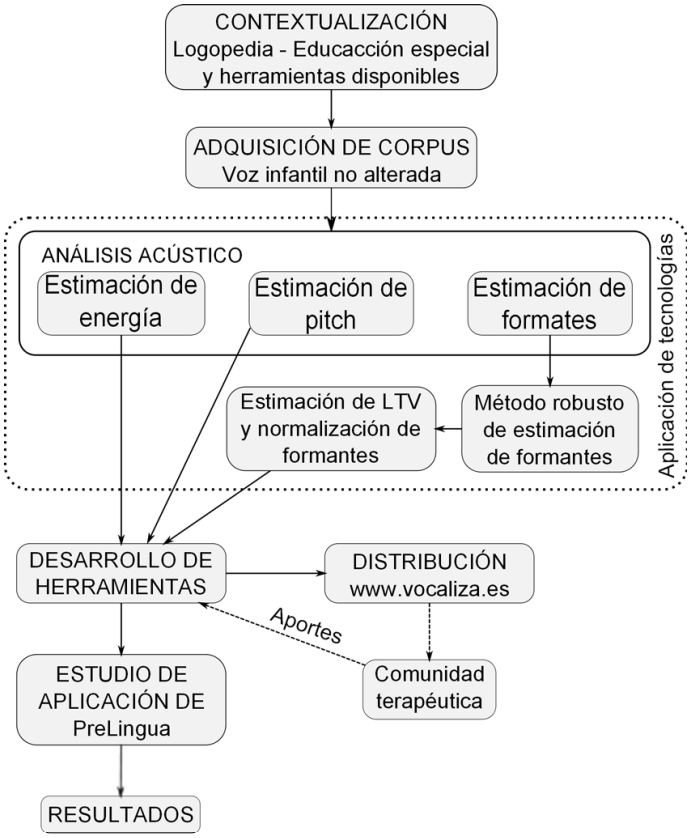


Figura 1.1: *Procedimiento metodológico.*

El punto de partida fue entrar en contacto con instituciones del mundo de la logopedia y la educación especial, como también conocer las herramientas disponibles en el mercado para trabajar alteraciones de la voz, este contacto con instituciones especializadas permitió la adquisición de un corpus de voz infantil no alterada que dio inicio formal a la etapa de experimentación e investigación, con el análisis acústico de la señal de voz para obtener sus parámetros acústicos. Se estimaron parámetros como la energía y la frecuencia fundamental o pitch, y con mucha dificultad y poca fiabilidad los formantes vocálicos, debido a la alta tonalidad presente en este tipo de voz. Utilizando técnicas tradicionales en procesamiento de voz, se propuso un método para estimar de manera robusta los formantes vocálicos infantiles, y reducir su alta variabilidad por medio de una normalización en la que se utiliza una estimación de la longitud del tracto vocal de cada locutor.

Con los parámetros acústicos de energía, pitch, y formantes estimados de manera robusta, se inició la etapa de desarrollo de herramientas libres para terapia de voz, las cuales están disponibles en www.vocaliza.es para la comunidad en general, y cuyos resultados y aportes recibidos de quienes las están utilizando han permitido la mejora y continua evolución de las herramientas.

Con el objetivo de evaluar la herramienta principal fruto de ésta investigación denominada *PreLingua*, se diseño un estudio para aplicar la herramienta en casos reales de población con discapacidad y voz alterada en dos instituciones de educación especial, y cuyos resultados muestran los beneficios y limitantes de la herramienta para tratar la

voz, así como las muchas otras aplicaciones de la tecnología propuesta en la comunicación pre-lingüística y otras áreas de la discapacidad y educación especial.

1.4 Organización

La presente tesis se organiza en 5 partes con 12 capítulos como lo muestra la Figura 1.2:

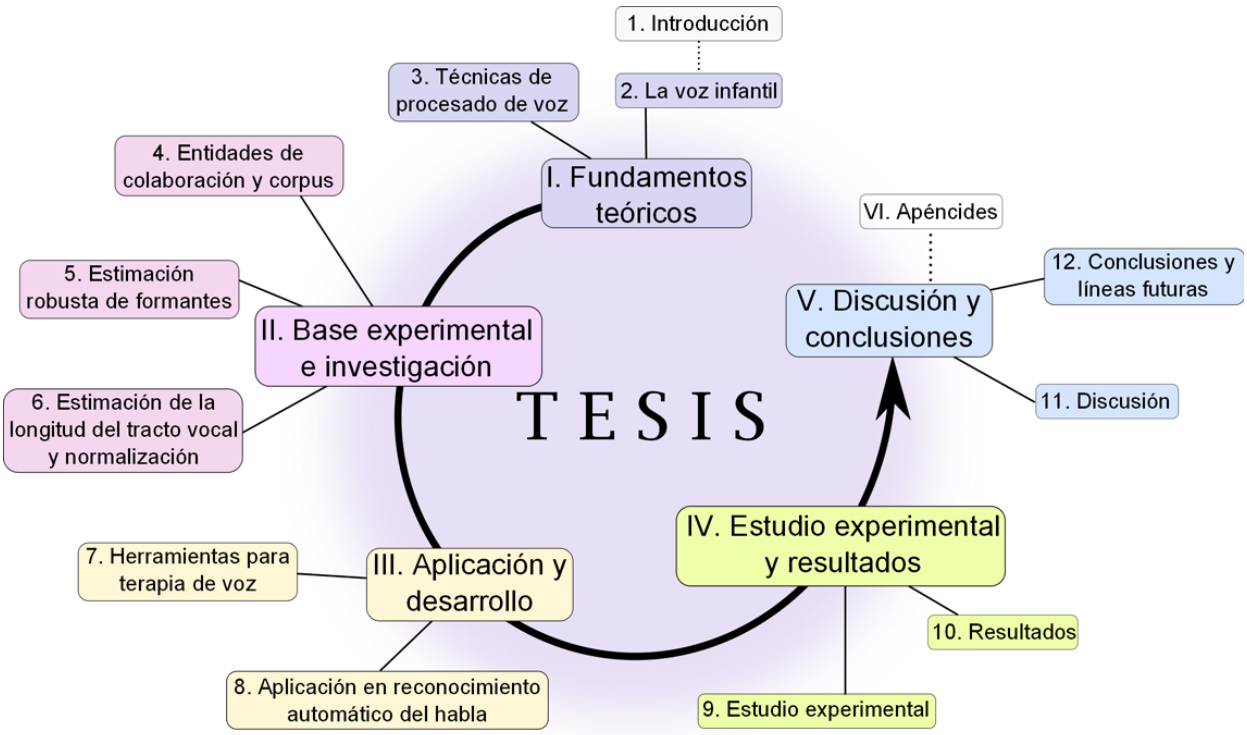


Figura 1.2: *Organización de la tesis.*

- La primera parte de *Fundamentos teóricos* comprende los Capítulos 2 y 3. El primero de ellos reúne generalidades sobre la voz infantil desde la perspectiva terapéutica y algunas herramientas encontradas en el mercado para trabajar con voz infantil, pero con un uso mínimo por parte de los profesionales de ésta área. El tercer capítulo esboza las técnicas de procesado de voz implicadas en la investigación, y que en su conjunto permitieron la consecución de los objetivos propuestos.

- En la segunda parte denominada *Base Experimental e Investigación*, está el Capítulo 4 en el que se citan los colegios e instituciones de educación especial que apoyaron la investigación, y que permitieron la grabación del corpus de voz infantil no alterada. El Capítulo 5 muestra por su parte las dificultades técnicas encontradas en la estimación fiable de formantes en la voz infantil, también, como aplicando técnicas como el análisis LPC y homomórfico es posible mejorar éstas estimaciones de manera robusta. En el Capítulo 6, se describe como obtener una estimación fiable de la longitud del tracto vocal de un locutor determinado a partir de sus propia información formántica,

lo que permite hacer una normalización de los formantes y así reducir la alta variabilidad inter-locutor presente en la población infantil, debida fundamentalmente a las diferentes longitudes de sus tractos vocales.

- La tercera parte: *Aplicación y Desarrollo*, reúne en el Capítulo 7 el conjunto de herramientas desarrolladas para trabajar con voz infantil, las cuales permiten una aplicación real de la tecnología propuesta y que cuentan con una gran potencial de beneficio al ser éstas de libre distribución. El Capítulo 8 muestra la aplicación de la tecnología propuesta en la tarea de reconocimiento automático del habla, en donde a partir de la longitud del tracto vocal estimada para un locutor determinado, se propone un factor de deformación de frecuencia para su aplicación en tiempo real.

- El *Estudio Experimental y Resultados* de la cuarta parte, describe un estudio realizado en dos instituciones de educación especial en España y Colombia en casos reales con discapacidad y los resultados obtenidos. El Capítulo 9 describe las características del estudio y de la población participante, y en el Capítulo 10 se muestran los resultados obtenidos de manera cuantitativa y cualitativa.

- La quinta parte de *Discusión y Conclusiones*, analiza en el Capítulo 11 hasta que punto la herramienta *PreLingua* puede ser considerada como una herramienta para el tratamiento y evaluación de la voz infantil, según los resultados obtenidos en el estudio realizado. También se discute el impacto y difusión obtenidos por la herramienta en la comunidad terapéutica, y el potencial de la tecnología en otras aplicaciones. El Capítulo 12 muestra por su parte, un breve resumen del trabajo realizado en la tesis, los aportes realizados por la misma y como se cumplieron los objetivos propuestos, finalmente, describe algunas líneas de trabajo futuras y diferentes ponencias y publicaciones como indicio de calidad de la tesis.

- Finalmente, la sexta parte con una breve sección de *Apéndices* en donde se cita el motor gráfico *Allegro* en el apéndice A, el cual posibilitó el desarrollo de las herramientas al ser un conjunto de librerías gratuitas para videojuegos escritas en código C, y en el apéndice B, se muestra la evaluación logopédica realizada por los terapeutas a la población participante del estudio, y de la que se obtuvieron los resultados cualitativos.

Parte I

Fundamentos Teóricos

Capítulo 2

La Voz Infantil

La especial problemática que atañe a esta tesis hace que conocer las alteraciones de la voz y en especial la voz infantil, sea tan necesario como conocer el estado del arte del ámbito tecnológico en el procesado de voz. La voz aparte de ser el principal canal de comunicación entre los humanos es el más eficaz y por la misma razón requiere de una especial atención en la población infantil, ya que en general si el infante posee una alteración en su voz él no es consciente de dicha situación.

2.1 Consideraciones Sobre la Voz Infantil

La voz es percibida por el bebe desde su estancia en el vientre materno durante embarazo, la voz de la madre se transmite por su estructura ósea hasta la cavidad pélvica donde es percibida por el bebe a partir de la semana 24 de gestación aproximadamente. Este hecho contribuye de modo definitivo a estrechar los vínculos afectivos entre madre e hijo y se crea formalmente el primer canal de comunicación entre ellos.

Después del nacimiento el sistema fonatorio y articulatorio del bebe evoluciona desde su función estricta de supervivencia hasta las funciones comunicativas. En situaciones de estrés o que le puedan generar algún tipo de angustia al bebe, el hecho de escuchar la voz de la madre le produce sosiego, calma y sensación de bienestar, puesto que le recuerda la placidez y la ausencia de necesidades de su vida intrauterina. Y que mejor para llamar la atención de la madre que el llanto y los gritos, el descubrimiento de la acción que tiene la voz propia sobre el interlocutor en este caso la madre y posteriormente otras personas, fija su gran valor comunicativo para el bebe.

Es así como el llanto y los gritos forman parte de la vida cotidiana del bebe y es cuando comienza el desarrollo prelingüístico que tiene lugar durante los primeros 12 meses de vida aproximadamente. Aparecen entonces las primeras producciones que los adultos interpretan como palabras (protopalabras) con formas vocales bastante estables. En este primer año ocurre una gradual sintonización hacia la lengua del entorno, tanto en el nivel productivo como en el perceptivo [Bosch, 2004].

2.1.1 Adquisiciones Prelingüísticas

Las primeras habilidades de comunicación se denominan adquisiciones o conductas prelingüísticas, ya que no se trata de lenguaje en sentido estricto antes de que el niño empiece a utilizar los recursos convencionales del lenguaje, es decir, antes de que empiece el segundo año de vida. En la tabla 2.1 se muestran los principales aspectos del desarrollo prelingüístico. Modificado de [Puyuelo et al., 2004].

Tabla 2.1: *Aspectos Principales del Desarrollo Prelingüístico.*

1. Inicio de los mecanismos básicos de comunicación
6 *primeros meses* - Gritos y lloros - Pueden determinar la aparición del adulto para satisfacción de necesidades - Adulto: Interlocutor privilegiado - 4.° o 5.° mes: El niño es capaz de seguir con los ojos la dirección de la mirada del adulto, situación que esta ligada a la actividad verbal (detección de actividad de voz)
7 *a* 12 *meses* - En la relación personal no verbal se pasa de la modalidad de demanda, a una modalidad de intercambio y reciprocidad - Se crean varios juegos y rutinas que se organizan según un modelo de intercambio (rutinas de aseo, alimento..) - Se asumen papeles de conducido y actuado, y correlativamente de conductor y agente
2. De la expresión global e indiferencia al balbuceo controlado y a un principio de comprensión verbal
6 *primeros meses* - Hacía el 2.º mes el adulto familiar puede reconocer los gritos y lloros según la razón que lo ocasiona: hambre, dolor, incomodidad - Hacia el 3.º o 4.º mes empieza el balbuceo, actividad vocal poco diferenciada pero se pueden reconocer algunas vocales
7 *a* 12 *meses* - Aparición progresiva en el balbuceo del niño de elementos tipo consonántico con bloqueo y después relajación de la corriente de aire espirado, lo que incluye una modulación de la intensidad - Hacia el 8.º mes pueden aparecer combinaciones de consonantes y vocales - Hacia el final del primer año el balbuceo del niño gana claridad articulatoria, vocalizaciones más cortas y más numerosas - Reproducción de la entonación del lenguaje escuchado y repetición de elementos vocales del mismo lenguaje - Comprensión de ciertas tonalidades (adulto bien intencionado/adulto mal intencionado)

En este primer año el niño aprende del adulto y de otros niños más mayores que forman parte de su entorno, y utiliza los mecanismos básicos de la comunicación a nivel preverbal. Pasa progresivamente de una forma global de expresión y de comunicación, utilizando todo

el cuerpo, a una forma más diferenciada que recurre principalmente a la actividad vocal y que tiene como telón de fondo la expresión y la comunicación gestual. La actividad vocal evoluciona considerablemente durante los 15 primeros meses, desde los gritos y los lloros de las primeras semanas al balbuceo y al control articulatorio observable en la producción de las primeras palabras y la capacidad de reproducción inmediata (aunque aproximada) de las palabras producidas por el interlocutor adulto.

Finalmente, durante el primer año de vida y el principio del segundo se desarrolla la comprensión verbal. El niño comprende ciertas palabras y algunas expresiones que aparecen en contextos apropiados antes de empezar a expresarse a través de palabras [Puyuelo et al., 2004]. Respecto a los cambios estructurales, al nacer la laringe se sitúa entre la 3ª y 4ª vértebras cervicales posibilitando que la respiración sea nasal y que al mismo tiempo se permita la deglución. La posición alta de la laringe genera una voz muy aguda y nasalizada. Progresivamente comienza el descenso de la laringe permitiendo el retroceso de la lengua y la liberación de los movimientos articulatorios dando lugar a los inicios del lenguaje articulado. Gracias a estos cambios estructurales de la laringe, el cambio acústico más significativo de la voz es el frecuencial, pasa de los 400 Hz del llanto del bebe a los 110 Hz en los niños y a los y 220 Hz en las niñas tras la pubertad [Vila, 2009].

2.2 Interpretación terapéutica de la voz

La voz enferma y nos enferma. La disfonía o alteración de la voz es un fenómeno corriente muy habitual en el mundo de los niños. Normalmente las disfonías que se presentan en los niños son de corta duración y asociadas con facilidad por los padres a estados gripales o resfriados, o bien a excesos vocales en fiestas o prácticas deportivas. La disfonía infantil no es un fenómeno fácilmente observable de manera objetiva por los padres ni, a veces, por parte de pediatras. Los padres se habitúan a ciertas características de la voz de sus hijos y no entienden esas características como alteradas.

Algunas manifestaciones de la disfonía infantil son más evidentes que otras y hacen que sus padres soliciten ayuda médica. Algunos padres se alertan por el elevado esfuerzo que hacen sus hijos al hablar, otros detectan que su hijo no grita y no se le escucha de lejos, o del lado contrario, que su hijo siempre habla muy fuerte. También es conocido que muchas de las alteraciones acústicas de la voz tienen su origen en manifestaciones histológicas o morfológicas del aparato fonador, y algunas de estas pueden tener origen congénito o funcional [Vila, 2009].

En el caso de niños con discapacidad, son innumerables los diagnósticos que pueden tener repercusiones en la calidad de la voz del niño. Encontramos por ejemplo el retardo mental, la sordera, síndrome de Down, parálisis cerebral, distrofia muscular, hipotonía, hipertonía, etc. También las malformaciones o problemas anatómicos comprometen de manera importante las cualidades de la voz especialmente la articulación.

En el campo de la logopedia y educación vocal, se entiende por voz alterada o disfonía la alteración de sus cualidades acústicas, estas son: la Intensidad, el Tono, el Timbre y la Duración. Sea una de ellas o diversas combinaciones de ellas, una modificación significativa

de los valores respecto a los considerados normales puede ser vivida por el sujeto o por su entorno como una alteración.

2.2.1 Intensidad

La intensidad o volumen de la voz es la característica física resultante de la presión del aire en su paso por las cuerdas vocales que dificultan su salida. Se debe tener en cuenta que la intensidad se incrementa por la participación de los espacios y paredes de resonancias que, amplifican las frecuencias, y la sensación de volumen. Sus unidades son los decibelios (dB) y en la exploración infantil se suele registrar la intensidad mínima, conversacional, proyectada y de grito. Habitualmente los niños disfónicos tienen dificultad para producir sonidos de baja intensidad. [Vila, 2009].

2.2.2 Tono

El tono o pitch, es el resultado de la vibración de las cuerdas vocales en la fonación. Sus unidades son los hercios (Hz) y suelen tomarse tres valores en la exploración infantil: la frecuencia mínima, la frecuencia máxima y la espontánea. Este valor va descendiendo lentamente desde la infancia hasta la pubertad, y a partir de allí ocurre un descenso brusco en el caso de los hombres y menos acentuado en las mujeres.

2.2.3 Timbre

El timbre es la personalidad de la voz, propio de cada persona. Esta constituido por la frecuencia fundamental, sus armónicos y formantes cuando el sonido inicial de la laringe pasa por el tracto vocal. Estos formantes dependen de la disposición variable de los órganos vocales, (lengua, mandíbula, labios, velo del paladar) y a su vez varían según la talla, genero y raza del niño entre otros factores.

2.2.4 Duración

Es el tiempo de permanencia de las vibraciones sonoras durante la emisión de la voz. Los tiempos máximos de fonación son de gran valor diagnóstico y permiten valorar la evolución del paciente. También es de interés conocer el tiempo máximo de espiración es decir sin la generación de sonidos sonoros, ya que el cociente entre el tiempo máximo de fonación y el tiempo máximo de espiración, permite una valoración de la eficiencia del cierre glótico.

La voz es el resultado de un complejo proceso en el que participan en mayor o menor medida muy distintos elementos de nuestro cuerpo, de manera que valorar los aspectos acústicos anteriormente descritos, no son en absoluto los únicos elementos a tener en cuenta por un profesional a la hora de explorar la voz infantil alterada.

2.3 Exploración Profesional

La valoración de la voz es una tarea compleja que requiere la intervención de varios profesionales. Entre ellos están los logopedas o fonoaudiólogos, médicos foniatras u otorrinolaringólogos que en conjunto permiten establecer un correcto diagnóstico vocal. En

este diagnóstico no solo es necesario analizar con rigor las características acústicas de la voz sino también debe observarse como utiliza el niño diversos elementos que intervienen en la fonación como la postura corporal, la respiración, la relajación y la movilidad orofacial entre otros. Una correcta exploración profesional debe incluir: la historia clínica, una valoración subjetiva de la discapacidad vocal y una exploración del gesto vocal general y acústica de la voz.

2.3.1 Historia Clínica

El profesional debe conocer en detalle el historial médico del niño, su familia, recoger información sobre la evolución del lenguaje, la evolución psicomotriz, escolar y emocional. La historia clínica en patología vocal infantil también debe incluir la salud general e hipótesis de causalidad de la alteración, ritmo de vida, usos de la voz y hábitos de higiene [Vila, 2009].

2.3.2 Valoración Subjetiva

La valoración subjetiva plantea dificultades en cuanto a la descripción, interpretación y escalas de ciertos parámetros. Hoy día no existe una escala estándar para valorar la voz en lengua española, sin embargo, para la población adulta y en los casos donde el lenguaje de un niño lo permita, la unión europea de foniatras ha propuesto la escala en ingles Grade, Rough, Breathy, Asthenic (GRBAS) de 1981 [Hirano, 1981], [Arias and Estape, 2005]. También existen cuestionarios de calidad de vida relacionados con la voz de los que no existen versiones en español [Vila, 2009]. En España, comunidades de audición y lenguaje recopilan y actualizan periódicamente en Internet y de manera libre, documentos y pruebas para evaluación de lenguaje[1], en un intento de apoyar la falta de información y el trabajo diario de estos profesionales.

2.3.3 Exploración del gesto vocal general

Este examen es tan importante como el de la valoración acústica de la voz, permite que el niño descubra la mecánica fonatoria, el origen de las dificultades y lo que hay que corregir mediante la terapia de voz. Se tienen en cuenta aspectos como:

- **Actitud Vocal**, es decir si el niño muestra una actitud activa y de colaboración o pasiva si se muestra desinteresado para comunicarse.
- **Postura y Verticalidad**, donde se valora la columna vertebral en su parte dorsal y lumbar (plano vertical) y el correcto apoyo de la pelvis y de las extremidades inferiores en el plano horizontal. Se analiza también si hay laxitud o tensión con o sin desplazamiento del tronco hacia adelante.
- **Respiración**, un fenómeno ligado al trabajo corporal donde se valoran los ciclos de inspiración y espiración. Durante la inspiración, se observa si la respiración es por la nariz o por la boca, el tipo de respiración (diafragmático-abdominal, torácica, o con tiraje), y si la inspiración es ruidosa. En la espiración, se analiza el golpe de la glotis, el desplazamiento de la glotis y si se presenta respiración invertida. El análisis de

[1] http://usuarios.multimania.es/maestrosayl/evaluacion-lenguaje.htm

la respiración se complementa con la medición de algunos parámetros aerodinámicos como la espirometría, y el registro de el Tiempo Máximo de Fonación (TMF) y el Tiempo Máximo de Espiración (TME). Cuando el TMF es inferior a los valores de referencia existe la posibilidad de un escape de aire durante la fonación, mientras que si ocurre en el TME, el problema puede deberse a una insuficiencia glótica [Arias and Estape, 2005].

- **Gestos Bucofaciales** Esta valoración examina aspectos como la deglución, el soplo, la masticación, la movilidad y control (praxias) sobre la lengua, labios y mejillas. El correcto control de estos elementos en conjunto permite disminuir el esfuerzo laríngeo y da como resultado un buen timbre de voz.

2.3.4 Valoración acústica de la voz

Es en este apartado donde las tecnologías del habla pueden ser más útiles, no solo en la valoración profesional sino en la educación o reeducación misma de la voz. En general, los profesionales de la voz observan en esta valoración aspectos como la intensidad, el tono, el timbre y la duración. Idealmente debe realizarse de manera objetiva con instrumentos de medida o por medio de herramientas informáticas, la otra manera y de hecho la más utilizada es la subjetiva mediante la valoración perceptual a través del oído.

- **Intensidad.** Se suele registrar la intensidad mínima, conversacional, proyectada y en el grito. Algunos profesionales se interesan por conocer también el rango dinámico definido como la diferencia entre los valores máximos y mínimos. Si se tiene la posibilidad de utilizar instrumentos de medida se utiliza entonces un sonómetro.

- **Tono.** Se registra la frecuencia fundamental mínima, máxima y espontánea. Para obtener el valor numérico de la frecuencia, se utiliza un estroboscopio unido a un fonendoscopio que se ubica en la laringe del niño, el estroboscopio muestra el valor en hercios y luego el profesional busca la equivalencia a notas musicales en tablas para este fin. Otra manera de obtener la equivalencia entre la nota musical y el valor en hercios, consiste en utilizar un piano o teclado eléctrico y perceptualmente el profesional ubica en el teclado el tono más próximo a la emisión vocal del niño [Arias and Estape, 2005].

- **Timbre.** El análisis del timbre de la voz se realiza perceptualmente a través del oído, a nivel laríngeo se evalúa el cierro glótico y la calidad de la vibración de las cuerdas vocales, en el tracto vocal se aprecia la correcta adecuación de los órganos de articulación. Esta valoración se puede hacer con la escala GRBAS o alguna equivalente.

- **Duración.** Se registran los tiempos TMF y TME. También se obtiene el cociente TME/TMF cuyo valor permite valorar la eficiencia del cierre glótico, se consideran valores normales los cercanos a 1 y si es superior a 1.4 se considera un indicador clínico de atención [Vila, 2009].

2.4 Terapia de Voz

El tratamiento de la disfonía infantil puede ser desde el ámbito médico y/o quirúrgico o desde el terapéutico específicamente. Si el tratamiento precisa intervención médica

la rehabilitación terapéutica es igualmente necesaria. En aquellos pacientes cuya voz se encuentra alterada temporalmente, la rehabilitación vocal o reeducación en general trabaja aspectos como: la higiene vocal, relajación, postura, discriminación auditiva, respiración, elementos acústicos de la voz, expresión corporal y comunicación [Bonet, 2009].

En pacientes con discapacidad donde la alteración de la voz se origina o se deriva de la discapacidad misma, la educación vocal es mucho más difícil, se requiere de más tiempo y recursos para lograr resultados no siempre satisfactorios. En este tipo de población la terapia sigue los mismos aspectos descritos con anterioridad pero con las modificaciones necesarias que se requieren al trabajar con discapacidad, esta población demanda un elevado nivel de personalización independiente de si su condición discapacitante es mental o motora.

A continuación se describe someramente en que consiste la terapia de voz.

- **Higiene Vocal**. Consiste en enseñar al niño como cuidar su voz, enseñarle a comunicarse con gestos, miradas o gesticulaciones para que evite gritar o esforzar la voz en tanto esta mejora. También indicarle que evite el humo, el ruido, el polvo y que deje descansar su voz.

- **Relajación**. Se trabaja relajación del cuerpo en general o de forma segmentaria, la idea es trabajar las las partes del cuerpo implicadas en la fonación, como son el cuello, mandíbula, hombros, cara, lengua, labios, mejillas y cabeza. El terapeuta enseña al niño la manera de hacer los ejercicios trabajando los dos frente a un espejo.

- **Postura**. Se busca controlar la postura del cuerpo, tomando conciencia del correcto apoyo del cuerpo en el plano horizontal y en el eje vertical. También es útil que el niño busque y analice diferentes posturas para valorar aquellas que son incorrectas.

- **Discriminación Auditiva**. Es importante que el niño aprenda a escuchar e identificar sonidos y voces para poder discriminarlos según su intensidad, tono, timbre y duración. Igualmente que el niño diferencie una voz sana y conservada de otra con alguna patología.

- **Respiración**. Se busca que el niño tome conciencia de los dos tiempos de la respiración en situaciones con intención vocal o sin ella. También trabajar la salida voluntaria del aire, enseñando al niño a controlar el inicio del soplo espiratorio, su dirección y velocidad. Finalmente que aprenda a regular las presiones glóticas y así controlar el ataque vocal. El terapeuta suele usar velas o globos para trabajar respiración y soplo, siempre sirviendo el mismo de modelo.

- **Trabajo Vocal**. Se busca en general eficacia y calidad vocal. Que el niño regule la respiración y adapte las cavidades de resonancia en función de la intensidad, el tono y la extensión de la secuencia sonora que quiera utilizar. Que ejercite el ataque vocal preciso, y que adquiera destreza en la articulación de sonidos. La terapia suele ser frente a un espejo trabajando praxias de lengua, labios, mejillas y velo del paladar. En la intensidad el terapeuta da un ejemplo al niño con su propia voz para que este le imite, para tonalidad se utiliza un modelo similar o se le pide al niño que repita o lea determinadas palabras y frases donde también se trabaja ataque vocal, articulación y entonación.

- **Expresión Corporal y Comunicación**. Se busca el desarrollo de la voz como vehículo eficiente de comunicación, se trabaja la voz proyectada, la conversación y la voz susurrada. También es útil trabajar entonación en contexto y estados de ánimo.

Las anteriores técnicas no suponen en momento alguno la utilización de tecnologías informáticas como apoyo, por una parte porque sencillamente no ha sido tradición y porque hay pocas herramientas disponibles para lengua española. En su lugar, los recursos audiovisuales tradicionales como vídeos, láminas o cintas de audio, han sido en muchos casos las únicas herramientas disponibles.

Actualmente, con el creciente acceso a las Tecnologías de la Información y la Comunicación (TIC), la principal barrera se presenta en el elevado costo de adquisición de las herramientas disponibles, adicionalmente, estas herramientas no trabajan todos los aspectos deseables en la educación vocal, o no lo hacen de una manera clara para el niño en algunas actividades. Citando de nuevo la población con discapacidad, esta última desventaja tiene mayores implicaciones, además, que una institución de educación especial haga un esfuerzo importante en la adquisición de alguna de estas herramientas, no significa que todos puedan beneficiarsen en igual medida, pues las licencias de uso suelen ser para un solo ordenador.

2.5 Herramientas Informáticas para Terapia de Voz

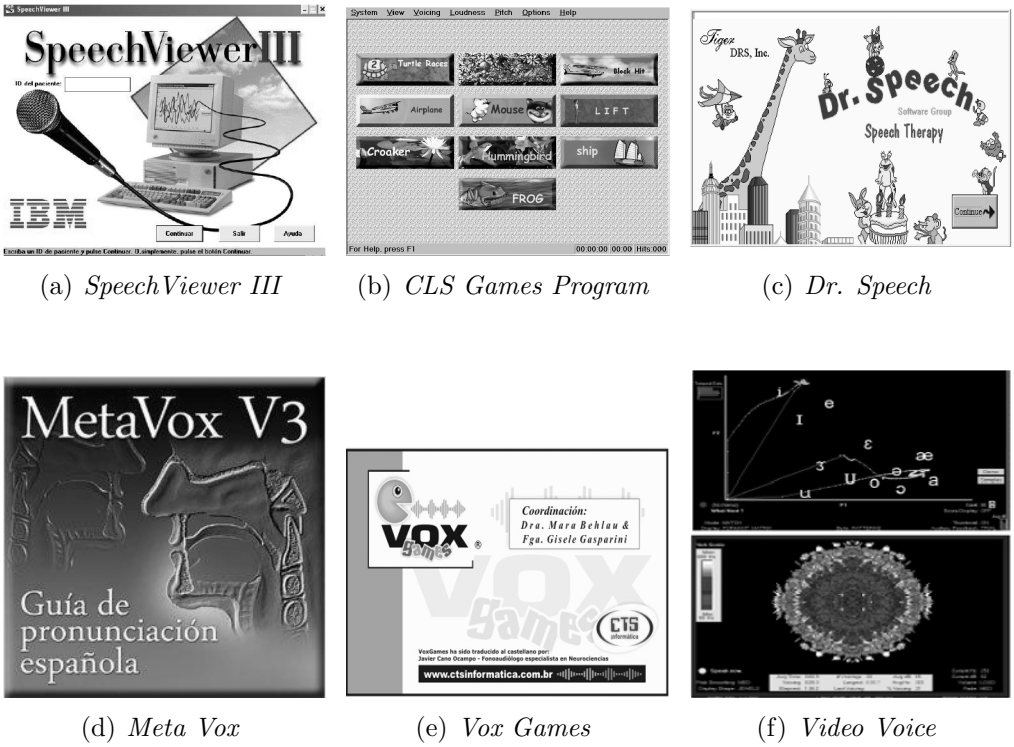
Este apartado describe algunas generalidades de las herramientas disponibles en la actualidad. Es bien sabido que los recursos informáticos presentan las mismas ventajas y beneficios que cualquier otro medio audiovisual, respecto a que potencian y desarrollan los procesos cognitivos básicos en la educación como: atención, percepción, identificación, discriminación, memoria y motivación [Cabero et al., 2008].

La incorporación de las nuevas tecnologías a la intervención en disfonía, suele ir encaminada a la obtención de información relevante de las características acústicas de la voz del paciente. Dichas herramienta trabajan en general cinco aspectos: el ataque vocal, la duración del sonido, el control de la intensidad, el control de tono y la precisión de fonemas [Hurtado and Soto, 2005]. Estas herramientas también han mostrado ser de gran utilidad en casos de personas sordas o hipoacúsicas, ya que en la mayoría de personas sordas las cuerdas vocales son funcionales, pero al no poder oír su propia voz no pueden generar y modular los sonidos adecuadamente [Sánchez, 2002].

2.5.1 Speech Viewer

Esta herramienta como la mostrada en la Figura 2.1(a), es una aplicación desarrollada por IBM[2] dirigida a logopedas, educadores y otros profesionales del área. Con esta aplicación se pueden tratar desordenes de comunicación en diferentes edades, se puede elegir entre: control de tono, intensidad, sonoridad, duración de la voz, análisis de espectros y pronunciación de fonemas. Es una herramienta de pago y actualmente descontinuada y sin soporte.

[2]http://www.axistive.com/speechviewer-iii.html



(a) *SpeechViewer III* (b) *CLS Games Program* (c) *Dr. Speech*

(d) *Meta Vox* (e) *Vox Games* (f) *Video Voice*

Figura 2.1: *Herramientas informáticas para terapia de voz.*

2.5.2 CLS Games Program

Desarrollada por KAY Pentax[3] y mostrada en la Figura 2.1(b), ofrece juegos para el tono, intensidad, y sonoridad, la más reciente versión incluye una nueva actividad para trabajar tiempo de fonación. Es una herramienta de pago para lengua inglesa y que adicionalmente requiere de un hardware externo exclusivo para su funcionamiento.

2.5.3 Speech Theraphy Dr. Speech

Desarrollada por Tiger DRS, Inc.[4] Figura 2.1(c), es un sistema que cuenta con varios juegos interactivos, donde el niño recibe realimentación del cambio de tono, intensidad, tiempo de fonación y ataque vocal. Es una herramienta de pago para lengua inglesa y que permite también trabajar producción vocálica, muestra información visual de la posición de la lengua pero el trabajo vocal del niño se manifiesta únicamente en una imagen con la posición de los formantes vocálicos (F1 vs F2) solo entendible por el logopeda.

2.5.4 Meta Voz

Herramienta desarrollada por Euphonía Ediciones[5] Figura 2.1(d), es una guía interactiva de pronunciación Española. Recrea visualmente todos los elementos lingüísticos que intervienen en la mecánica articulatoria tanto en vocales como en consonantes. Es una herramienta de

[3]http://www.kayelemetrics.com/Product Info/3950/3950.htm
[4]http://www.drspeech.com
[5]http://www.euphoniaediciones.com

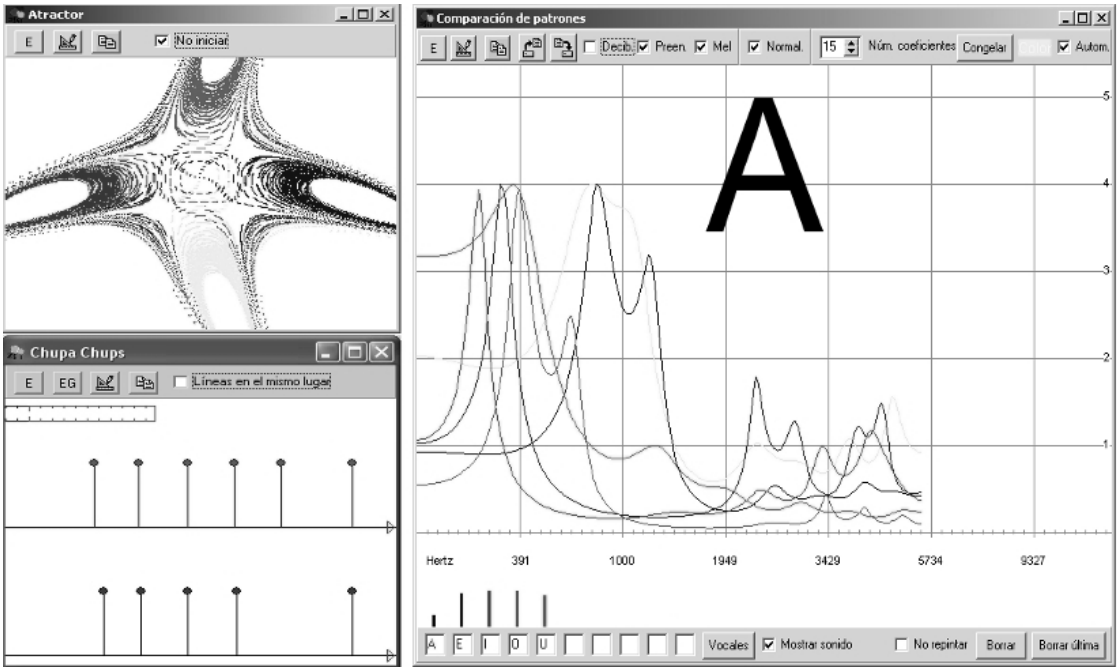


Figura 2.2: *Globus3.*

pago que muestra información pre-grabada en vídeos sin tener en cuenta la producción oral del usuario (funciona sin micrófono), tiene aplicación en logopedia, lingüística y en el aprendizaje del español como segundo idioma.

2.5.5 VoxGames

Desarrollada por CTS Informática[6] en Brasil, es una herramienta para trabajar intensidad, tono, tiempo de fonación, sonidos sordo/sonoro y ataque vocal. Esta herramienta mostrada en la Figura 2.1(e), es de pago y carece de aplicaciones para trabajar articulación vocálica en tiempo real. La herramienta tiene versiones para ingles, portugués y español.

2.5.6 VideoVoice

Desarrollada por Micro Video Corporation[7] USA, es un conjunto de herramientas para terapia de voz diseñadas para niños, que permite trabajar la intensidad, ataque vocal, duración, tonalidad, y posee un apartado especial para producción vocálica en Ingles. Esta última sección muestra información entendible únicamente por el terapeuta. Como se aprecia en la Figura 2.1(f)-arriba, al pronunciar sonidos vocálicos el sistema muestra en pantalla un trazado de puntos que se corresponden a la posición de los formantes, lo que no ofrece mayor información al usuario final. En la parte inferior de la Figura, se puede apreciar otra actividad de esta herramienta donde se trabajar la tonalidad por medio de un caleidoscopio.

[6]http://www.ctsinformatica.com.br
[7]http://www.videovoice.com/

Durante las experiencias recogidas en el transcurso de esta tesis visitando centros y colegios de educación especial es España y Latinoamérica, no se encontró una sola institución que tuviera instalada algunas de las herramientas descritas anteriormente, es evidente que sin considerar la limitante del idioma, la gran dificultad que se presenta es el elevado costo de adquisición de estas herramientas.

Una herramienta de libre distribución y disponible en idioma Español es Globus3[8], perteneciente al proyecto Fresa del Departament d'Educacio de la Generalitat de Catalunya y diseñada por Jordi Lagares. Es un conjunto de herramientas pensado para personas con deficiencias auditivas donde a través de una interfaz gráfica muy sencilla, se puede mostrar gráficamente al usuario los efectos producidos por la intensidad de la voz y presencia/ausencia de sonido, permite también trabajar ritmo y hacer comparación de patrones vocálicos.

La Figura 2.2 muestra en la parte izquierda-arriba, un ejemplo de la imagen mostrada por el sistema ante la presencia de voz, en la parte inferior la actividad para trabajar ritmo alineando figuras verticales y, en la parte derecha, la comparación de patrones vocálicos por medio de líneas espectrales. Esta última herramienta es pues la única para idioma español de libre uso, que permite trabajar algunos aspectos acústicos de la voz. Dentro de las instituciones participantes en esta investigación, solo un centro de educación especial en España conoce y utiliza esporádicamente la herramienta.

[8]http://www.xtec.cat/ jlagares/f2kesp.htm/

Capítulo 3

Técnicas de Procesado de Voz

Este capítulo hace una revisión sobre las técnicas tradicionales de procesado de voz que han servido como punto de partida para esta investigación. Entre ellas encontramos el modelo digital de producción de voz, las técnicas de pre-procesado de la señal de voz, la estimación de energía, la autocorrelación, el análisis de predicción lineal, la estimación de pitch y formantes, y finalmente el análisis homomórfico.

3.1 Sistema Fonador Humano

La generación del habla por parte de los humanos, consiste en la creación de una onda de presión acústica sonora que se propaga a través del aire a una velocidad de unos 340 metros por segundo, se origina voluntariamente a partir de movimientos de la estructura anatómica del sistema humano de producción de voz.

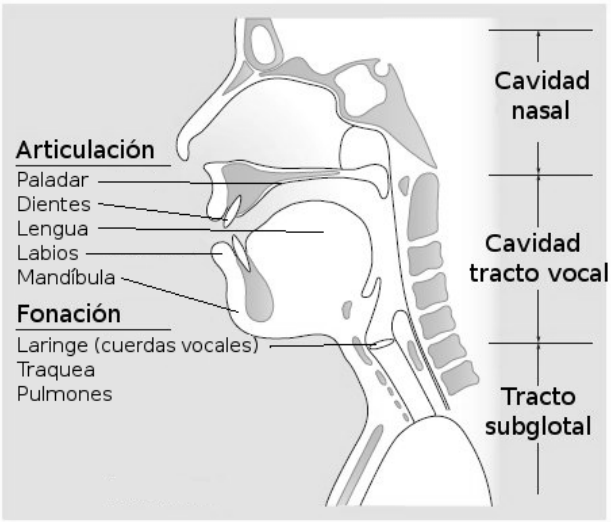


Figura 3.1: *Sistema Humano de Producción de Voz.*

Se destacan dos subsistemas principales el fonatorio y el articulatorio, en el fonatorio los componentes principales son los pulmones, la traquea, la laringe, hasta la región subglotal a la altura de las cuerdas vocales; el sistema articulatorio se compone por el paladar, la

lengua, los dientes, los labios y las mandíbulas. Los distintos sonidos se producen al pasar el aire emitido por los pulmones, a través de todo el sistema de producción en una determinada posición de cada parte articulatoria. La Figura 3.1 (modificada de [j. Benesty, 2008]) muestra un esquema del sistema humano de producción de voz.

Este sistema físico puede modelarse como un filtro lineal todo polos, cuya función de transferencia depende del sonido articulado y, por tanto, de la posición de los diversos órganos involucrados en la producción del habla. La entrada al filtro se puede modelar mediante una señal de excitación, que se corresponde con el paso del aire generado por los pulmones a través de la traquea y las cuerdas vocales [Faúndez, 2000].

Existen dos grandes clasificaciones de los sonidos generados, los sonoros y los sordos o no sonoros; en los sonoros las cuerdas vocales vibran y el aire pasa a través del tracto vocal sin impedimentos importantes, además poseen alta energía y contenido frecuencial en el rango del los 100 Hz a 4000 Hz. En los sordos las cuerdas vocales no vibran y existen restricciones importantes al paso del aire que proviene de los pulmones, tienen baja energía y contenido frecuencial uniforme a manera de ruido blanco. La Figura 3.2 muestra la forma de onda típica para un sonido sonoro y uno sordo, puede verse que su principal diferencia es la presencia de periodicidad en la señal sonora y ningún patrón en la sorda.

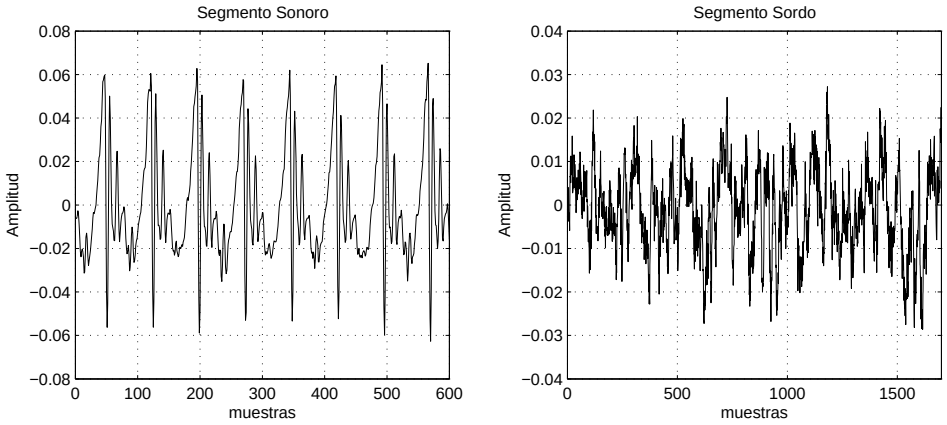


Figura 3.2: *Sonido Sonoro VS Sordo.*

El modelo del filtro que es variable en el tiempo, tiene entonces dos posibles señales de entrada, sonora o sorda. Para señales sonoras la excitación será un tren de impulsos de frecuencia controlada, mientras que para las señales no sonoras la excitación será ruido aleatorio. La combinación de estas señales modelizan el funcionamiento de la glotis. El espectro de frecuencias de la señal vocal puede obtenerse a partir del producto del espectro de la excitación por la respuesta en frecuencia del filtro. Este modelo denominado modelo digital de producción de voz puede apreciarse en la Figura 3.3.

El tracto vocal manifiesta un número grande de resonancias, sin embargo, las importantes son las dos o tres primeras ya que éstas son las contienen más información sobre la producción sonora.

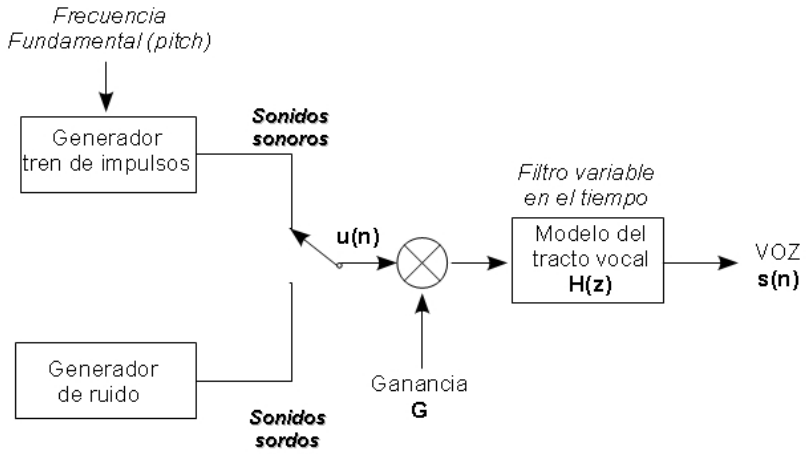


Figura 3.3: *Modelo Digital de Producción de Voz.*

En el campo de procesado de la señal de voz, existen técnicas que permiten extraer de ésta sus parámetros más relevantes. Para hacerlo se hace necesario un pre-procesado de la señal de voz que la adecua para su posterior tratamiento. Es cuando se aplican técnicas como la estimación de la intensidad de la señal voz, el análisis de Predicción Lineal LPC para estimación de pitch y formantes y, el análisis homomórfico entre otras.

3.2 Pre-procesado

El procesado digital de la señal de voz por medio de un ordenador requiere previamente la conversión de la señal acústica en eléctrica mediante un micrófono, y la conversión de la señal analógica resultante en una señal digital para poder procesarla computacionalmente. Para esta conversión es necesario realizar un muestreo o discretización de los valores de la señal cada cierto intervalo de tiempo, denominado periodo de muestreo cuyo inverso es la frecuencia de muestreo. Es importante tener en cuenta que el teorema de Nyquist establece que, para evitar fenómenos de *aliasing* en señales, es necesario muestrear como mínimo al doble de la frecuencia de la señal de entrada para no perder información frecuencial.

Teniendo en cuenta el ancho de banda de la voz humana, donde la información frecuencial se concentra en los 8000 Hz de frecuencia, una frecuencia de muestro de 16000 Hz resulta suficiente para extraer dicha información. La Figura 3.4 muestra en bloques el procesamiento típico realizado por un sistema para extraer parámetros acústicos de la voz. En síntesis, el pre-procesamiento comprende la compensación DC, el pre-énfasis y el enventanado Hamming. A partir de la etapa de pre-procesamiento, se puede hacer análisis en tiempo y frecuencia.

En el análisis temporal, después del pre-énfasis se puede hallar la intensidad de la señal de voz y a partir de allí conocer si el segmento analizado corresponde a un segmento sonoro o sordo. Para el análisis frecuencial se hace necesario preparar la señal para este fin utilizando un enventanado tipo Hamming, es cuando se realiza el análisis LPC y derivado de este será posible conocer la frecuencia fundamental o pitch y las resonancias del tracto vocal o formantes.

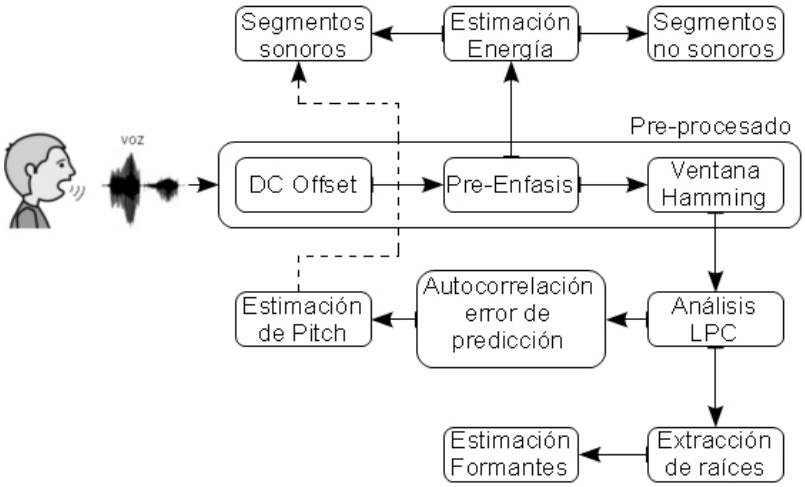


Figura 3.4: *Procesamiento sobre la Señal de Voz.*

El bloque DC offset de la Figura 3.4, elimina posibles componentes DC de la señal aplicándole un filtro de banda eliminada en frecuencia 0 como el descrito en la ecuación 3.1:

$$H(z) = \frac{1 - Z^{-1}}{1 - 0.9995Z^{-1}} \tag{3.1}$$

Después de la compensación DC, la etapa de pre-énfasis se realiza para compensar la caída de -6 dB por octava que experimenta el espectro de la señal de voz por el efecto combinado del pulso glotal y la radiación en los labios. Para esto se usa un filtro digital de primer orden cuya función de transferencia esta descrita por:

$$H(z) = 1 - aZ^{-1}, a = 0,95 \tag{3.2}$$

Ya que la señal de voz tiene un comportamiento pseudo-estacionario solo a corto plazo (decenas de ms), se hace necesario un análisis localizado de la señal durante periodos cortos de tiempo por medio de tramas. El mecanismo que nos permite, dada una señal de voz, realizar un análisis localizado mediante tramas consecutivas se denomina enventanado. Dentro de las ventanas posibles en procesado de voz destaca la tipo Hamming cuya estructura temporal se define como:

$$W(n) = \begin{cases} 0.54 + 0.46 \cdot cos\frac{2\pi n}{(N-1)} & 0 \leq n \leq N-1 \\ \\ 0 & otro\ caso \end{cases} \tag{3.3}$$

La ventana de Hamming tiene un lóbulo ancho cuyo efecto convolutivo producirá un suavizado espectral. La Figura 3.5 muestra el efecto de multiplicar una señal de voz (arriba) por la ventana de Hamming (centro), cuyo resultado mostrado en la parte inferior es el realce de la información central de la ventana y la minimización de la información presente en los extremos, situación que facilita su posterior procesamiento.

Para compensar el efecto de minimización de información en los extremos de la ventana, se suelen tomar ventanas solapadas en las que las muestras de los extremos de una ventana sean las centrales en ventanas consecutivas, de esta manera no se pierde información.

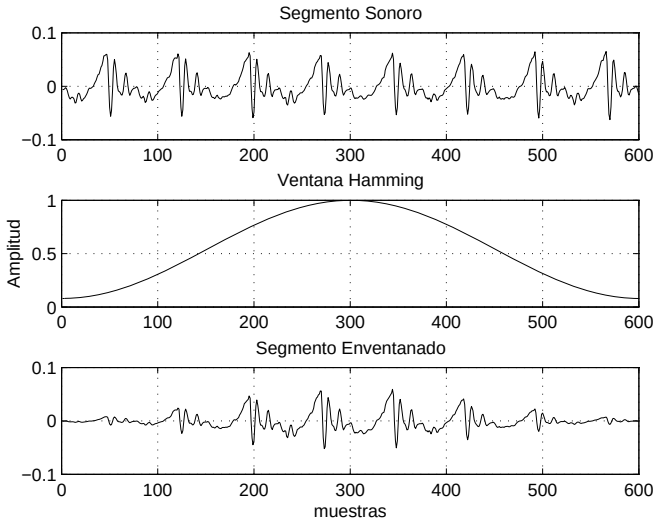


Figura 3.5: *Efecto del enventanado tipo Hamming.*

Después de pasar la señal de voz por la etapa de pre-procesamiento, esta se encuentra lista para ser analizada más fácilmente en etapas posteriores.

3.3 Estimación de Energía

Dentro de las técnicas de análisis localizado en tiempo está la estimación de la energía de la señal. Es una técnica muy útil pues brinda información sobre cambios importantes de amplitud de la señal, permite conocer si las tramas analizadas son sonoras ya que estas son de alta energía, o si corresponden a tramas sordas cuya energía es menor [R. Schafer, 1978].

La energía localizada $E[m]$ para una señal $s(n)$ será:

$$E_s[m] = \sum_{n=-\infty}^{n=\infty} (s[n] \cdot w[n-m])^2 \tag{3.4}$$

$$E_s[m] = \sum_{n=m-N+1}^{m} s^2[n] \cdot w^2[n-m] \tag{3.5}$$

expresando $w^2(n) = h(n)$, $w(n)$ puede quedar:

$$E_s[m] = \sum_{m=n-N+1}^{m} s^2[n] \cdot h[n-m] \tag{3.6}$$

El efecto de enventanado produce entonces una convolución de la energía con un filtro $h[n]$ igual al cuadrado de las muestras de la ventana. En la Figura 3.6 puede apreciarse la evolución de la energía para una señal sonora correspondientes a la vocal /a/.

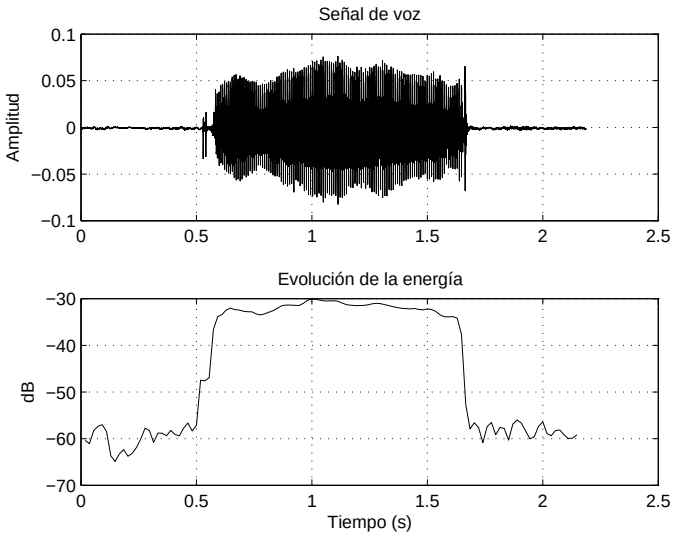


Figura 3.6: *Energía de una Señal sonora.*

3.3.1 Detector de Actividad de voz

La característica de alta energía de los sonidos sonoros es aprovechada por algunos detectores de actividad de voz o Voice Activity Detector (VAD) simples, basados en umbral de energía, para proveer una indicación de la presencia de voz y facilitar el procesamiento de la misma en diversas aplicaciones. Los sistemas VAD basados en umbrales sobre la energía localizada o sobre la relación señal a ruido localizada dan unas prestaciones más que aceptables cuando las condiciones de ruido son altamente estacionarias [R. Schafer, 1978].

Estos sistemas VAD comparan la estimación de la energía del segmento en análisis con un umbral preestablecido, si la energía es mayor que el umbral, el segmento analizado se considera sonoro ya que este posee alta energía; si la energía es menor que el umbral el segmento se considera como no sonoro. El VAD entrega entonces una señal binaria de alto nivel en los segmentos donde hay presencia de voz, y de bajo nivel en los segmentos de silencio.

La Figura 3.7 muestra en la parte inferior una señal de voz con intervalos de silencio, el trazado de la energía y el umbral en la parte central y, en la parte superior de la imagen la señal VAD de activación. La salida del VAD será entonces una señal cuadrada con valor 1 en las tramas de voz y valor 0 en las tramas de silencio.

3.4 Autocorrelación

Con frecuencia es necesario cuantificar el grado de similitud entre varias señales o entre si mismas. Este grado de dependencia o similitud denominado correlación se puede obtener matemáticamente. La correlación existente entre dos señales o correlación cruzada r_{xy} se define como:

$$r_{xy}(k) = E[x(n)y(n-k)] \tag{3.7}$$

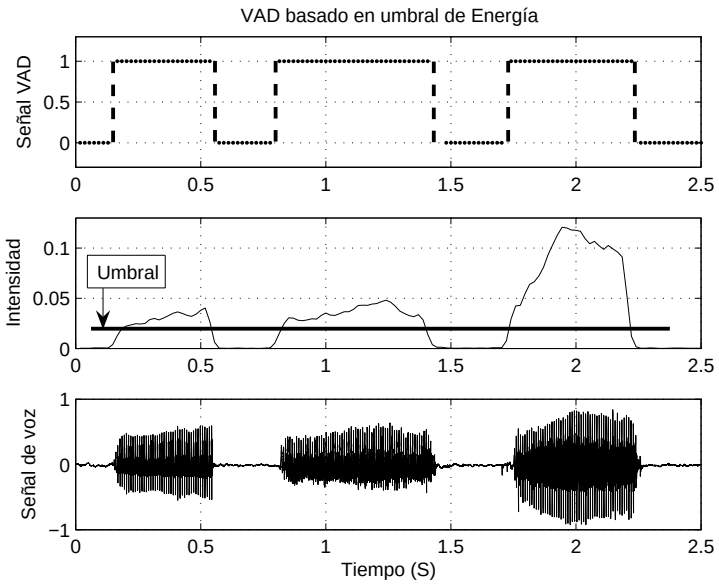


Figura 3.7: *VAD basado en Umbral de Energía.*

En la práctica el operador estadístico de la esperanza matemática $E[\ ]$ se aproxima por el promediado temporal $\sum_{n=-\infty}^{\infty}$, de esta forma la autocorrelación la estimaremos como:

$$r_{xy}(k) = \sum_{n=-\infty}^{\infty} [x(n)y(n-k)] \tag{3.8}$$

De esta manera valores grandes y positivos indicarían que ambas señales son parecidas y crecen a la vez, y valores negativos indican que el crecimiento de una variable esta asociado con el decrecimiento de la otra. Por otra parte valores próximos a cero indican que las señales no tienen parecido [E. Soria, 2003].

Un caso de correlación cruzada es la autocorrelación, cuando las secuencias $x(n)$ e $y(n)$ coinciden. Particularizando la ecuación 3.8 tendremos:

$$r_x(k) = \sum_{n=-\infty}^{\infty} x(n)x(n-k) \tag{3.9}$$

Para señales periódicas se verificará que para valores de desplazamiento k iguales al periodo de la señal, la autocorrelación tendrá un máximo local, por lo que la autocorrelación de señales periódicas será también una señal periódica del mismo periodo, así pues la autocorrelación se puede emplear para detectar la frecuencia fundamental de señales sonoras. La Figura 3.8 muestra la forma de onda de una señal sonora y una señal sorda con sus respectivas autocorrelaciones.

En la parte superior de la imagen se aprecia la señal sonora y su autocorrelación, allí se observa claramente la periodicidad de ambas señales en contraste con las formas de onda de la parte inferior, donde las señales carecen de algún patrón o periodicidad. La longitud de la ventana en muestras debe ser lo suficientemente grande para que haya varios periodos de

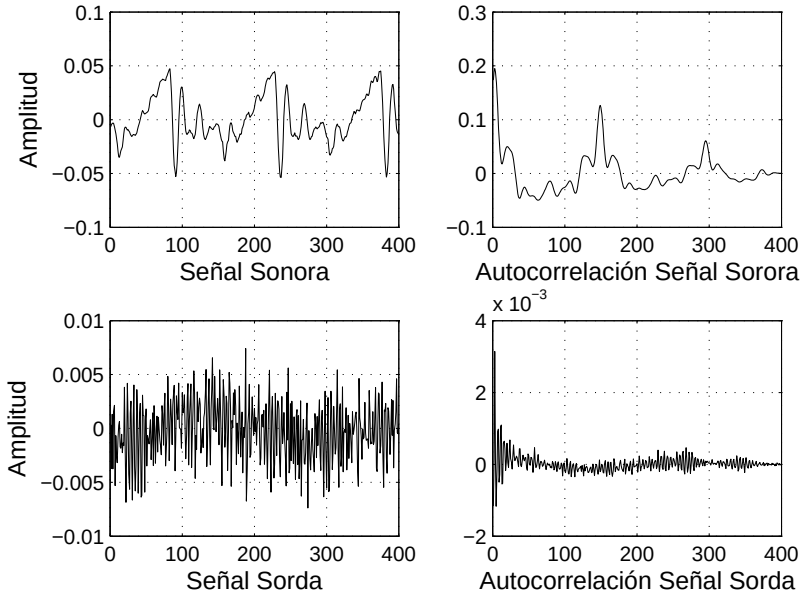


Figura 3.8: *Autocorrelación de una Señal Sonora y Sorda utilizando una ventana rectangular con N=400.*

la señal dentro y lo suficientemente pequeña para que no haya variaciones del pitch dentro de la ventana misma.

3.5 Análisis de Predicción Lineal LPC

Una de las herramientas más poderosas para analizar la voz es el método de codificación por predicción lineal o Linear Prediction Coefficients (LPC). Este método se ha convertido en la técnica predominante para analizar parámetros básicos de la voz como: pitch, formantes y área del tracto vocal con estimaciones muy aproximadas y relativo bajo costo computacional.

La idea básica del análisis de predicción lineal es expresar o predecir una señal en un instante determinado, como una combinación lineal de muestras en instantes anteriores, minimizando el error cometido entre la señal original y la predicha [R. Schafer, 1978]. La filosofía de la predicción lineal esta íntimamente relacionada con el modelo de producción de voz de la Figura 3.3.

En la técnica de análisis LPC, la muestra actual es aproximada o predicha mediante una combinación lineal de muestras anteriores así:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{3.10}$$

el error cometido en la aproximación será:

$$e(n) = s(n) - \hat{s}(n) \tag{3.11}$$

donde a_k son los coeficientes de predicción lineal, $s(n)$ es la señal real, $\hat{s}(n)$ la señal predicha y $e(n)$ el error de predicción lineal (error residual o residuo). Podemos expresar el error cometido como:

$$e(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{3.12}$$

tomando transformada Z tendremos:

$$E(z) = S(z)[1 + \sum_{k=1}^{p} a_k z^{-k}] \tag{3.13}$$

denominando $A(z)$ a la expresión:

$$A(z) = \frac{E(z)}{S(z)} = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{3.14}$$

podemos decir que $A(z)$ al que llamaremos filtro inverso, será la función de transferencia de un sistema como el mostrado en la Figura 3.9 parte izquierda y el inverso de éste será un sistema como el mostrado en la parte derecha.

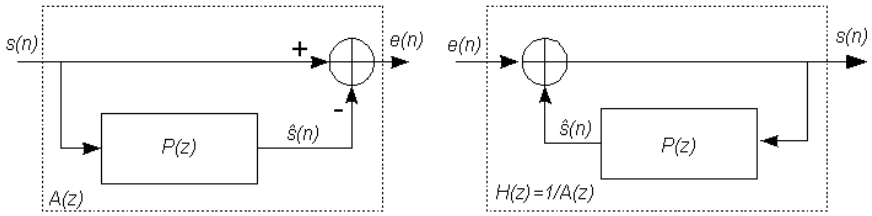


Figura 3.9: *Filtro $A(z)$ y su Inverso.*

Excitando con el error de predicción un sistema cuya función de transferencia sea $1/A(z)$, obtendremos a la salida la señal deseada de voz $s(n)$. En el modelo simplificado de producción, podemos asumir que $H(z)$ sigue un modelo todo-polos con q polos, es decir:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{k=1}^{q} a_k Z^{-k}} \tag{3.15}$$

Identificando $H(z)$ con el filtro de predicción lineal $1/A(z)$ y asumiendo que el número de polos del modelo es igual al orden de predicción lineal, $p = q$, tendremos que:

$$H(z) = \frac{G}{A(z)} \tag{3.16}$$

donde G es la ganancia y que $E(z) = GU(z)$.

Como consecuencia de este análisis será posible: obtener la función de transferencia del filtro equivalente calculando los coeficientes, conocer si el segmento es sonoro o sordo, estimar el filtro $H(z)$ para poder producir la señal deseada, además si los parámetros del tracto vocal representados por $H(z)$ se pueden modelar mediante $1/A(z)$, entonces el error

de predicción representará la excitación.

Debemos encontrar por consiguiente, un conjunto de parámetros a_k que minimicen el error de predicción cuadrático medio en cada trama de análisis. Utilizando la notación $s_n(m) = s(n+m)$ y $e_n(m) = e(n+m)$ la expresión del error cuadrático medio será:

$$E_n = \sum_m e_n^2(m) \tag{3.17}$$

y desarrollando el error de predicción tenemos:

$$E_n = \sum_m (s_n(m) + \sum_{k=1}^{p} a_k s_n(m-k))^2 \tag{3.18}$$

Con el objeto de minimizar el error de predicción respecto al conjunto de parámetros a_k, tendremos que derivar E_n parcialmente respecto a cada coeficiente a_k e igualar a cero, esto es:

$$\frac{\partial E_n}{\partial a_k} = 0, k = 1, 2, ..., p \tag{3.19}$$

Resultando así:

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} \hat{a_k} \sum_m s_n(m-i)s_n(m-k) \tag{3.20}$$

Teniendo en cuenta la expresión de la covarianza localizada:

$$\Phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k) \tag{3.21}$$

Por tanto, podemos expresar de forma compacta los coeficientes óptimos:

$$\Phi_n(i,0) = \sum_{k=1}^{p} \hat{a_k} \Phi_n(i,k) \tag{3.22}$$

lo que constituye un conjunto de p ecuaciones con p incógnitas.

Para la resolución del sistema de ecuaciones lineales planteado, debemos calcular $\Phi_n(i,k)$ para $1 \leq i \leq p$, $0 \leq k \leq p$. Un método clásico para la resolución es el método de la autocorrelación, que permite una resolución recursiva y además exige poca carga computacional.

Supongamos que la señal de voz $s_n(m)$ se anula en el intervalo $0 \leq m \leq N-1$, esto equivale a asumir que la señal $s_n(m)$ ha sido multiplicada en el tiempo por una ventana $w(m)$ que vale cero fuera del intervalo: $0 \leq m \leq N-1$

De este modo, podemos expresar la señal de voz como:

$$s_n(m) = s(n+m) \cdot w(m), 0 \leq m \leq N-1 \tag{3.23}$$

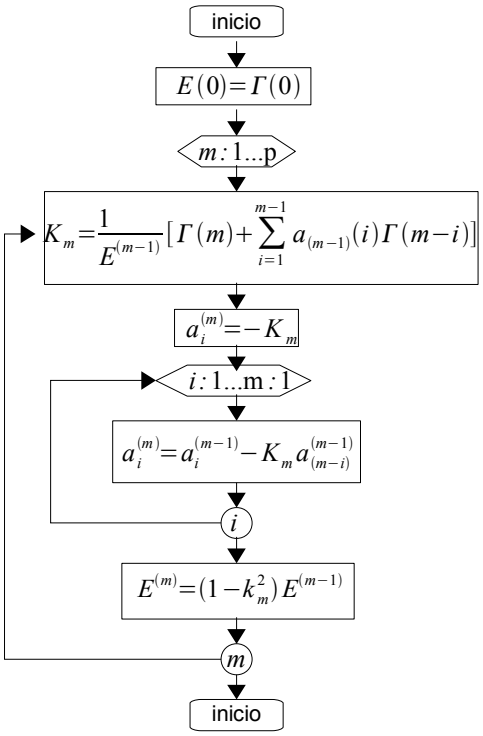


Figura 3.10: *Algoritmo de Levinson-Durbin.*

Así, el error cuadrático medio será distinto de cero en el intervalo $0 \leq m \leq N-1+p$, por lo que podremos expresarlo como:

$$E_n = \sum_{m=0}^{N-1+p} e_n^2(m) \tag{3.24}$$

y la expresión de la covarianza localizada quedará:

$$\Phi_n(i,k) = \sum_{m=0}^{N-1+p} s_n(m-i) \cdot s_n(m-k), 1 \leq i \leq p, 0 \leq k \leq p \tag{3.25}$$

o también

$$\Phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m) \cdot s_n(m+i-k), 1 \leq i \leq p, 0 \leq k \leq p \tag{3.26}$$

Esta última expresión es sólo función de $i-k$, por lo que $\Phi_n(i,k)$ se reduce sencillamente a la expresión de la función de autocorrelación:

$$\Phi_n(i,k) = \Gamma_n(i-k) \tag{3.27}$$

$$\Phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m) \cdot s_n(m+i-k), 1 \leq i \leq p, 0 \leq k \leq p \tag{3.28}$$

y, puesto que la función de autocorrelación es simétrica, es decir $\Gamma_n(-k) = \Gamma_n(k)$, las ecuaciones se pueden expresar como:

$$\sum_{k=1}^{p} \Gamma_n(i-k)\hat{a_k} = \Gamma_n(i), i = 1, ..., p \tag{3.29}$$

expresión que en forma matricial resulta:

$$\begin{bmatrix} \Gamma_n(0) & \Gamma_n(1) & \Gamma_n(2) & \dots & \Gamma_n(p-1) \\ \Gamma_n(1) & \Gamma_n(0) & \Gamma_n(1) & \dots & \Gamma_n(p-2) \\ \Gamma_n(2) & \Gamma_n(1) & \Gamma_n(0) & \dots & \Gamma_n(p-3) \\ \vdots & & & \vdots & \\ \Gamma_n(p-1) & \Gamma_n(p-2) & \Gamma_n(p-3) & \dots & \Gamma_n(0) \end{bmatrix} \cdot \begin{bmatrix} \hat{a_1} \\ \hat{a_2} \\ \hat{a_3} \\ \vdots \\ \hat{a_p} \end{bmatrix} = \begin{bmatrix} \Gamma_n(1) \\ \Gamma_n(2) \\ \Gamma_n(3) \\ \vdots \\ \Gamma_n(p) \end{bmatrix} \tag{3.30}$$

La matriz de autocorrelación, de dimensión p x p, es una matriz tipo Toeplitz (simétrica, con las diagonales principal y secundarias de elementos iguales). Debido a este tipo de matriz resultante, este conjunto de ecuaciones se puede resolver de forma recursiva utilizando el algoritmo de Levinson-Durbin.

Este algoritmo que se aprecia en la Figura 3.10 nos entrega los coeficientes de predicción a_i y los coeficientes de reflexión del tracto vocal.

Se inicializa el algoritmo con $E(0) = \Gamma_n(0)$ y de forma recursiva, para $m = 1, 2, ..., p$ tenemos:

$$k_m = \frac{\Gamma(m) - \sum_{i=1}^{m-1} a_{m-1}(i) \cdot \Gamma_n(m-i)}{E_m - 1} \tag{3.31}$$

Las soluciones parciales, para $m < p$ permitirán calcular los coeficientes óptimos del filtro $H(z)$ de orden m. La solución final buscada, para $m = p$, dará como resultado los coeficientes óptimos del filtro de orden p, esto es:

$$\hat{a_i} = a_p(i), 1 \leq i \leq p \tag{3.32}$$

El orden de predicción p controla el número de polos con el que modelamos la envolvente espectral y según crece p, aumenta el detalle del modelo, por lo que su elección varía en función de la aplicación. Para el caso de la información formántica de la voz se suele utilizar un par de polos complejos conjugados por cada formante.

3.6 Estimación de Pitch

El análisis LPC permite la extracción de la señal de error o residual a partir de una trama de voz. Esta señal residual, que se corresponde con la excitación vocal, permite una estimación de la frecuencia fundamental o pitch (dominio de la correlación) y de la estructura armónica *blanqueada* (dominio espectral) de alta precisión. Para obtener la señal residual o de error $e(n)$ es necesario multiplicar la señal de voz $s(n)$ por un filtro FIR todo polos $A(z)$ según el esquema de la Figura 3.9-izquierda.

Una vez se tiene la señal de error $e(n)$, y tomando la autocorrelación de esta es posible calcular la frecuencia de pitch, ya que la distancia entre el origen y el primer pico que aparece

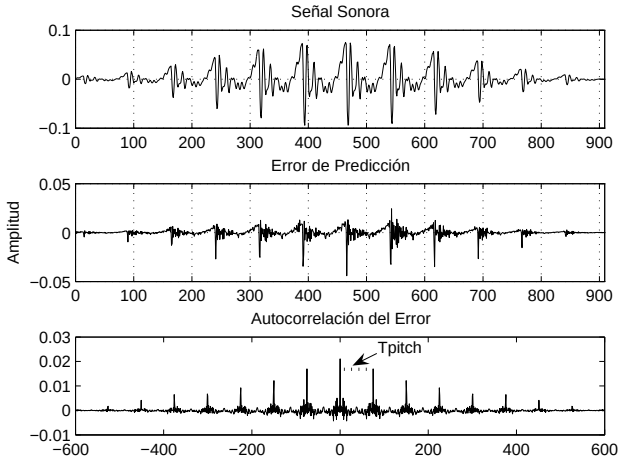


Figura 3.11: *Estimación de Pitch por Análisis LPC.*

en la señal de autocorrelación corresponde al periodo de pitch.

La parte superior de la Figura 3.11 muestra una señal sonora y su error de predicción obtenido del análisis LPC en la parte central, la parte inferior muestra la autocorrelación de la señal de error donde se aprecia claramente la periodicidad de la señal. La distancia entre el origen y el primer pico corresponderá entonces al periodo de pitch $Tpitch$ de cuyo inverso obtenemos la frecuencia de pitch $Fpitch$.

Como etapa de pos-procesado tenemos un filtrado de mediana para corregir errores en la estimación de frecuencia de pitch. El proceso consiste en ordenar los valores estimados y tomar el que queda en el medio (si el número de datos es impar). Si el número de datos es par, se elige la media de los dos datos centrales. Así por ejemplo un filtro de mediana de orden 5, ordena 5 valores consecutivos y selecciona el tercero y luego avanza una posición; si hubiesen datos espurios, el ordenamiento de los valores los dejaría en los extremos, por lo que con la selección del valor central no se tendrían en cuenta.

La Figura 3.12 resume el proceso para obtener el frecuencia fundamental o pitch. Ante una señal sonora el sistema adecua la señal de voz en la etapa de pre-procesamiento, realiza el análisis LPC para obtener el error de predicción, de la autocorrelación de este obtenemos el periodo de pitch y con su inverso la frecuencia de pitch. En la parte izquierda de la imagen puede observarse la estimación inicial de la frecuencia de pitch con algunas estimaciones espurias, y, en la parte derecha, el resultado de aplicar un filtro de mediana de orden=5 a dichas estimaciones, puede apreciarse el suavizado ante los espurios inicialmente estimados.

3.7 Estimación de Formantes

Ya que la técnica de predicción lineal permite separar la influencia del tracto vocal de la señal de excitación, podemos centrarnos en la estructura formántica de la señal hablada. Según la Ecuación 3.14, $A(z)$ es un polinomio que puede ser representado en términos de

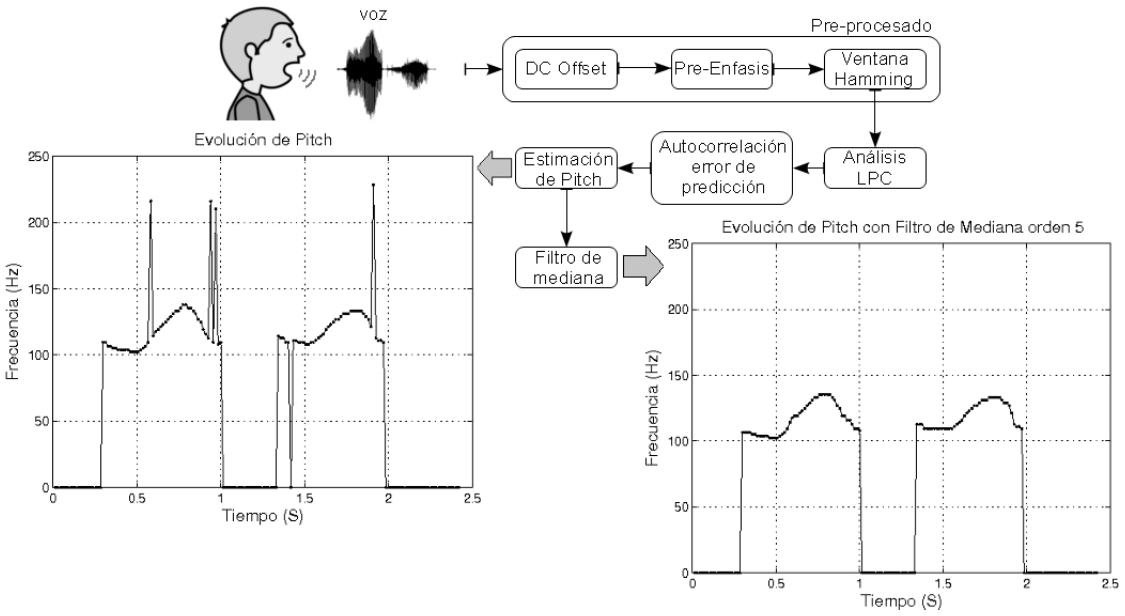


Figura 3.12: *Proceso de Estimación de Pitch con Filtro de Mediana.*

ceros como:

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i} = \prod (1 - z_k z^{-}1) \tag{3.33}$$

Y de acuerdo a la Ecuación 3.16 los ceros de $A(z)$ son los polos de $H(z)$. De manera que con un orden de predicción adecuado se puede esperar que aproximadamente $\frac{Fs}{1000}$ de las raíces estarán cerca en frecuencia a las frecuencias de resonancia (en el plano z), siendo Fs la frecuencia de muestreo en Hz. Es decir, las raíces (pares complejos conjugados) que están cerca de la circunferencia de radio unidad, son los polos de $H(z)$ que modelan los formantes. [Rabiner and Shafer, 2007].

Tomando entonces los coeficientes de predicción a_i, se pueden hallar las raíces del polinomio, convertirlos a frecuencia analógica y ordenarlos de menor a mayor; los tres primeros valores corresponderán a los tres primeros formantes de la trama de análisis.

La Figura 3.13 muestra en la parte izquierda el espectro y la envolvente LPC para una trama sonora, los tres formantes F1, F2 y F3 corresponden a los tres polos cercanos a la circunferencia de radio unidad en el plazo z como muestra la misma figura en la parte derecha.

La Figura 3.14 resume el proceso de estimación de formantes para las cinco vocales del español. Ante la emisión sonora, el sistema adecua la señal de voz en la etapa de pre-procesamiento, realiza el análisis LPC y se obtienen las raíces de los coeficientes de filtro, finalmente se convierten a frecuencia analógica y después de ordenarlos ascendentemente se obtienen los formantes.

Dibujando el Formante 1 contra el Formante 2 se obtiene lo que se conoce como el triángulo vocálico. Debido a que las resonancias del tracto vocal o formantes dependen principalmente de las condiciones geométricas de este, resulta difícil establecer valores

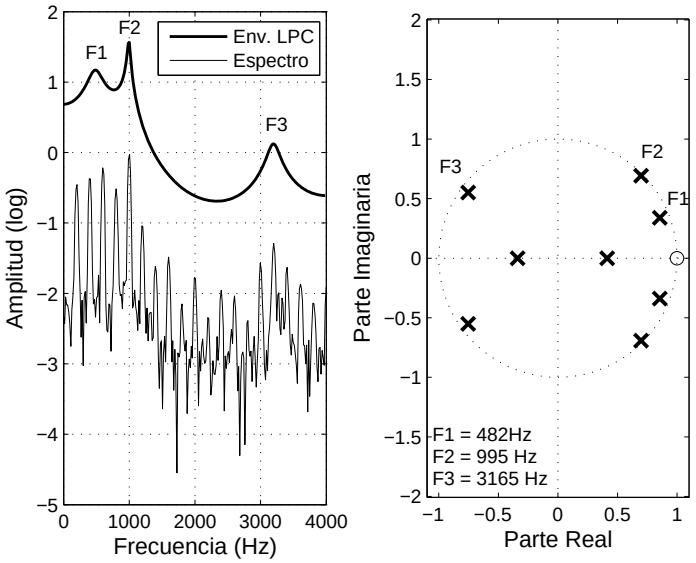


Figura 3.13: *Formantes y envolvente espectral para una /a/ sonora*

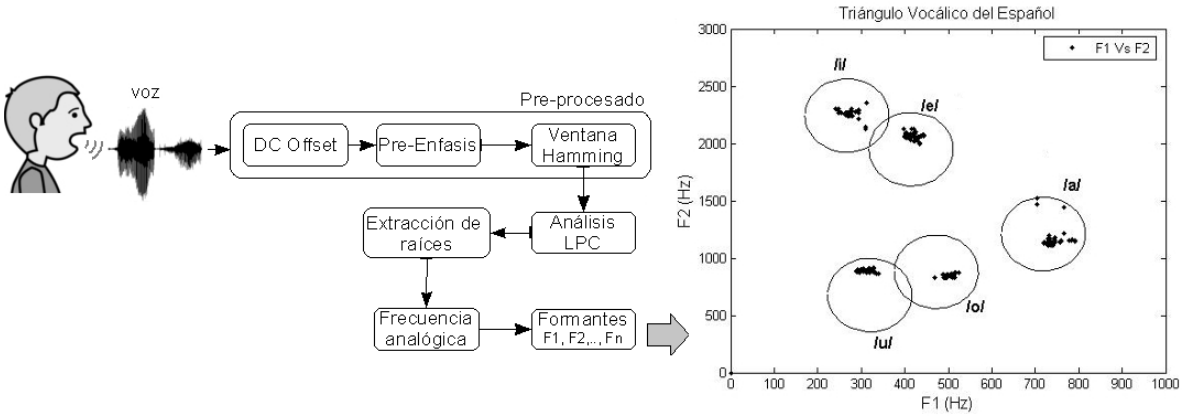


Figura 3.14: *Proceso de Estimación de Formantes.*

estándar de dichos formantes, lo que ha motivado en gran medida la presente tesis en busca de un método para normalizar dichas estimaciones. Los formantes estimados en el triángulo en la figura 3.14 corresponden pues a un adulto varón de 33 años de edad.

3.8 Análisis Homomórfico

Un homomorfismo consiste en convertir un elemento matemático en otro, por ejemplo convertir una convolución en una suma o viceversa, es una herramienta que resulta especialmente útil en el tratamiento de la voz. La tarea requerida será entonces la deconvolución de un segmento de voz $s[n]$ en una componente que representa la señal de excitación $e[n]$ y una componente que representa la respuesta impulsional del tracto vocal $h[n]$, es decir,

$$s[n] = e[n] * h[n] \Longrightarrow \hat{s}[n] = \hat{e}[n] + \hat{h}[n] \tag{3.34}$$

Esta separación no se puede conseguir mediante filtrado ya que ambas componentes no están combinadas linealmente en el dominio temporal. El análisis cepstral permite obtener una representación de la señal de voz en un dominio donde la excitación $\hat{e}[n]$ y el modelo del tracto vocal $\hat{h}[n]$ se combinan linealmente y aparecen separadas.

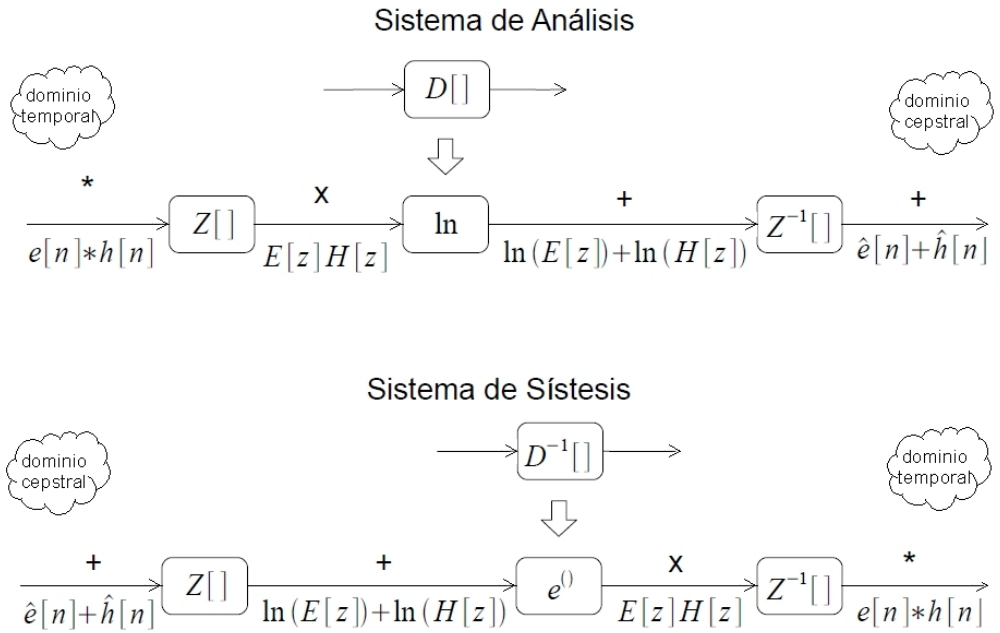


Figura 3.15: *Análisis Homomórfico.*

Como muestra la Figura 3.15 en la parte superior, en el análisis cepstral, D transforma la señal de entrada a un espacio tal que la convolución se convierta en producto, por ejemplo a través de la transformada Z o la transformada de Fourier, luego aplicando el homomorfismo del logaritmo el producto se convierte en sumas, de manera que si se aplica la transformada inversa, se obtienen unas secuencias discretas en el dominio cepstral que se relacionarán a través de la suma y no a través del producto como antes de la transformación.

El sistema inverso o de síntesis D^{-1} mostrado en la parte inferior de la Figura 3.15, devuelve las secuencias al dominio temporal. Consiste en aplicar la misma secuencia de pasos pero en sentido inverso, primero estando en el dominio cepstral se aplica transformada Z directa para obtener los logaritmos sumados, luego aplicando la operación inversa al logaritmo es decir la exponencial, se recupera la relación a través de productos de la excitación y el filtro, de nuevo aplicando transformada Z inversa se obtiene la relación entre las dos secuencias en el dominio temporal de la excitación y respuesta impulsional del tracto vocal.

La definición formal de cepstrum complejo, es decir cuyo resultado es una secuencia de valores complejos en el caso general, es la transformada inversa de Fourier del logaritmo del módulo de la transformada de Fourier de la señal original, es decir:

$$\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} ln\left[X(e^{jw})\right] e^{jwn} dw \tag{3.35}$$

y teniendo en cuenta la definición de logaritmos para números complejos, en donde son fácilmente diferenciables las partes del logaritmo:

$$ln\left[X(e^{jw})\right] = ln\left|X(e^{jw})\right| + j\ arg\left[X(e^{jw})\right] \tag{3.36}$$

$\hat{x}[n]$ será:

$$\hat{x}[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} ln\left|X(e^{jw})\right| e^{jwn} dw + \frac{j}{2\pi}\int_{-\pi}^{\pi} ln\left[X(e^{jw})\right] e^{jwn} dw \tag{3.37}$$

donde el primer sumando es completamente real, y la parte restante, y gracias a la simetría impar de la fase de la transformada de Fourier para señales reales, el cepstrum complejo de una señal real es real. Del primer sumando de la expresión anterior surge la definición de cepstrum real, el cual de define como la anti-transformada de Fourier del logaritmo del módulo de la transformada de Fourier, aunque se debe tener en cuenta que el cepstrum real no es la parte real del cepstrum complejo, simplemente son dos definiciones diferentes, en el cepstrum real solo interviene el módulo.

De manera que el cepstrum real se define como:

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} ln\left|X(e^{jw})\right| e^{jwn} dw \tag{3.38}$$

el cepstrum real también puede verse como la parte par del cepstrum complejo:

$$c[n] = \frac{\hat{x}[n] + \hat{x}[-n]}{2} \tag{3.39}$$

El cepstrum complejo contiene información de la magnitud y fase del espectro inicial, por lo que la señal se puede reconstruir completamente, mientras que el cepstrum real solo utiliza la información de la magnitud del espectro. Si el sistema considerado no tiene polos y ceros fuera de la circunferencia de radio unidad, es decir un sistema de fase mínima, el cepstrum toma valores de cero para índices negativos y toma valores distintos de cero para índices positivos, es decir, en sistemas de fase mínima el cepstrum complejo esta determinado de forma unívoca por el cepstrum real y decae a razón de $1/n$ donde la mayor parte de la información se concentra en el origen [R. Schafer, 1978].

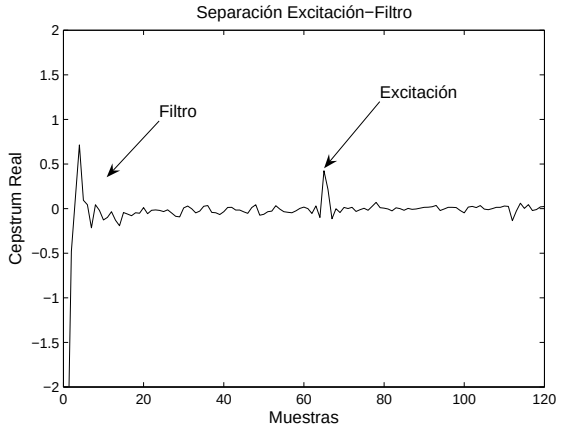


Figura 3.16: *Separación en el Dominio Cepstral.*

La Figura 3.16 muestra el cepstrum real para una trama sonora, allí, la parte baja de $c[n]$ corresponde entonces a la información del tracto vocal (filtro), mientras que la parte alta se debe principalmente a la excitación. De esta manera, la representación de la excitación y el filtro en el dominio cepstral permite una fácil separación de estas componentes por medio de un liftado, que es un proceso equivalente al filtrado pero realizado en el dominio cepstral.

Considerando una señal de voz, su cepstrum real discreto $c[n]$ queda definido entonces por la ecuación:

$$c[n] = \frac{1}{N}\sum_{k=0}^{N-1} ln\,|X(k)|\,e^{j\frac{2\pi}{N}kn}, 0 \leq n \leq N-1 \tag{3.40}$$

donde $X(k)$ es la transformada de Fourier de N puntos la señal de voz.

Parte II

Base Experimental e Investigación

Capítulo 4

Entidades de Colaboración y Corpus

Esta investigación surge de la colaboración existente entre el Grupo de Tecnologías de las Comunicaciones (GTC) de la Universidad de Zaragoza y el centro de Educación Especial *Alborada* en Zaragoza. Ellos manifestaron al grupo de investigación la sentida necesidad de poder disponer de herramientas libres para trabajar la voz y aspectos prelingüísticos en los niños que asisten a dicha institución.

Es así como la investigación se inició trabajando sobre aspectos prelingüísticos como la intensidad y tonalidad. Al llegar a la etapa de vocalización se presento una gran dificultad al no poseer bases de datos con voz infantil que mostraran el entorno real de trabajo y explicaran las dificultades técnicas hasta ese momento encontradas. Para poder continuar con la investigación se planteo como solución adquirir un corpus de voz infantil no alterada que permitiera experimentar, diseñar, y personalizar algoritmos para el tratamiento de este tipo de voz, y finalmente, poderlos aplicar en casos de usuarios con alteraciones en su voz.

En el transcurso de la investigación también se establecieron otros convenios de colaboración que se describen en la Sección 4.1, y que posibilitaron la aplicación y evaluación de la tecnología propuesta. En la Sección 4.2 se describe el corpus de voz infantil no alterada adquirido, los requerimientos, entorno de la adquisición, y finalmente las características de los locutores.

4.1 Entidades de Colaboración

Colegio Público de E.E.
ALBORADA
Zaragoza

Si en este momento estas líneas pueden ser leídas se debe en parte a la colaboración del personal del Colegio Público de Educación Especial (CPEE) **Alborada**. Esta institución viene trabajando en los últimos años con los diferentes grupos de investigación del Instituto de Investigación en Ingeniería de Aragón (I3A) en diferentes proyectos de aplicación de la investigación a la discapacidad [Falcó et al., 2006, Negre, 2005, Negre et al., 2006, Martínez et al., 2007, Vaquero, 2006] y tiene una gran experiencia en el desarrollo de ayudas técnicas para educación especial.

Una muestra de esta colaboración es la tesis doctoral denominada: *Personalización y Adaptación On-line a Trastornos y Variaciones de la Voz en Sistemas de Reconocimiento Automático del Habla* del Dr. Oscar Saz Torralba. En dicha investigación se propone el uso de técnicas de personalización para mejorar los resultados de los sistemas de reconocimiento automático del habla o RAH, en tareas propias de terapia del habla alterada.

Por otro lado en la presente tesis, el grupo de profesionales de la **Alborada** participó desde el principio en contextualizar la investigación dentro de la educación especial y la logopedia. Facilito las periódicas visitas al centro para probar y discutir los avances obtenidos en cada etapa, y algunos de sus estudiantes participaron en el estudio final de aplicación de las herramientas que se trata en detalle en la Sección 9.4.

El equipo de **Alborada** gestionó también el acceso al colegio de educación infantil y primaria **Rio Ebro**, lugar donde se llevo a cabo la grabación del corpus de voz infantil no alterada. También se contó con el apoyo del Institución de Educación Secundaria (IES) **Elaios** para la grabación del corpus con locutores adolescentes.

En el trascurso de la investigación se establecieron también diferentes convenios de colaboración con entidades dedicadas a la educación especial en Latinoamérica. Es el caso de el **Centro de Enseñanza Especial y Rehabilitación de Alajuela** en Costa Rica, quién realizó una rigurosa evaluación de la primera versión de la herramienta llamada *PreLingua* (descrita en la Sección 7.1) desde un punto de vista como profesionales experimentados en educación especial, y teniendo en cuenta también las necesidades de esta región Centroamericana. Sus valiosos aportes permitieron ampliar y mejorar la siguiente versión de la herramienta.

En el último año de investigación se contó con el valioso apoyo de la fundación **Centro de Educación Especial del Niño Diferente, CEDESNID** en Bogotá Colombia. Es una entidad sin ánimo de lucro que brinda intervención terapéutica en todos los niveles a personas con discapacidad. Ha aportado sus experiencias y apoyado la investigación con sus profesionales en fonoaudiología y participado activamente en el estudio de aplicación de la herramienta *PreLingua*. Es el centro con mayor número de usuarios participantes en el estudio.

4.2 Corpus de Voz Infantil no Alterada

Como en todas las tareas de investigación en Tecnologías del Habla, y especialmente cuando se trabaja en una situación tan específica, se depende mucho de la existencia de bases de datos que reflejen las características de la tarea. Existen algunos corpus para investigación en habla disártrica [Menéndez-Pidal et al., 1996, Green et al., 2003, Hawley et al., 2003] u otros tipos de hablas alteradas [Navarro-Mesa et al., 2005], pero que no son totalmente útiles para esta investigación; bien por estar adquiridas en inglés o bien porque no están diseñadas para los estudios que se pretenden llevar a cabo en este trabajo, es decir un corpus con información vocálica específica.

4.2.1 Requerimientos de la Adquisición

Trabajar articulación vocálica en población infantil requiere conocer con buen grado de detalle como cambian las resonancias del tracto vocal en la producción vocálica a medida que un individuo crece. Es bien sabido que estas resonancias cambian no solo por las condiciones geométricas del tracto vocal donde afecta mucho el crecimiento, sino también por factores como el sexo, la talla y la raza entre otros. Para la investigación era entonces necesario contar con un corpus con emisiones vocálicas del idioma español generadas de manera aislada y sostenida, estas emisiones debían abarcar en lo posible un rango de edades tal que permitiera conocer la evolución formántica desde la infancia a la adolescencia.

Otro aspecto importante a considerar fue que los locutores tuviesen voz sin alteraciones para poder estudiarla y conocerla mejor desde un punto de vista científico y así poder enfrentarse posteriormente a voces alteradas. Este corpus se diseñó teniendo en cuenta los recursos humanos disponibles en ese momento en el colegio de educación infantil y primaria **Rio Ebro**. Esta institución cuenta con alumnos en la etapa infantil con edades entre los 4 y 6 años y la etapa primaria con edades entre los 6 y 12 años. Es evidente que después de los 12 años las estructuras fonatorias siguen cambiando y por ende siguen variando la información formántica, de manera que gracias a la al apoyo de la IES **Elaios** fue posible tener como locutores a jóvenes entre los 12 y 16 años de edad.

La grabación del corpus tuvo lugar en las instalaciones del colegio **Rio Ebro** teniendo en cuenta la comodidad de los locutores y evitando grandes desplazamientos. Adicionalmente para la grabación del corpus se tuvieron en cuenta los siguientes aspectos:

- En lo posible mínimas condiciones de ruido aditivo y convolutivo.
- Una debida instrucción a los locutores respecto a la generación de las vocales a grabar, es decir enfatizando que estas fueran aisladas, sostenidas, y con una entonación natural o espontánea.
- En lo posible un adecuado balance en el número de locutores femenino y masculino, partiendo desde la niñez donde el requisito fue que el niño ya supiese las vocales, hasta la adolescencia donde los cambios hormonales afectan en gran medida el tono y el timbre de la voz.

Tabla 4.1: *Formulario de Registro de Datos.*

Locutor No.	Registro de Audio	Sexo	Edad	Talla
1	xx	xx	xx	xx
⋮	⋮	⋮	⋮	⋮
235	xx	xx	xx	xx

Ya que la finalidad del corpus es meramente investigativa y no se utilizaron los datos personales de los locutores, las voces fueron donadas por ellos y en su lugar se registraron otros datos de interés para la investigación como los mostrados en la Tabla 4.1 para cada locutor.

4.2.2 Entorno de la Adquisición

Figura 4.1: *Entorno de grabación.*

Como herramienta de grabación se utilizó una interface especialmente diseñada basada en la herramienta *PreLingua*, ya que en ella los efectos de la voz se traducen en movimientos de elementos gráficos muy llamativos para los niños, situación que ayudo a que fuese una experiencia motivadora. Se utilizo un computador portátil convencional con el sistema operativo WindowsXP y un micrófono de escritorio. La Figura 4.1 muestra el entorno real de trabajo con la disposición de estos elementos.

4.2.3 Características de los Locutores

El corpus de voz adquirido se compone de 235 registros de audio correspondientes a 235 locutores, de los cuales 110 corresponden al sexo femenino y 125 al sexo masculino. Cada

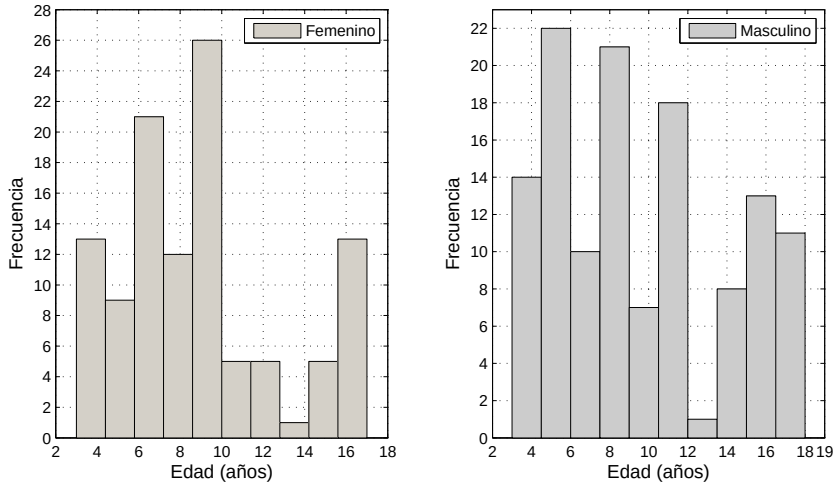


Figura 4.2: *Histograma de Edad de los Locutores.*

grabación contiene la producción sonora de las cinco vocales del español y cada vocal fue pronunciada de manera aislada, sostenida, y con un breve intervalo de silencio entre cada vocal.

La Figura 4.2 muestra el histograma con la distribución por edades para locutores femeninos en la parte izquierda y masculinos en la parte derecha. Las edades se distribuyen entre los 3 y 17 años de edad para locutores femeninos y desde los 3 a los 18 años en locutores masculinos. En el intervalo de los 12 a 14 años de edad solo se pudo contar con un locutor por cada género debido a una actividad particular en el curso de estudiantes de esa edad que impidió su asistencia a las grabaciones.

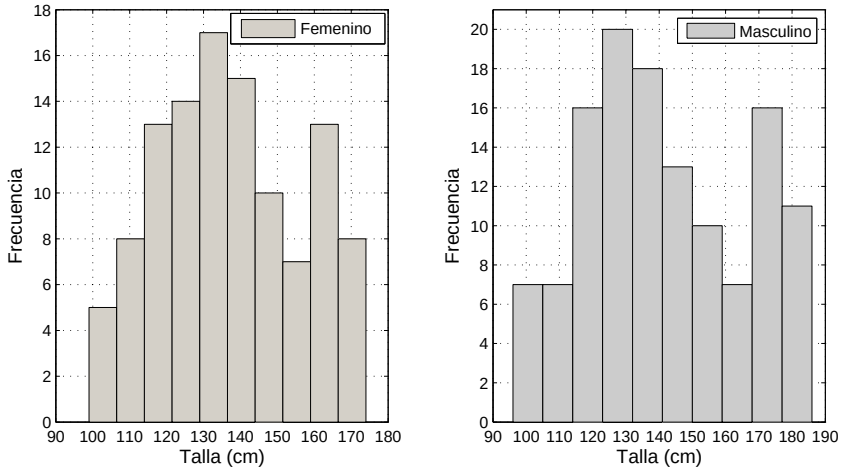


Figura 4.3: *Histograma de Talla de los Locutores.*

Como se explica en el Capítulo 6, otra finalidad de este corpus es encontrar una correlación entre la talla del locutor y la longitud de su tracto vocal estimada a partir de los

formantes vocálicos. Por esta razón es más relevante para la investigación caracterizar a los locutores de acuerdo a su talla, ya que es bien sabido que las personas no tienen la misma talla por el hecho de tener el mismo sexo y edad. La Figura 4.3 muestra el histograma de tallas para locutores femeninos en la parte izquierda y para masculinos en la parte derecha.

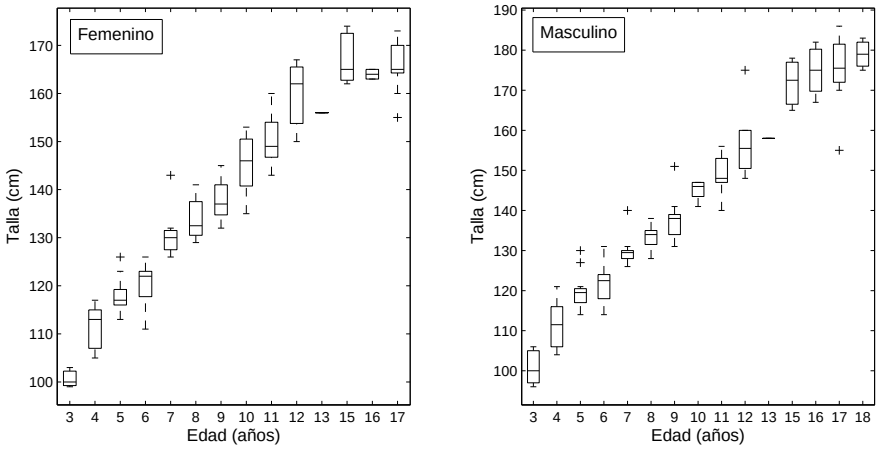


Figura 4.4: *Diagrama de caja para Edad vs Talla.*

Finalmente, la Figura 4.4 muestra el diagrama de cajas donde se relaciona la edad y la talla de los locutores. Allí se puede apreciar que evidentemente cuando aumenta la edad de una persona aumenta también su talla pero en diferente medida. Se observa por ejemplo que para esta población existe una mayor dispersión en las tallas de los locutores femeninos que en los locutores los masculinos. También, que aproximadamente a partir de los 11 años de edad la dispersión de tallas entre el primer y tercer cuartil tiende a aumentar sobretodo en el caso de los locutores masculinos, alcanzando una mayor estatura que en los casos de locutores femeninos.

Capítulo 5

Estimación Robusta de Formantes

Una vez obtenido el corpus de voz infantil no alterada, la siguiente tarea fue analizarlo en detalle para conocer los valores de los parámetros de la voz de esta población en función del sexo y la talla de los locutores. Se aplicaron técnicas tradicionales en procesado de voz como las descritas en el Capítulo 3 para estimar: la intensidad de la señal de voz en los segmentos sordos y sonoros, la frecuencia fundamental y los formantes vocálicos. De las técnicas mencionadas, la estimación de formantes fue la técnica que más presento dificultades mostrando estimaciones erróneas sobretodo en aquellas voces con valores altos de pitch, especialmente entre los 3 y los 9 años de edad. Esta dificultad técnica que genera estimaciones erróneas es tratada en la Sección 5.1, allí se explica como influye la alta tonalidad en la estimación formántica. La Sección 5.2 describe una técnica alternativa para eliminar esta influencia basada en el análisis homomórfico y que permite estimar de una manera más robusta los formantes vocálicos en la voz infantil

5.1 Dificultad Técnica de la Voz Infantil

La estimación de formantes es de por si una tarea difícil, situación que disminuye considerablemente cuando la frecuencia fundamental igualmente disminuye como en los casos de voces masculinas [Traunmuller and Eriksson, 1997]. Una manera de apreciar las diferencias entre voces adultas e infantiles es por medio de un espectrograma, ya que es una representación bidimensional que muestra la evolución temporal de la caracterización espectral de la señal de voz [Faúndez, 2000].

La Figura 5.1 muestra dos espectrogramas de voz con las cinco vocales del español pronunciadas de manera aislada y con un breve espacio de silencio entre ellas. El espectrograma mostrado en (a) corresponde a un adulto varón de 33 años de edad con una talla de 170 cm y con una media de pitch para toda la grabación de 110Hz. El espectrograma mostrado en (b) corresponde a una niña de 5 años de edad de 117 cm de altura y con una media de pitch para toda la grabación de 303 Hz. Cada segmento sonoro de los espectrogramas esta etiquetado con la vocal respectiva, así mismo, cada vocal tiene indicada la región donde hay mayor energía y se localizan los dos primeros formantes $F1$ y $F2$ que caracterizan dicha vocal.

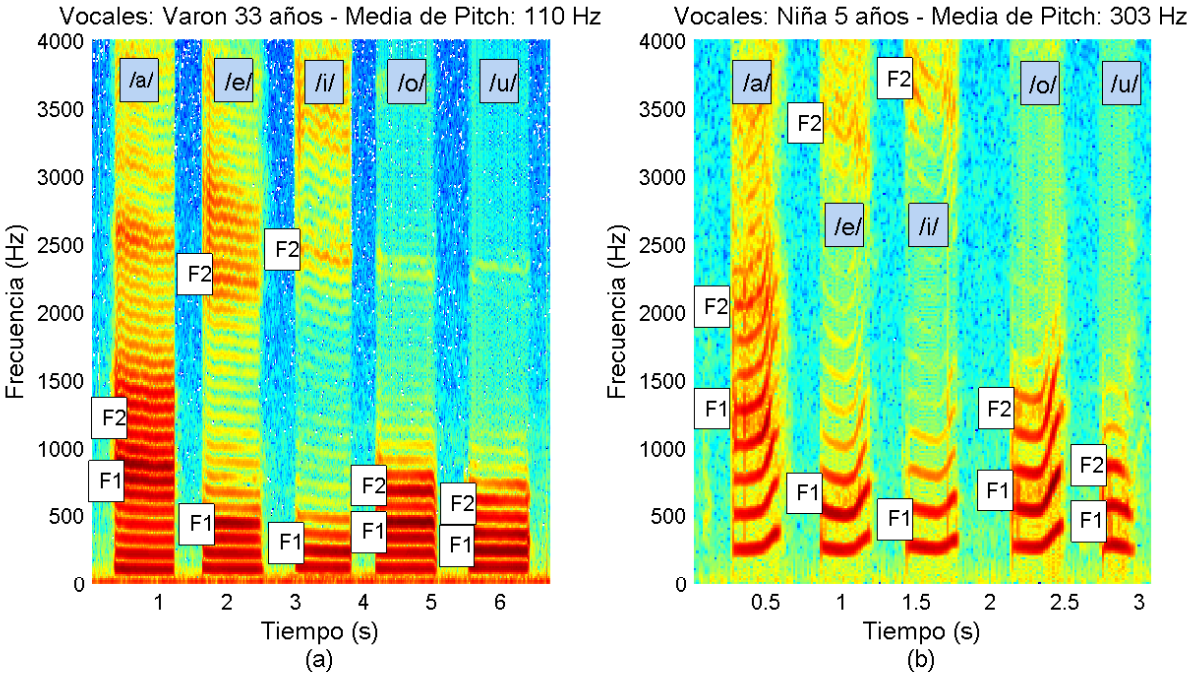


Figura 5.1: *Espectro de vocales en voz de adulto (a) y en voz infantil (b).*

En los espectros se pueden observar grandes diferencias entre los dos tipos de voz, diferencias fundamentalmente en la posición del pitch y la distribución de sus armónicos y la posición de los formantes para cada vocal. En el caso de la voz adulta la posición del pitch es más baja y por la misma razón la distancia entre sus armónicos es menor, mientras que en la voz infantil el pitch es mayor y sus armónicos están mucho más espaciados y acentuados quedando los formantes más difusos y de alguna manera ocultos entre los armónicos. Como se explicará posteriormente, al momento de estimar un formante lo que realmente se esta detectando es el pitch o uno de sus armónicos si el formante se encuentra cercano a este.

Comparando los valores de los formantes, se evidencia que los formantes de la voz infantil son superiores a los formantes de la voz adulta para todas las vocales, debido principalmente a que el tracto vocal de los niños es más corto y por ende sus frecuencias de resonancias son mayores, por ejemplo, mientras que en la voz adulta el segundo formante (F2) de la vocal /i/ esta sobre los 2500Hz, en el caso de la voz infantil este mismo formante F2 alcanza los 3650 Hz.

Son evidentes las grandes diferencias que existen entre la voz de un adulto y la de un infante, lo que de alguna manera indica que la estimación de formantes en voz infantil por técnicas tradicionales, puede no reflejar la realidad. Para entender lo que ocurre al estimar formantes en voces con alta tonalidad, retomaremos el análisis LPC de la Sección 3.5. Se sabe que este análisis es una solución eficiente y estable de los coeficientes AR pero con algunas limitaciones según lo indican:[Makhoul, 1975], [El-Jaroudi and Makhoul, 1991], [Vallabha and Tuller, 2002], donde enfatizan que los picos de la envolvente espectral estimados durante segmentos que tienen alto pitch, se encuentran sesgados hacia los armónicos del pitch.

Abordando el problema desde la óptica de procesado de señal, vamos a crear vocales sintéticas donde tenemos control para establecer los formantes y la frecuencia de excitación, y así mostrar las dificultades en la estimación de formantes. Considerando la vocal /u/, se sintetizaron cuatro señales diferentes convolucionando una respuesta impulsional $h[n]$ cuyos formantes de establecieron en F1=570Hz y F2=860Hz, y un tren de deltas como excitación a frecuencias de: 100Hz, 200Hz, 300Hz y 350Hz, con una frecuencia de muestreo de 8KHz. Posteriormente se estimaron los formantes de éstas cuatro señales utilizando el método LPC tradicional con un orden de predicción P=8, y los resultados obtenidos se muestran en la Figura 5.2.

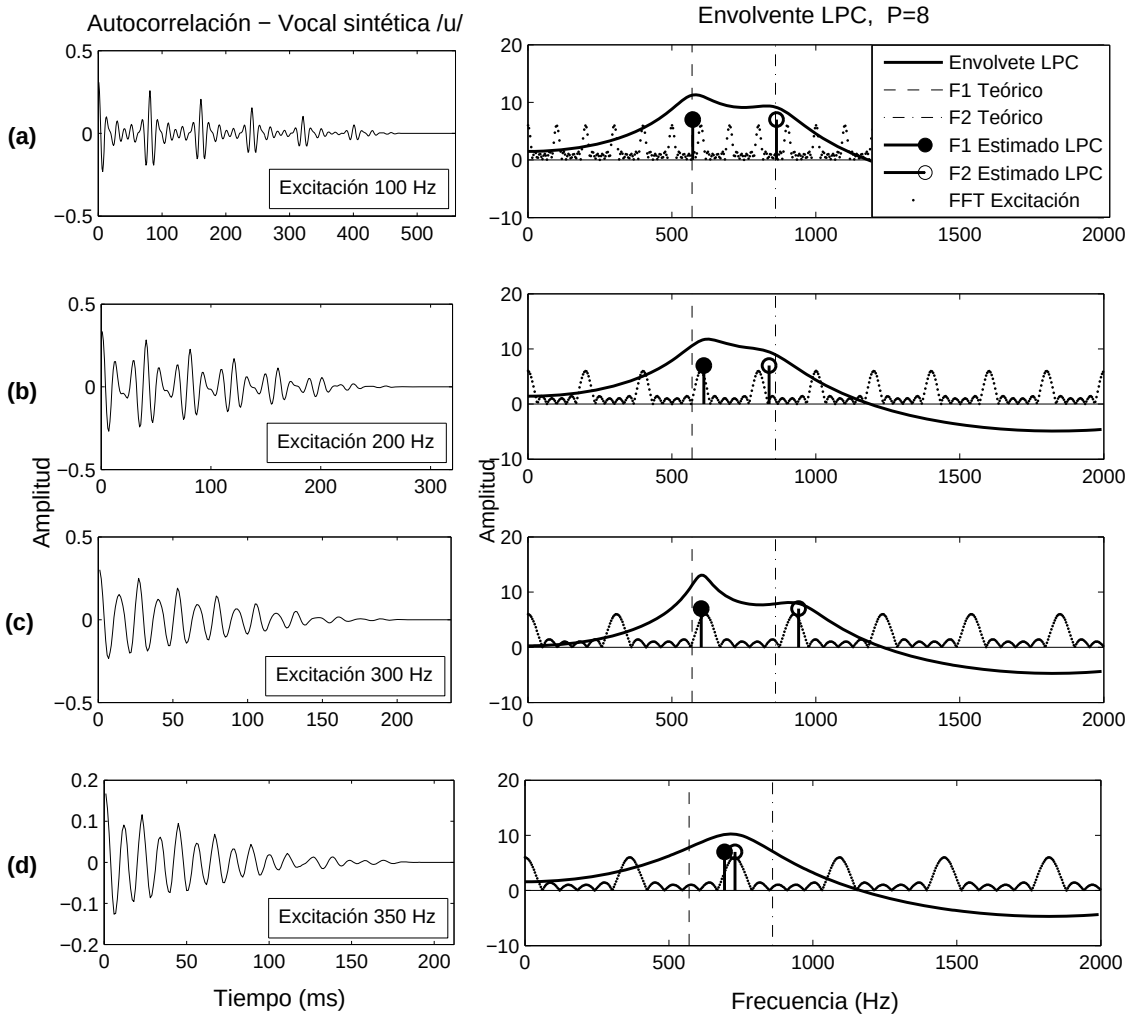


Figura 5.2: *Funciones de autocorrelación y estimación de formantes para vocales /u/ artificiales, sintetizadas con diferentes frecuencias de excitación.*

La parte (a), muestra la secuencia de autocorrelación para la señal sintetizada a 100Hz en la parte izquierda, y en la parte derecha, la envolvente LPC junto con los formantes teóricos, los formantes estimados, y la Transformada Rápida de Fourier (FFT) de la señal de excitación. Las partes (b), (c), y (d), muestran la misma información para las señales sintetizadas a 200HZ, 300Hz, y 350Hz respectivamente.

La respuesta impulsional $h[n]$ tiene una duración t_h de 80ms, de manera que al hacer la convolución con el tren de deltas de 100Hz, es decir un periodo t_e de 80ms, es posible estimar los formantes sin dificultad como muestra en la parte (a), es decir, que para señales con $t_e \geq t_h$, la secuencia de autocorrelación no presenta aliasing y la estimación de formantes es fiable. En la parte (b) de la figura, t_e es de 40ms siendo menor que t_h produciendo aliasing en la secuencia de autocorrelación, y la estimación de formantes tiende a tomar valores próximos al armónico de pitch más cercano alejándose de los reales. En (c), donde t_e es de 27ms, la dificultad en la estimación es más evidente donde el segundo formantes alcanza los 940Hz,y finalmente, en la parte (d) de la figura donde t_e es mucho menor que t_h (23ms), los formantes estimados se aproximan a los 700Hz que corresponden en realidad al segundo armónico de la excitación. La tendencia de los formantes de tomar valores cercanos a los armónicos del pitch cuando éste es alto, se puede apreciar también en la Figura 5.3 donde se muestran dos vocales sintetizadas con patrones variables de pitch, cuando el pitch se incrementa, la estimación es errática como en las zonas de las elipses hasta el punto de coincidir como se muestra en la parte derecha, y como se demostró en la parte (d) de la Figura 5.2.

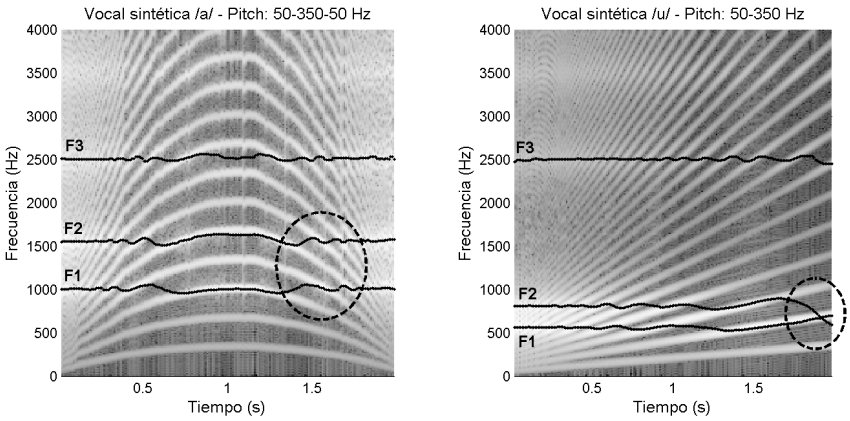


Figura 5.3: *Estimación de formantes en vocales sintéticas con patrones variables de pitch*

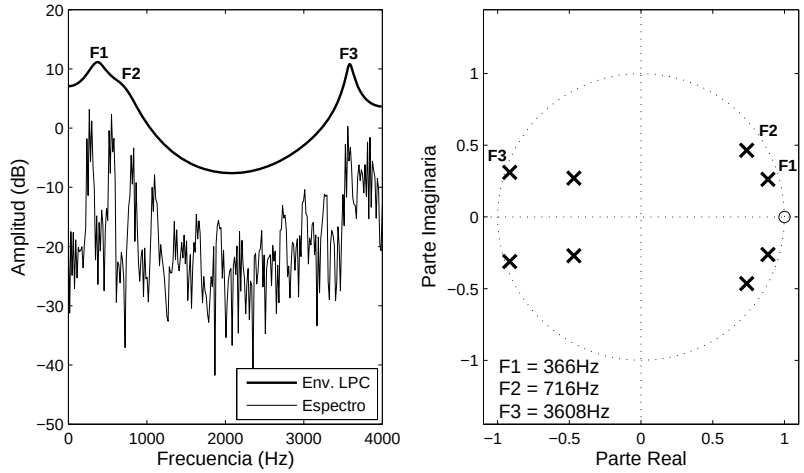


Figura 5.4: *Estimación de formantes para una trama de voz infantil de la vocal /i/*

Tomando de nuevo la señal de voz real de la Figura 5.1, se analizó la trama sonora de la vocal /i/ ubicada a los 1.6 segundos de tiempo para estimar los formantes aplicando el método LPC tradicional, los formantes obtenidos para dicha trama y su ubicación sobre el plano z se muestran en la Figura 5.4. El espectro muestra como el primer armónico del pitch influyen en la estimación tomando el valor F2, mientras que el valor real de éste es estimado como F3. Después de analizar ésta trama, se encuentra que los formantes son: F1=366Hz, F2=716Hz, y F3=3608Hz, lo que naturalmente no corresponde a los formantes de la vocal /i/.

Finalmente, estimando los formantes para la totalidad de la grabación de voz infantil, se encuentran estimaciones erróneas en las vocales /a/, /i/ y /o/, y en menor medida en las vocales /e/ y /u/, tal y como muestra el espectro de la Figura 5.5. Teniendo en cuenta lo anterior, es fácil hacerse una idea de la dificultad técnica para trabajar voz infantil y más aun, en población con voz alterada.

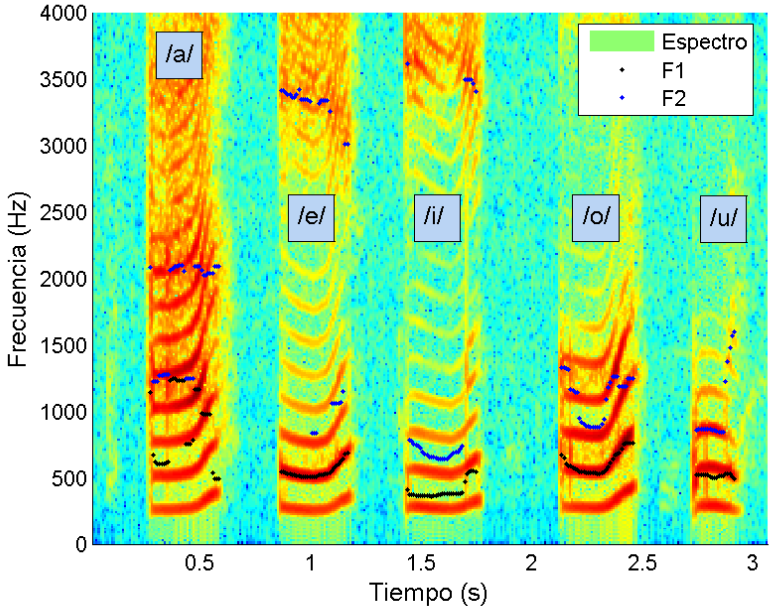


Figura 5.5: *Estimación de formantes para las cinco vocales en un locutor femenino de 5 años de edad.*

5.2 Eliminación de la Influencia de Pitch

Conociendo el problema, la influencia del alto pitch (fuente de excitación $e[n]$) en la estimación de formantes (respuesta impulsional del tracto vocal $h[n]$), se hace necesario separar estas componentes para obtener mejores estimaciones de formantes libres de la influencia del pitch. Ya que estas componentes que se hallan convolucionadas en tiempo, la tarea es entonces la deconvolución del segmento de voz de manera tal que las componentes queden combinadas linealmente y puedan separarse.

La técnica de deconvolución por análisis homomórfico tiene una larga historia de aplicaciones donde se requiere separar componentes periódicos de señales combinadas no linealmente [Oppenheim and Schafer, 1968]. Utilizando esta técnica será posible llevar una señal de voz $s[n]$ que se encuentra en el dominio temporal, al dominio cepstral en donde $\hat{s}[n]$ tendrá sus componentes $\hat{e}[n]$ y $\hat{h}[n]$ combinadas linealmente y podrán ser tratadas por separado como lo indica la expresión:

$$\hat{s}[n] = \hat{e}[n] + \hat{h}[n] \tag{5.1}$$

Trabajos previos como en [Shahidur and Shimamura, 2005] han mostrado que no todas las formulaciones de deconvolución cepstral son apropiadas para la estimación de formantes. En el caso del cepstrum complejo $c_c[n]$ definido por:

$$c_c[n] = \frac{1}{N}\sum_{k=0}^{N-1} log\left[X(k)\right] e^{j\frac{2\pi}{N}kn}, 0 \leq n \leq N-1 \tag{5.2}$$

donde $X(k)$ es la transformada de Fourier de N puntos de la señal de voz $s[n]$ y $log[X(k)] = log|X(k)| + jarg[X(k)]$, es un planteamiento no adecuado para estimar formantes debido a su alta sensibilidad a la fase [Quatieri, 1979], y porque la estimación del cepstrum complejo varia significativamente dependiendo de la posición de la ventana de análisis.

En el caso del cepstrum real $c_r[n]$ definido por:

$$c_r[n] = \frac{1}{N}\sum_{k=0}^{N-1} ln\left|X(k)\right| e^{j\frac{2\pi}{N}kn}, 0 \leq n \leq N-1 \tag{5.3}$$

no se tiene en cuenta la fase y la magnitud contiene información suficiente sobre la trama de voz para su posterior análisis. Una vez obtenido el cepstrum real de la señal de voz, las componentes $\hat{e}[n]$ y $\hat{h}[n]$ estarán linealmente combinadas (como muestra la Figura 3.16) y podrán ser tratadas de manera independiente.

Con las componentes separadas, es entonces necesario hacer un filtrado en el dominio cepstral conocido como liftado. Como la información correspondiente a la respuesta impulsional del tracto vocal $\hat{h}[n]$ se encuentra concentrada en la parte baja de $c_r[n]$, y la fuente de excitación $\hat{e}[n]$ se encuentra en la parte alta y es justamente esta la que se quiere eliminar, se puede utilizar el valor estimado previamente del periodo de pitch $Tpitch$ para hacer dicho liftado utilizando una ventana de liftado $w[n]$ y así eliminar la parte alta de $c_r[n]$.

La longitud de la ventana de liftado $w[n]$ puede introducir errores en la estimación de formantes ya que los coeficientes cepstrales cercanos al periodo de pitch pueden ser distorsionados [Verhelst and Steenhaut, 1986], de manera que es importante seleccionar adecuadamente su longitud. Una ventana de liftado con longitud $0.5Tpitch$ (50% del periodo de pitch) ha sido propuesta en [Verhelst and Steenhaut, 1986], en el caso de voces con alta frecuencia de pitch [Shahidur and Shimamura, 2005] han propuesto incrementar la longitud de la ventana de liftado a $0.7Tpitch$ para voces con frecuencia de pitch superior a 250Hz, y

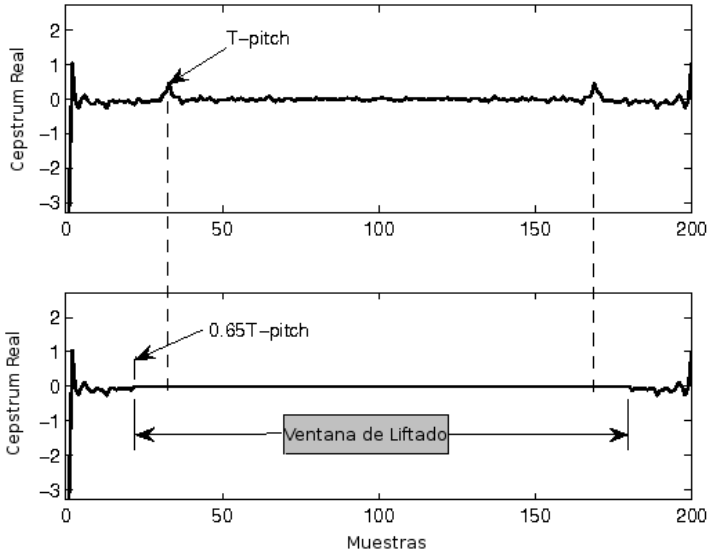


Figura 5.6: *Efecto del liftado en el dominio cepstral.*

de $0.6Tpitch$ para frecuencia menores a este valor.

En las pruebas iniciales aplicando el liftado sobre el corpus de voz adquirido, se aplico una ventana $w[n]$ con una longitud de $0.65Tpitch$ para valores de frecuencia de pitch superiores a 250Hz. $w[n]$ descrita en la ecuación 5.4 y su efecto al aplicarla sobre $c_r[n]$ se puede apreciar en la Figura 5.6.

$$w[n] = \begin{cases} 0 & 0.65T_{pitch} \leq n \leq N - 1 - 0.65T_{pitch} \\ 1 & otro\ n \end{cases} \tag{5.4}$$

Después de analizar los datos obtenidos en la totalidad del corpus respecto a la frecuencia de pitch, se encontró que este valor varía significativamente dependiendo de la talla y sexo del locutor. Como se aprecia en la Figura 5.7-a, este valor variar desde 100Hz hasta los 340Hz en los caso de locutores masculinos, y desde los 175Hz hasta 310Hz para locutores femeninos (Figura 5.7-b). De manera que establecer un valor fijo para la longitud de la ventana de liftado tendrá los efectos deseados cerca de los 250Hz y valores cercanos, para valores muy lejanos bien sea por encima o por debajo de este valor, el proceso de liftado no tendría el efecto deseado.

Para lograr que el sistema se adapte mejor a las características propias de cada usuario, se ha reemplazado el valor de la longitud de la ventana de liftado de 0.65 por un alfa (α), donde α se calcula por medio de una interpolación lineal obtenida de los datos directamente. En la recta para esta interpolación, mostrada en la Figura 5.7-c, α varía desde 0.5 hasta 0.74 para valores de pitch entre los 100Hz y 420Hz respectivamente. De esta forma al incrementarse el valor de pitch la longitud ventana de liftado estará más cerca de este valor y afectará menos a los coeficientes cepstrales cercanos.

Una vez realizado el liftado de cada trama de voz, la tarea a seguir es la estimación de formantes a partir de esta nueva secuencia liftada $\hat{c}_r[n]$. Hallando la densidad espectral de

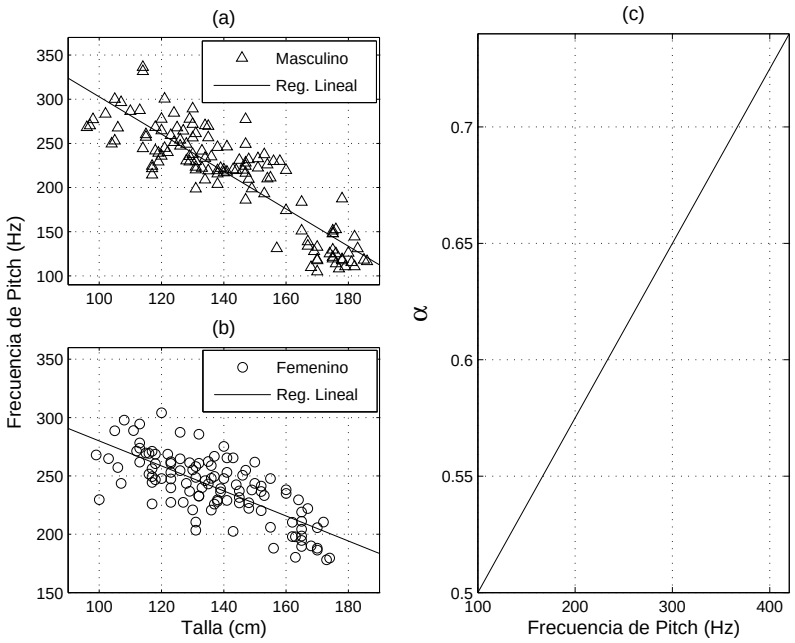


Figura 5.7: *Frecuencia de pitch VS talla, para locutores masculinos (a), locutores femeninos (b), y valor alfa para la ventana de liftado.*

potencia $S_x(e^{jw})$ de la secuencia $\hat{c}_r[n]$, y asumiendo que se trata de un proceso estacionario en sentido amplio (WSS), podemos obtener la nueva función de autocorrelación $\Gamma_x[k]$ gracias al teorema de Wiener-Khintchine (ecuación 5.5) aplicado en sentido inverso.

$$S_x(e^{jw}) = \sum_{k=-\infty}^{\infty} \Gamma_x[k] e^{-jwk} \tag{5.5}$$

Es decir, hallando la transformada inversa de Fourier de la densidad espectral de potencia $S_x(e^{jw})$, se puede obtener la función de autocorrelación $\Gamma_x[k]$, como muestra la ecuación 5.6 [Deller et al., 1993], [Proakis and Manolakis, 2007].

$$\Gamma_x[k] = \frac{1}{2\pi} \int_{2\pi} S_x(e^{jw}) e^{jwk} dw \tag{5.6}$$

Ahora, a partir de la función de autocorrelación obtenida $\Gamma_x[k]$ podemos estimar los nuevos formantes $\hat{F}_k$ libres de la influencia del pitch por el método LPC convencional descrito en la sección 3.7, y con los mismos parámetros establecidos de orden de predicción $P = 8$, frecuencia de muestreo de 8KHz y una ventana de análisis de 25ms.

Para comprobar que el método propuesto es robusto y fiable a la hora de estimar formantes en voz infantil, éste se aplicó de nuevo a las mismas vocales sintéticas de la Figura 5.2, para ver hasta que punto el método eliminaba la influencia de la alta tonalidad, los resultados obtenidos en dicha prueba se muestran en la Figura 5.8. En la parte (a) de la figura, se observa que para $t_e \geq t_h$ el método propuesto (en rojo) tiene los mismos resultados en la estimación de formantes que el método tradicional LPC, en (b), F1 se ve igualmente afectado pero en cambio F2, logra mantener una estimación correcta del formante. Para el

caso de la señal sintetizada con una excitación de 300Hz (c), en donde la estimación por el método tradicional se aleja bastante de la realidad, el método propuesto estima F1 de manera acertada y F2 con una pequeña variación. Finalmente en la parte (d) en donde el método tradicional falla completamente, el método basado en el liftado se acerca bastante a los valores reales de los formantes demostrando que éste funciona.

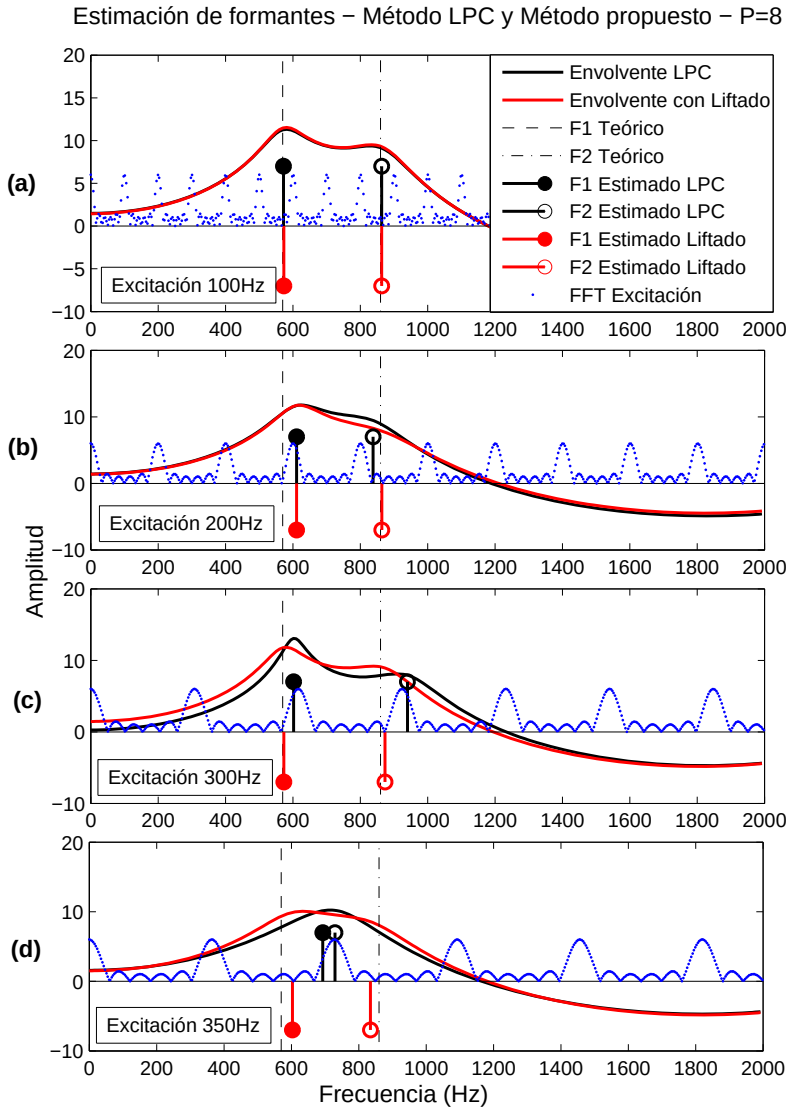


Figura 5.8: *Estimación de formantes por el método LPC y el método propuesto con liftado, para frecuencias de excitación de: (a) 100Hz, (b) 200Hz, (c) 300Hz y (d) 350Hz.*

Aplicando ahora el método propuesto en las vocales sintéticas de la Figura 5.3, la nueva estimación de formantes que se aprecia ahora en la Figura 5.9, es más robusta frente a los incremento de pitch y la estimación se mantiene en valores cercanos o reales a los establecidos en cada vocal. En ésta figura, se aprecia también el efecto del liftado sobre toda la señal eliminando el pitch y sus armónicos lo que beneficia en general la estimación.

Tomado de nuevo la señal de voz infantil de la Figura 5.1 y aplicando la técnica de liftado expuesta, los nuevos formantes estimados para la trama de la vocal /i/ se muestran

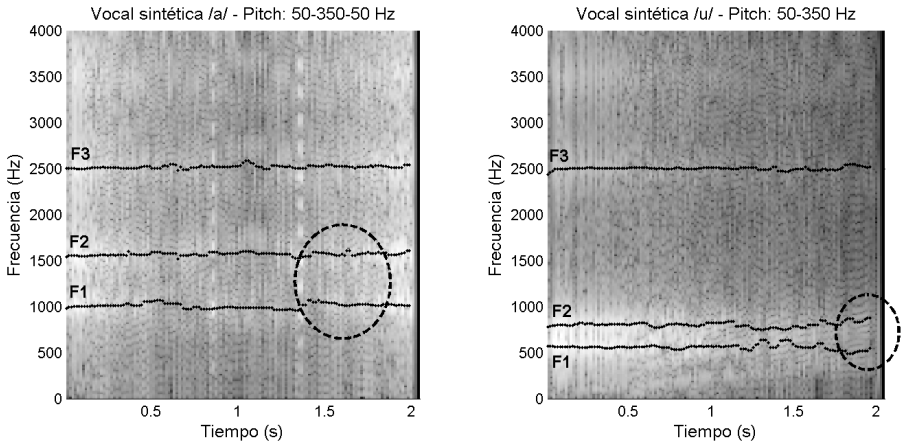


Figura 5.9: *Estimación de formantes en vocales sintéticas con el método propuesto.*

en la Figura 5.10. Allí se observa en el espectro que después del proceso de liftado se han eliminado los armónicos del pitch que influían principalmente en la estimación de $F1$ y $F2$ si se compara con la Figura 5.4, situación que ha permitido estimar el valor real del segundo formante $\hat{F}_2$. Observando el plano z, los polos que se encuentran cercanos a la circunferencia de radio unidad si pertenecen a la vocal /i/, además, el polo $F2$ de la Figura 5.4 se ha desplazado alejándose de la circunferencia de radio unidad permitiendo la estimación real del segundo formante. En resumen, después de aplicar la técnica de liftado expuesta para esta trama de voz infantil, los formantes estimados $\hat{F}_1 = 392Hz$ y $\hat{F}_2 = 3679Hz$ si corresponden a la vocal /i/ en contraste con las estimaciones iniciales de $F1 = 366Hz$ y $F2 = 716Hz$.

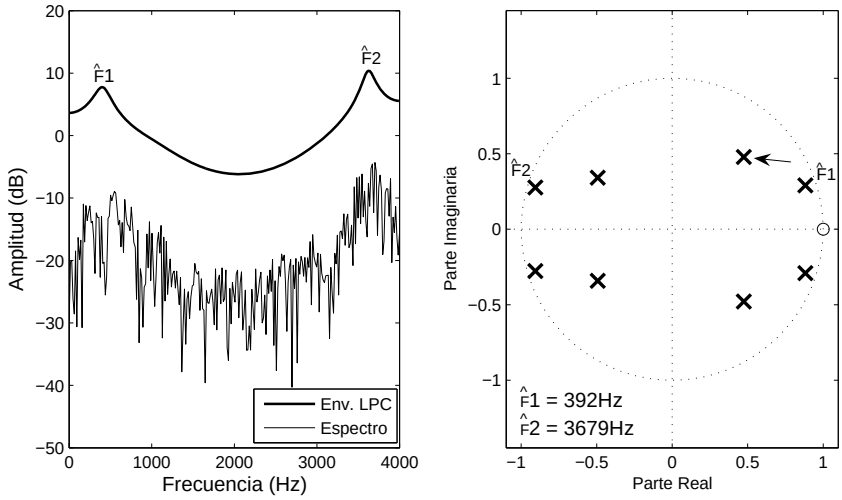


Figura 5.10: *Estimación de formantes para una trama de voz infantil de la vocal /i/ con el método propuesto.*

Considerando la grabación completa con las cinco vocales para el mismo locutor infantil, se estimaron los formantes para cada vocal después de aplicar la técnica de liftado. El resultado de la nueva estimación es mostrado en la Figura 5.11. Después de aplicar el

proceso del liftado las estimaciones formánticas han mejorado considerablemente como lo muestra la figura. En primer lugar, al eliminarse el pitch y sus armónicos el espectro muestra más claramente el lugar donde se concentra la energía para cada formante en cada vocal. Las vocales /a/ y /o/ muestran una marcada diferencia en la estimación de $\hat{F}_1$ y $\hat{F}_2$ respecto a las iniciales (Figura 5.5) porque las estimaciones se ubicación en una zona de alta energía y, porque la trayectoria de los formantes ya no se ve afectada por el contorno del armónico del pitch.

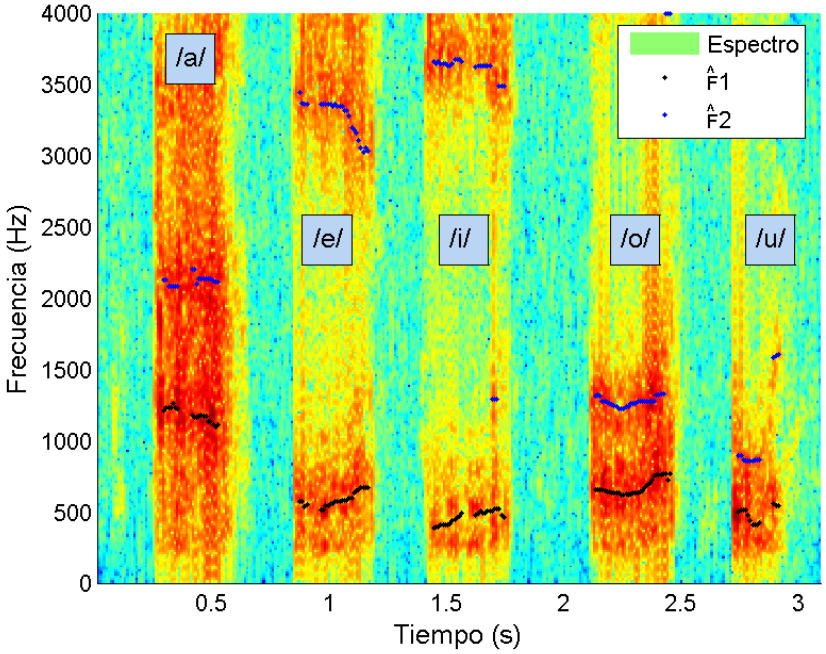


Figura 5.11: *Formantes estimados para las cinco vocales (Niña 5 años, talla 117cm) antes y después de aplicar el método propuesto.*

La vocal /e/ y en especial la /i/ son las vocales que más se benefician de la aplicación de la técnica de liftado, ya que al poseer los formantes más extremos el primero tiende a ser estimado sobre la frecuencia de pitch, y el segundo formante en su primer armónico, quedando el formante numéricamente bastante alejado de su valor real. En general todas las vocales presentan alguna mejora inclusive la /u/, en donde sus valores reales de los formantes pueden estar muy cercanos al pitch y sus primeros armónicos, pero la ganancia se evidencia en que las trayectorias de los formantes evitan seguir las trayectorias de los armónicos del pitch.

Partiendo del hecho de que con la técnica de liftado se estima de manera más fiable los formantes en voz infantil, se aplico esta técnica a las grabaciones de los 235 locutores del corpus. Los resultados obtenidos fueron bastante más satisfactorios e interesantes que antes de aplicar el liftado en el sentido de que explican la alta variabilidad inter-locutor de los formantes.

La Figura 5.12-(a) muestra el triángulo vocálico formado por $\hat{F}_1$ VS $\hat{F}_2$ estimados para los 125 locutores masculinos, y la media y varianza para cada vocal en la parte (b) de la

misma figura. Para el caso de los locutores femeninos, la misma información se muestra en (*c*) y (*d*) respectivamente. Comparando la distribución de los formantes entre ambos sexos, se aprecia que hay una mayor varianza en el segundo formante para las vocales /e/ e /i/ en los locutores masculinos debida principalmente a la disminución de los valores de estos formantes cuando los locutores alcanzan la adolescencia, momento en el que ocurre un incremento importante en la longitud del tracto vocal. Esta situación es mucho menos marcada en el caso de los locutores femeninos en donde el tracto vocal crece en menos medida que en los hombres. Para las vocales /a/, /o/ y /u/ este fenómeno también se presenta en mayor medida en los locutores masculinos que en los femeninos pero, fundamentalmente en el primer formante.

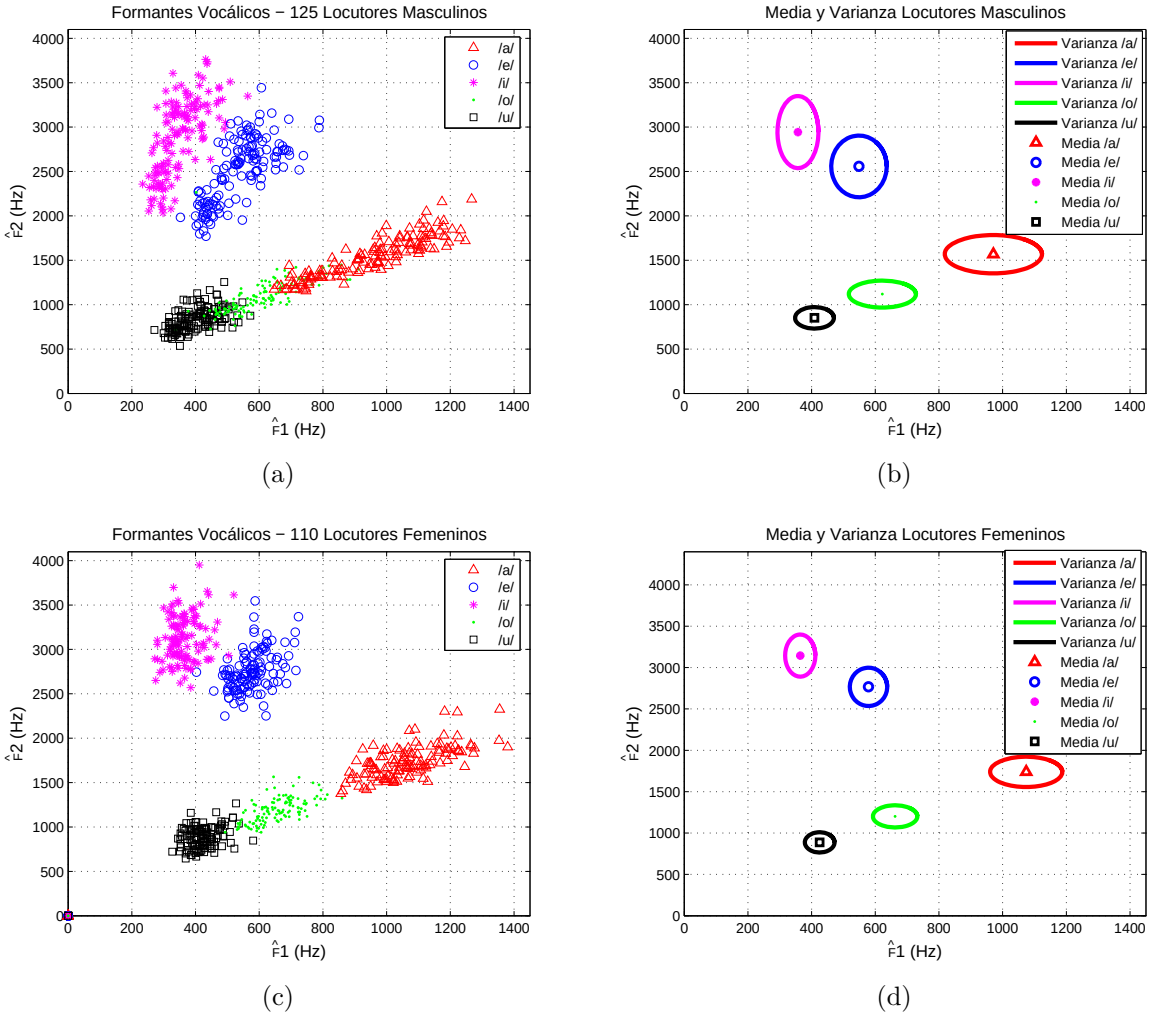


Figura 5.12: *Formantes vocálicos, media y varianza estimados para locutores masculinos (arriba), y locutores femeninos (abajo).*

En general, las medias de los formantes para cada vocal en los locutores femeninos son mayores que las medias de los locutores masculinos, y las varianzas son menores en todas las vocales femeninas respecto a las masculinas.

Este capítulo pone en evidencia la dificultad técnica al estimar formantes en el habla

infantil, también propone el uso del análisis homomórfico y el liftado como una solución para estimar con mayor fiabilidad y robustez estos formantes. Con mejores estimaciones es posible buscar una técnica de normalización que permita disminuir la gran variabilidad formántica entre locutores infantiles y poder así desarrollar herramientas para logopedia.

Capítulo 6

Estimación del la Longitud del Tracto Vocal y Normalización

Analizando los formantes del corpus de voz se puede observar la alta variabilidad de estos en ambos sexos y en especial en los locutores masculinos. En ellos, el hecho de llegar a la pubertad representa grandes cambios manifestados de manera física y especialmente en su voz, así lo muestran los triángulos vocálicos de la Figura 5.12(a) y (b) contrastados con los casos femeninos (c) y (d). En los casos femeninos esta situación es igualmente notoria aunque en menor medida pero en cambio en ellas, los formantes de las vocales abiertas /a/, /e/ e /i/ son mayores en media que los formantes de los casos masculinos. La anterior situación justifica la necesidad de encontrar una manera de poder llevar los formantes de la voz infantil a un espacio de trabajo más homogéneo y de menor variabilidad, de manera que las aplicaciones a desarrollar permitan trabajar en lo posible articulación vocálica independientemente del sexo y edad del infante.

En este capítulo se propone una técnica de normalización de formantes a través de la longitud del tracto vocal del usuario. Para lograrlo, se parte de un modelo de tracto vocal uniforme que es descrito en la Sección 6.1, que permitirá estimar la longitud del tracto vocal como se explica en la Sección 6.2 y, finalmente normalizar los formantes ya estimados como lo muestra la Sección 6.3.

6.1 Modelo del Tracto Vocal

El sistema de producción del habla puede verse como una excitación que atraviesa el canal del tracto vocal, y este canal se comporta como un filtro acústico que modifica la distribución espectral de energía de la señal de excitación. Una aproximación para estudiar este filtro acústico es modelarlo como un tubo de sección uniforme sin pérdidas o como una concatenación de tubos de sección variable. En cualquiera de los dos modelos se asume que las cuerdas vocales, es decir la excitación, son independientes del tracto vocal.

Ya que el objetivo aquí es obtener la longitud del tracto vocal para normalizar las estimaciones de los formantes, el modelo descrito es el de un tubo de sección uniforme sin pérdidas donde no se considera la cavidad nasal. Para este modelo suponemos un único tubo de área uniforme A como el de la Figura 6.1, el cual se encuentra cerrado en el extremo

donde se aplica la excitación (glotis, $x = 0$) y abierto en el otro extremo donde se encuentran los labios ($x = l$).

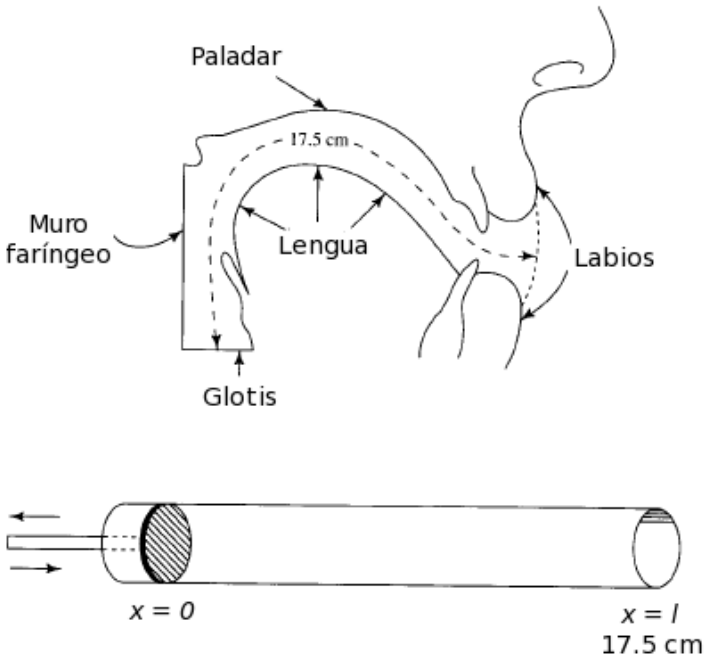


Figura 6.1: *Modelo de tubo uniforme sin pérdidas del Tracto Vocal.*

En el modelo, la longitud $l = 17.5cm$ es la asumida para un adulto varón estándar y es la distancia lineal comprendida entre la glotis y los labios. Para este modelo definiremos:

- $u(x,t) \Longrightarrow$ velocidad de una partícula de prueba
- $U(x,t) \Longrightarrow$ velocidad volumétrica ($U = uA$)
- $p(x,t) \Longrightarrow$ variación de la presión del sonido ($P = P\ o + p$)
- $\rho \Longrightarrow$ densidad del aire
- $c \Longrightarrow$ velocidad del sonido

Entendiendo la anterior notación como: por ejemplo $U(x,t)$ como la velocidad volumétrica a una distancia x del origen de la excitación en el tiempo t. Suponiendo una propagación de onda plana y un movimiento ondulatorio unidimensional, puede demostrarse que:

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial u}{\partial t} \tag{6.1}$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2}\frac{\partial p}{\partial t} \tag{6.2}$$

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \tag{6.3}$$

cuyas soluciones en el dominio del tiempo y la frecuencia son de la forma:

$$u(x,t) = u^{+}(t - \frac{x}{c}) - u^{-}(t + \frac{x}{c}) \tag{6.4}$$

$$u(x,s) = \frac{1}{\rho c}\left[P + e^{\frac{-sx}{c}} - P - e^{\frac{sx}{c}}\right] \tag{6.5}$$

$$p(x,t) = \rho c\left[u^{+}(t - \frac{x}{c}) + u^{-}(t + \frac{x}{c})\right] \tag{6.6}$$

$$p(x,s) = P + e^{\frac{-sx}{c}} + P - e^{\frac{sx}{c}} \tag{6.7}$$

tomando ahora la velocidad volumétrica de la glotis como $U_G(j\omega)$ y la velocidad volumétrica de los labios como $U_L(j\omega)$, la función de transferencia del tracto vocal $T(j\omega)$ sera:

$$T(j\omega) = \frac{U_L(j\omega)}{U_G(j\omega)} = \frac{U(l,j\omega)}{U(0,j\omega)} \tag{6.8}$$

Utilizando las condiciones de contorno $U(0,s) = U$ y $P(l,s) = 0$, la función de transferencia se expresa como:

$$T(s) = \frac{2}{e^{\frac{sl}{c}} + e^{\frac{-sl}{c}}} \tag{6.9}$$

$$T(j\omega) = \frac{1}{cos(\frac{\omega l}{c})} \tag{6.10}$$

Es fácil ver que polos de la función de transferencia $T(j\omega)$ están donde $cos(\frac{\omega l}{c})$ es igual a cero, es decir en aquellas frecuencias f_n de $T(j\omega)$ que tienden a infinito:

$$\frac{(2\pi f_n)l}{c} = \frac{(2n-1)}{2}\pi \tag{6.11}$$

$$f_n = \frac{c}{4l}(2n-1) \tag{6.12}$$

$$\lambda_n = \frac{4l}{(2n-1)} \qquad n = 1,2,3,... \tag{6.13}$$

Lo que implica que las resonancias ocurren en múltiplos impares de la frecuencia fundamental f_n. Tomando la velocidad del sonido como $c = 34000cm/seg$ y $l = 17.5cm$, las frecuencias de resonancia o formantes aparecen en 500Hz, 1500Hz, 2500Hz, etc, como lo muestra la Figura 6.2.

Del anterior análisis se concluye que la función de transferencia de un tubo sin ramas laterales, excitado en un extremo y con la respuesta medida en el otro extremo, únicamente posee polos. También que las frecuencias de resonancia tendrán un ancho de banda finito cuando se consideran las pérdidas del tracto vocal como la radiación, paredes, viscosidad o calor.

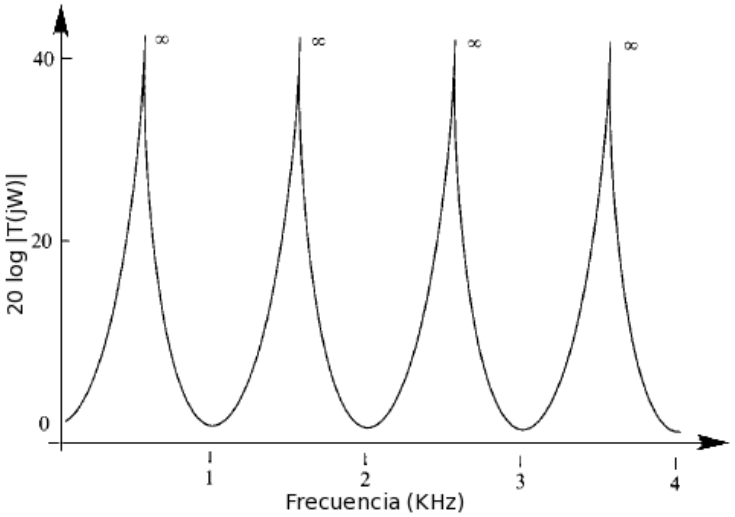


Figura 6.2: *Resonancias de un tubo uniforme de 17.5 cm de longitud.*

Finalmente, para el modelo anteriormente descrito la longitud del tracto vocal l corresponde a: $\frac{1}{4}\lambda_1$, $\frac{3}{4}\lambda_2$, $\frac{5}{4}\lambda_3$, ..., donde λ_i es la longitud de onda de la frecuencia natural i^{th}. El tubo uniforme descrito aquí cerrado en un extremo y abierto en el otro, es conocido como un resonador en cuarto de longitud de onda.

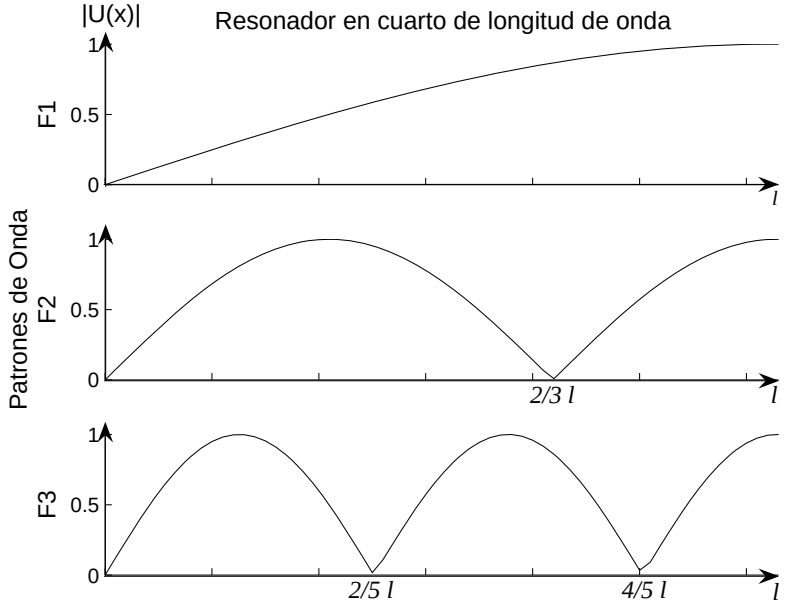


Figura 6.3: *Patrones de onda para un resonador en cuarto de longitud de onda.*

Para un resonador de este tipo con una longitud $l = 17.5cm$, los patrones de onda de la velocidad volumétrica para las tres primeras frecuencias naturales $F1$, $F2$ y $F3$ se muestran en la Figura 6.3, allí se aprecia como $|U(x)|$ es mínima en el extremo cerrado del tubo (glotis) y máxima en el extremo abierto para cada frecuencia natural. [Deller et al., 1993], [Stevens, 1998], [R. Schafer, 1978].

6.2 Estimación de la Longitud del Tracto Vocal

Uno de los objetivos de conocer la longitud del tracto vocal en niños en función del crecimiento, es poder correlarla con los formantes estimados de manera robusta y encontrar un modelo que refleje el comportamiento de estos en función de la talla y sexo del usuario. Por otra parte, conociendo los valores formánticos de un individuo de talla y sexo determinados, se podrían utilizar estos valores para trabajar articulación vocálica en otro individuo de características semejantes pero con alguna alteración en la articulación de sonidos vocálicos.

Lamentablemente hay muy pocos estudios que relacionen el crecimiento del tracto vocal de un individuo con el crecimiento mismo del cuerpo. En el estudio: [Tecumseh, 1997] por ejemplo, se correla la longitud del tracto vocal y la talla para 23 primates (monos rhesus). Estudios en humanos como: [Goldstein, 1980] y [Vorperian et al., 2005], se relaciona el crecimiento del tracto vocal con la edad, éste último trabajo muestra además el crecimiento de diferentes estructuras como el paladar blando y duro, la mandíbula y la lengua, todo basado en el análisis y mediciones sobre imágenes de resonancia magnética (o Magnetic Resonance Image (MRI)). En principio, pareciera que éste estudio es de gran ayuda para la investigación, sin embargo, la información disponible de los casos de estudio es mínima, el rango de edad va solamente desde el nacimiento hasta los 7 años de edad e incluye 12 casos de adultos, y las variables estudiadas se explican en función de la edad y no de la talla como se estableció desde un principio en ésta investigación.

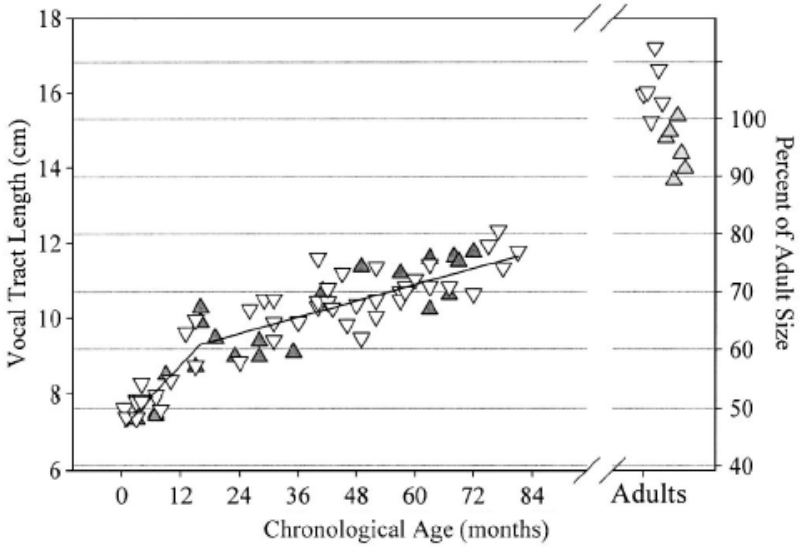


Figura 6.4: *Longitud del tracto vocal en casos pediátricos y adultos. (Tomado de [Vorperian et al., 2005], triángulos hacia arriba casos femeninos y triángulos hacia abajo casos masculinos)*

La Figura 6.4 muestra los resultados del estudio [Vorperian et al., 2005] donde se relaciona la longitud del tracto vocal en función de la edad en meses. Allí se observan dos zonas lineales con una alta correlación entre las variables pero resulta difícil separar el comportamiento para los casos femeninos y masculinos, además no se cuenta con información

a partir de los 7 años de edad.

Viendo las dificultades para encontrar información fiable de como crece el tracto vocal en niños, se abordó el problema tratando de obtener la longitud del tracto de un individuo directamente de la emisión sonora de éste. Una manera de conocer la longitud del tracto vocal de un locutor determinado, es pidiéndole a éste que genere un sonido sonoro procurando que todo su tracto vocal esté configurado con una sección homogénea como el modelo descrito en la Sección 6.1, luego se hace la estimación de los formantes y finalmente se puede obtener la longitud del tracto aplicando la expresión 6.14.

$$l = \frac{c}{4fn}(2n-1) \tag{6.14}$$

Éste sonido en particular que es muy próximo a la vocal /æ/ francesa y muy difícil de conseguir de manera voluntaria, se ubica frecuencialmente entre las vocales /e/ y /o/ en la zona del centro de masa del triángulo vocálico, éste centro de masa se obtiene calculando la media de todos los formantes F1 y todos los formantes F2 de las cinco vocales como muestra la Figura 6.5.

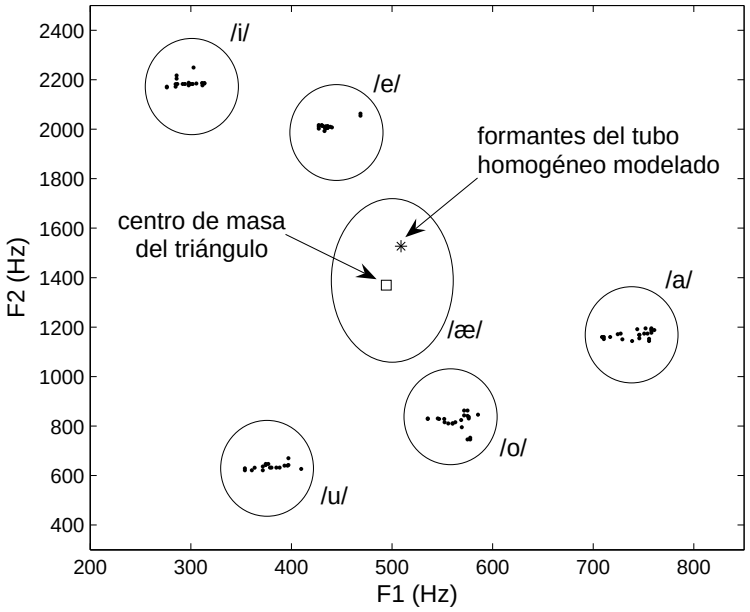


Figura 6.5: *Ubicación del centro de masa de un triángulo vocálico, y de los formantes de un tubo homogéneo modelado con los mismos formantes.*

Obtener la longitud de tracto vocal por éste método resulta poco preciso y muy difícil de conseguir, de manera que el problema debía ser abordado utilizando una alternativa numérica a partir de la propia información formántica de cada usuario, además, se contaba con el corpus de voz infantil y con sus formantes estimados de manera robusta.

El método para estimar la longitud del tracto vocal a partir de los formantes vocálicos y que se describe a continuación, da como resultado final la longitud de un tubo homogéneo cuyos formantes teóricos caen próximos al centro de masa del triángulo (* en la Figura 6.5),

demostrando que el método funciona y que se trata de una buena aproximación ya que se consideran los propios formantes del locutor obtenidos de las cinco vocales.

Trabajos previos para estimar la longitud del tracto vocal proponen hacerlo a partir de la impedancia de los labios como en: [Paige and Zue, 1969] o a partir de las áreas del tracto vocal en los modelos de concatenación de tubos como en: [Wakita, 1977], [Kirlin, 1978] y [Schroeder, 1967]. El método utilizado en ésta investigación fue propuesto por [Necioglu et al., 2000], en éste trabajo se estima la longitud del tracto vocal en adultos a partir de emisiones vocálicas en ingles, de donde se estiman los formantes y se evalúan diferentes métodos para obtener la longitud. Ya que el marco de trabajo es similar, en el sentido que se tienen grabaciones con las vocales del español para cada locutor y que la técnica permite obtener la longitud del tracto vocal directamente de los formantes, se seleccionó ésta técnica para estimar la longitud del tracto vocal de los locutores del corpus.

En ésta técnica se parte del modelo de tubo uniforme de la Sección 6.1, en donde las resonancias de éste tubo descritas por la ecuación 6.12 se encuentran uniformemente espaciadas. La estimación de la longitud se puede resumir a un ajuste de las frecuencias de resonancia medidas $\tilde{F}_k$, con las frecuencias de resonancia del tubo uniforme del modelo, las cuales están determinadas solamente por su longitud l. Es decir que el problema se puede aproximar reduciendo al mínimo el error ε:

$$\varepsilon = \sum_{k=1}^{M} D\left(\tilde{F}_k, (2k-1)f_1\right) = \sum_{k=1}^{M} D\left(\tilde{F}_k, (2k-1)\frac{c}{4l}\right) \tag{6.15}$$

donde $D(\tilde{F}_k, (2k-1)f_1)$ es una función que expresa la diferencia entre los formantes medidos $\tilde{F}_k$, y los formantes del tubo homogéneo. El error de la ecuación 6.15 puede construirse utilizando una función de distancia entre los formantes medidos $\tilde{F}_k$ y las resonancias impares de un tubo uniforme $(2k-1)f_1$, esta función puede ser:

$$\varepsilon = \sum_{k=1}^{M} \frac{\left(\frac{\tilde{F}_k}{2k-1} - f_1\right)^2}{f_1} \tag{6.16}$$

minimizando ahora la ecuación 6.16, se puede hallar la frecuencia de resonancia "fundamental" del tubo homogéneo f_1:

$$f_1 = \left(\frac{1}{M}\sum_{k}\left(\frac{\tilde{F}_k}{2k-1}\right)^2\right)^{1/2} \tag{6.17}$$

Finalmente, la Longitud del Tracto Vocal (LTV) se obtiene con la expresión 6.18 utilizando el valor obtenido en f_1:

$$VTL = \frac{c}{4f_1} \tag{6.18}$$

En [Necioglu et al., 2000] han aplicado ésta técnica en 164 locutores del corpus TIMIT, utilizando 8 frases del corpus para obtener la LTV por cada frase y por cada locutor, finalmente obtienen la LTV final para cada locutor como el promedio de las longitudes de las 8 frases. En otro trabajo como [Wakita, 1977], se grabaron a 26 adultos (14 masculinos

y 12 femeninos) pronunciando 10 palabras monosilábicas cada una comenzando con /h/ y finalizando con /d/ por ejemplo: /heed/, de modo que cada vocal quedara claramente pronunciada, y se obtuvo la LTV final para cada vocal como el promedio de las 26 longitudes estimadas de los 26 locutores.

Ya que el corpus de voz adquirido en ésta investigación contiene las cinco vocales pronunciadas de manera aislada y sostenida por cada locutor, se estimaron las longitudes para cada vocal como el promedio de las longitudes de todas sus tramas, y finalmente se halló la media de las longitudes de todas las vocales para obtener la LTV final de cada locutor.

Los resultados de aplicar ésta técnica en el corpus de voz infantil se muestra en la Figura 6.6 para locutores masculinos y en la Figura 6.7 para locutores femeninos.

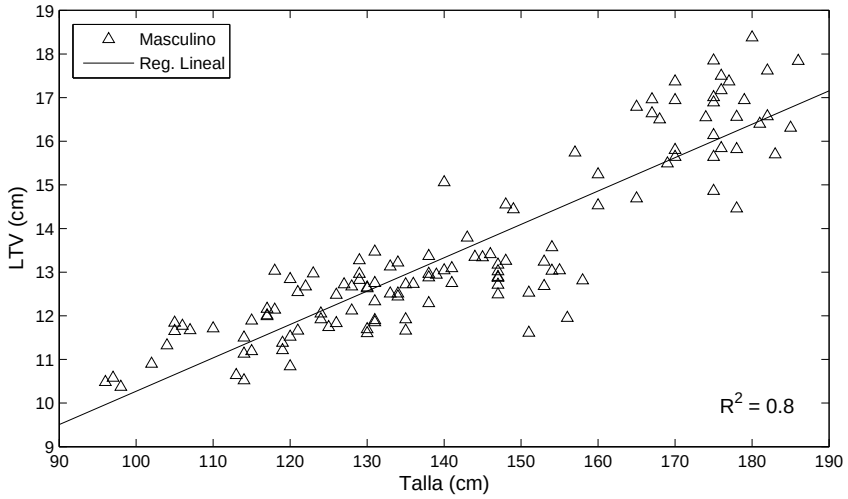


Figura 6.6: *LTV Estimada para 125 locutores masculinos.*

La imagen de los locutores masculinos muestra una alta correlación ($R^2 = 0.8$) entre la talla del locutor y la longitud de su tracto vocal, también se observa como ellos alcanzan una mayor talla respecto a los locutores femeninos reflejado también en las longitudes del tracto vocal estimadas. En los locutores femeninos, también existe una alta correlación entre la talla y la LTV ($R^2 = 0.66$) pero las longitudes máximas difícilmente superan los 15 cm de longitud únicamente, lo que explica que en general los formantes femeninos suelen estar por encima de los formantes masculinos ya que el tracto vocal de ellas es más corto.

Hasta el momento, se han estimado formantes fiables en la voz infantil y a partir de estos se obtuvo una relación lineal entre la talla y la longitud del tracto vocal, tanto para los locutores masculinos como los femeninos. Como lo explica la siguiente Sección 6.3, con ésta información es posible normalizar los formantes ya estimados con el objetivo de reducir la alta variabilidad formántica existente en la voz infantil debida principalmente a las diferentes longitudes de sus tractos.

Para conocer el comportamiento de la técnica de estimación de la longitud del tracto en voz adulta, ésta se aplicó a la base de datos AVACAR [Ortega et al., 2004] con en objetivo

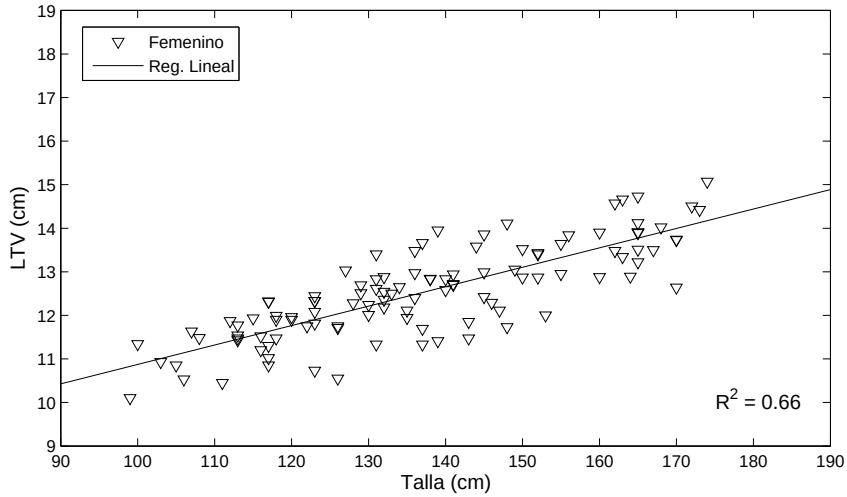


Figura 6.7: *LTV Estimada para 110 locutores femeninos.*

de estimar la LTV de los locutores que la componen. Éste corpus consta de 9 locutores femeninos y 11 locutores masculinos todos adultos y de cada uno de ellos se grabaron 18 frases en español leídas de manera espontánea.

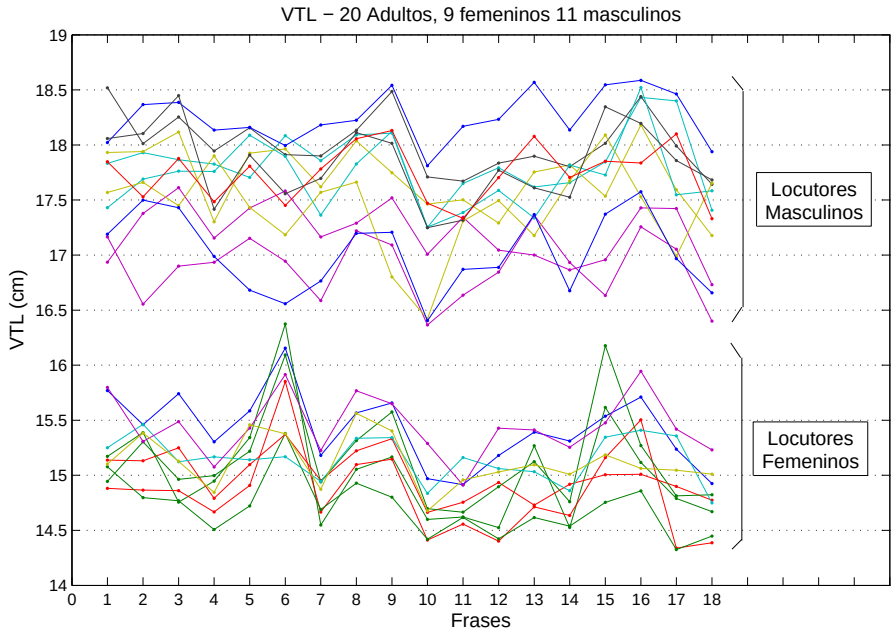


Figura 6.8: *LTV Estimada para 20 locutores adultos.*

La prueba se hizo de manera ciega, es decir, sin ningún orden específico y sin información a priori de los ficheros que pertenecen a locutores masculinos o femeninos. Los resultados de dicha prueba se muestran en la Figura 6.8, allí cada línea representa la estimación de la longitud del tracto vocal de cada locutor para las 18 frases. Se puede observar claramente como la estimación de las longitudes divide el conjunto de locutores en locutores femeninos y masculinos, información que fue corroborada manualmente con la documentación del corpus

y la coincidencia fue del 100%.

La variación en los valores de longitud depende principalmente del número de vocales abiertas o cerradas presente en la frase, pues en las vocales cerradas como la /o/ y la /u/ la estimación de la longitud del tracto es mayor por el efecto redondeado de los labios ([Wakita, 1977], [Paige and Zue, 1969]). Las pruebas realizadas en el corpus de voz infantil y en el de adultos con AVACAR demuestran, que la técnica planteada para estimar la LTV de un usuario a partir de los formantes de sus vocales, funciona y es un buen indicador de características acústicas del locutor, y que existe una alta correlación entre ésta longitud y la talla del usuario en el caso de la población infantil.

6.3 Normalización de Formantes

Con los formantes ya estimados, y teniendo una buena aproximación de la longitud del tracto vocal, es posible normalizar los formantes utilizando dicha longitud como lo propone [Wakita, 1977] en el caso de adultos. Tener los formantes en un espacio normalizado según el tracto vocal del usuario, disminuye la variabilidad y permite trabajar con estos formantes en un espacio más homogéneo, además de encontrar patrones de comportamiento de los formantes en función de la talla y sexo, ésto facilitará en gran medida el desarrollo de herramientas donde se utilicen formantes infantiles de una manera más precisa y realista.

La técnica de normalización se basa en la hipótesis de que la configuración del tracto vocal en emisiones vocálicas entre locutores es semejante pero difiere en términos de su longitud. Para ello, se calculan los formantes de un tubo acústico cuando su longitud LTV es variada a una longitud de referencia l_R sin alterar su forma. Como muestra la ecuación 6.19, los formantes normalizados F_{kN} se hallan multiplicando los formantes inicialmente calculados $\tilde{F}_k$, por el factor $\frac{LTV}{l_R}$, siendo l_R una longitud de referencia fijada en $17.5cm$.

$$F_{kN} = \frac{LTV}{l_R}\tilde{F}_k \qquad (k = 1, ..., M) \tag{6.19}$$

Aplicado la expresión 6.19 a los formantes del corpus de voz $\tilde{F}_k$, los nuevos formantes normalizados F_{kN} se muestran en la Figura 6.9. En (a) y (c) se encuentran los formantes normalizados para los 125 locutores masculinos y los 110 locutores femeninos respectivamente, y en (b) y (d) las medias y varianzas para cada vocal para los locutores masculinos y femeninos respectivamente.

La primera comparación apreciable se puede hacer entre los formantes no normalizados $\tilde{F}_k$ de la Figura 5.12 y los formantes normalizados F_{kN} de la Figura 6.9. Las partes (a) y (c) muestran notorias diferencias en especial en los locutores masculinos, donde la gran dispersión presente en las vocales /e/ e /i/ en la Figura 5.12, han disminuido en el espacio normalizado de la Figura 6.9, en general, los formantes se encuentran menos dispersos tras la normalización en todas las vocales para ambos sexos, aunque se debe tener en cuenta que la escala de frecuencia se ha alterado $\frac{LTV}{l_R}$ veces.

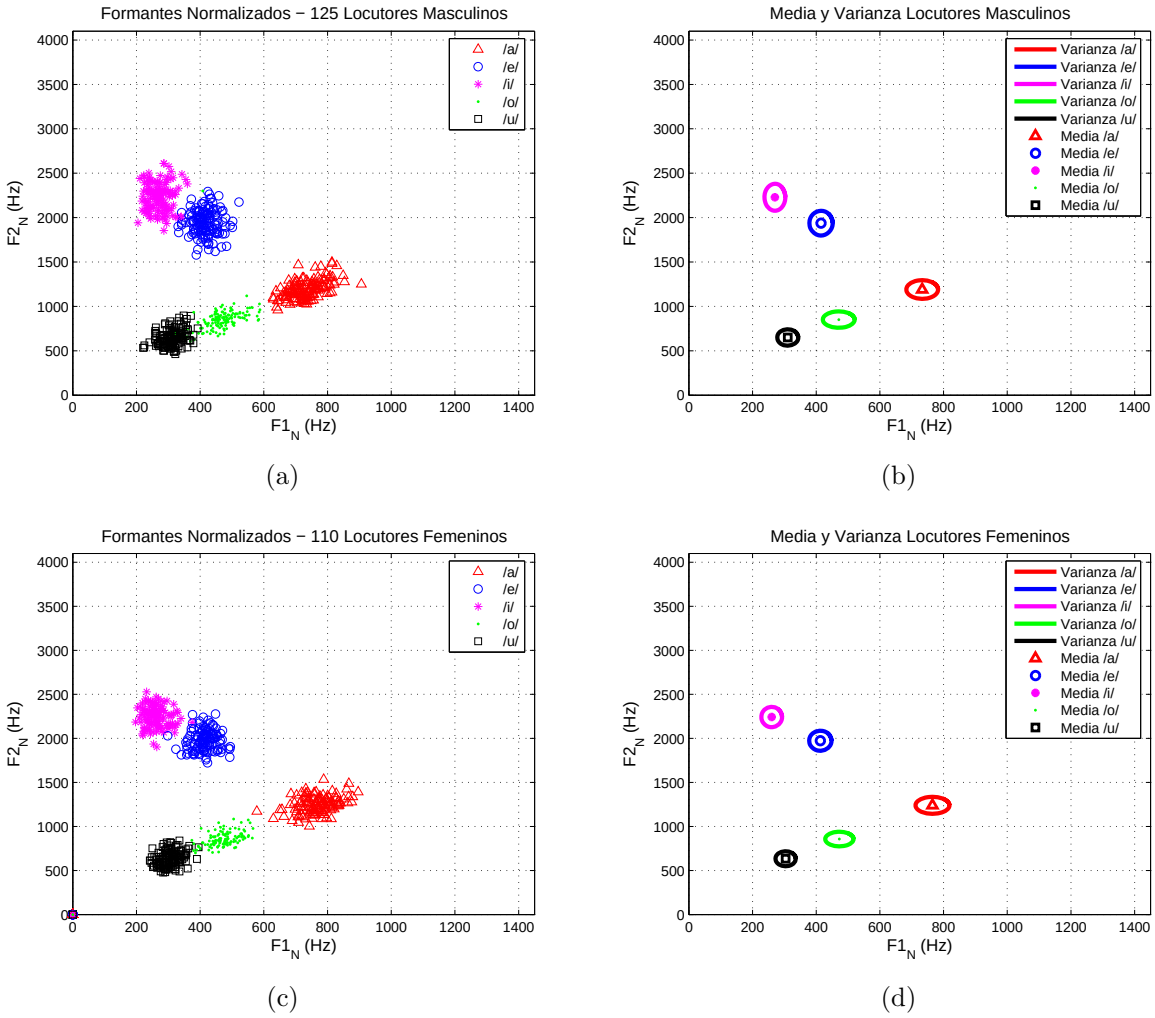


Figura 6.9: *Formantes vocálicos normalizados, media y varianza para locutores masculinos en (a) y (b), y locutores femeninos en (c) y (d).*

En cuanto a los valores de medias y varianzas de los formantes sin normalizar y normalizados (partes (*b*) y (*d*)), estos últimos se encuentran muy próximos y no se aprecian mayores diferencias entre los locutores femeninos y masculinos, en contraste con los valores sin normalización. Se puede concluir que el proceso de normalización de formantes reduce considerablemente la variabilidad inter-locutor, y brinda robustez a los algoritmos de tratamiento de voz a la hora de enfrentarse a voz infantil y diseñar herramientas indistintamente para niños o niñas con tallas diferentes.

Hasta aquí y a manera de resumen, el tratamiento aplicado sobre la señal de voz para obtener sus parámetros acústicos de manera robusta, puede apreciarse en la Figura 6.10. A partir de la etapa de pre-procesado se pueden estimar parámetros acústicos como: la energía de la señal de voz en los segmentos sordos y sonoros, la frecuencia fundamental o pitch a través del análisis LPC, y los formantes normalizados gracias al análisis LPC y homomórfico, y a la correlación encontrada entre la longitud del tracto vocal y la talla del locutor.

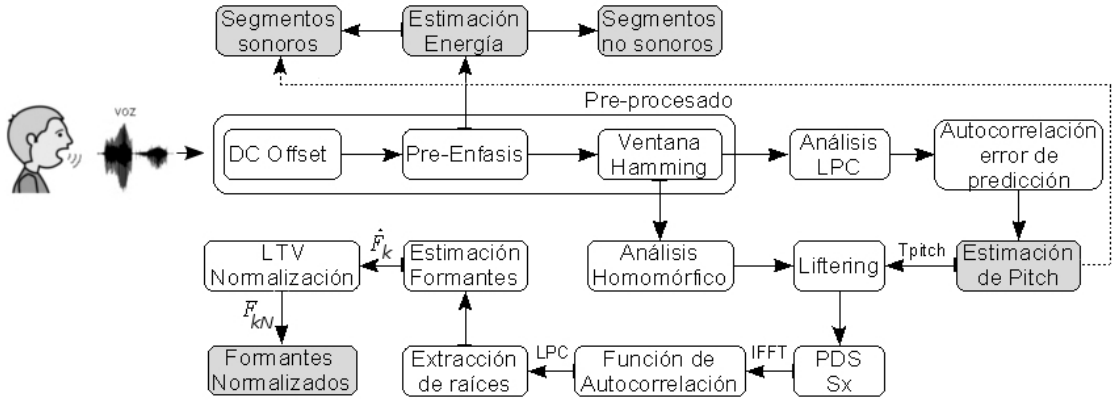


Figura 6.10: *Diagrama de bloques - Tratamiento sobre la señal de voz.*

Con la aplicación de ésta tecnología y una fase ardua de desarrollo, será posible crear herramientas para terapia de voz como se explica en el Capítulo 7. También, el trabajo aquí propuesto para estimar la longitud del tracto vocal directamente de los formantes de cada locutor, se aplicará en la estimación On-line de un factor de deformación frecuencial para reconocimiento automático del habla, comparando los resultados obtenidos con la técnica de normalización de máxima verosimilitud ML-VTLN en la base de datos TIDigits, los resultados obtenidos en dicha aplicación se describen en el Capítulo 8.

Parte III

Aplicación y Desarrollo

Capítulo 7

Herramientas para Terapia de Voz

El objetivo fundamental de esta tesis es el desarrollo de herramientas libres para profesionales de logopedia y de educación especial, con las que puedan educar la voz en población infantil. La educación de la voz implica por una parte la rehabilitación o re-educación de ésta cuando por alguna razón (como una enfermedad temporal) se ha alterado, ó, como en algunos casos de población infantil con discapacidad, en donde se requiere de la educación de la voz ya que ésta se encuentra alterada desde un principio por la condición discapacitante.

Este capítulo describe las herramientas desarrolladas las cuales aprovechan los avances obtenidos en el tratamiento de la voz infantil de los capítulos anteriores, y que también son posibles, gracias a las continuas retribuciones de los profesionales que han probado las versiones de desarrollo y de prueba. Estas herramientas se centran en el proyecto "COMUNICA" [Rodríguez et al., 2008], [Saz et al., 2008], el cual reúne un conjunto de aplicaciones que busca favorecer el desarrollo del lenguaje desde los niveles más básicos hasta los más elevados. Las herramientas del proyecto "COMUNICA" se encuentran disponibles en la página web www.vocaliza.es y hasta Agosto de 2010, se han registrado al rededor de 6000 descargas de sus herramientas, lo cual demuestra un elevado grado de aceptación por parte de los usuarios finales quienes en su mayoría pertenecen a países latinoamericanos.

La parte de tratamiento de señal de las herramientas aquí presentadas se desarrollo con algoritmos en lenguaje C, mientras que para la parte gráfica se utilizó un motor gráfico de uso libre llamado Allegro[1], cuyas generalidades se presentan en el Anexo A.

La primera herramienta descrita en la Sección 7.1 se denominada *PreLingua*, esta es la herramienta más completa y hace uso de todos los avances en las técnicas de procesado de señal alcanzados en esta tesis. En el último año de investigación surge la herramienta ARTICULA, la cual se describe en la Sección 7.2 y surge gracias a las mejoras hechas sobre los algoritmos iniciales de *PreLingua* y, sobre todo, atendiendo las continuas demandas de los profesionales del campo de poder contar con una herramienta para articulación vocálica por medio de una interfaz entendible para los niños. la Sección 7.3 describe la herramienta ViVo, destinada a todos aquellos interesados en trabajar y conocer los aspectos acústicos de la voz en tiempo real por medio de una interfaz simple. Finalmente, Sección la 7.4 describe la herramienta llamada VocalClick, la cual emula los movimientos y algunos eventos del puntero del ratón por medio de sonidos vocálicos, aprovechando la estimación robusta de

[1] http://www.liballeg.org Allegro

formantes del Capítulo 5 y su normalización descrita en el Capítulo 6.

7.1 PreLingua

Un niño sano durante su primer año de vida adquiere ciertas habilidades de comunicación conocidas como habilidades pre-lingüísticas o pre-lenguaje, tal y como se describe en la Sección 2.1.1. Dentro de ellas, las relacionadas con la voz incluyen la detección de Actividad de Voz, el control de la Intensidad, Tono, y Soplo y finalmente el control de la fonación con las primeras producciones vocálicas. Con estas habilidades adquiridas, el niño tiene las herramientas necesarias para continuar con la evolución de su lenguaje a nivel fonológico, semántico y pragmático en los posteriores años de vida.

Lamentablemente no siempre ocurre así, un niño con discapacidad puede tener problemas en el desarrollo de su pre-lenguaje, y por consiguiente en la adquisición posterior de su lenguaje, vemos casos de niños que llegan a la pubertad y no tienen un desarrollo pre-lingüístico adecuado, es decir, no controlan por ejemplo la tonalidad o la intensidad de su voz, o lo hacen con gran dificultad. También, en otras ocasiones debido a su condición discapacitante como las malformaciones, el niño presenta alteraciones en su voz que finalmente limitan sus habilidades de comunicación.

Es así como la aquí herramienta desarrollada busca trabajar aquellos aspectos del pre-lenguaje y acústicos de la voz, susceptibles de ser tratados por medio de Tecnologías del Habla, por ejemplo: La detección misma de Actividad de Voz, el control de la Intensidad, el Soplo, la Tonalidad y finalmente la Vocalización. También la herramienta se ha complementado con actividades para un mejor control vocal como el Ataque vocal y la Duración de sonidos.

La herramienta esta dividida en cinco niveles como muestra la Figura 7.1, y cubre todos los aspectos descritos anteriormente siguiendo tres enfoques: la detección de actividad de voz, el control y modulación de parámetros acústicos y la articulación vocálica.

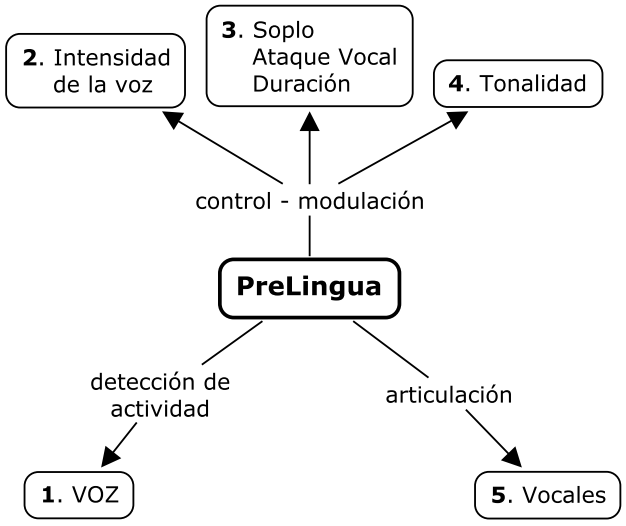


Figura 7.1: *Niveles en PreLingua*

La interfaz principal de la herramienta es mostrada en la Figura 7.2, en ella, los cinco niveles están organizados en forma piramidal en lo que a complejidad respecta. La base o nivel 1 trabaja la DETECCIÓN DE VOZ y busca que el niño cree conciencia de su propia voz y de que con ella puede interactuar con su entorno y comunicarse. Cuando el niño es consciente de su voz, el puede aprender a modular la INTENSIDAD utilizando los juegos del nivel 2, el cual le presenta diferentes escenarios y opciones de mayor o menor complejidad. El nivel 3 reúne los aspectos relacionados con la correcta respiración tan necesarios en una buena comunicación oral, se puede trabajar el SOPLO únicamente, es decir sin la generación de sonidos sonoros, el ATAQUE VOCAL para controlar la apertura y cierre glóticos y finalmente, la DURACIÓN de los sonidos sordos y sonoros.

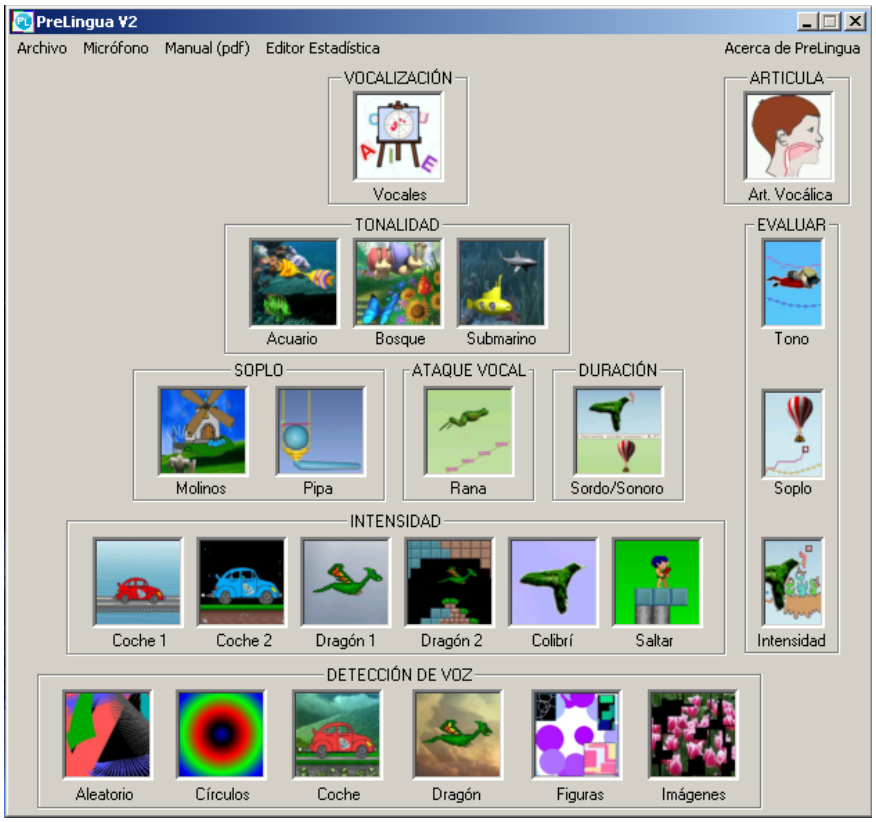


Figura 7.2: *Pantalla Principal de PreLingua*

El nivel 4 permite trabajar la modulación de la TONALIDAD, ya que la entonación es parte importante del mensaje a transmitir en la comunicación oral, finalmente, el nivel 5 trabaja la articulación de las vocales del español con VOCALIZACIÓN y ARTICULA.

En general, las actividades han sido diseñadas con la intención de iniciar en la base de la pirámide e ir ascendiendo en la medida de los avances de cada usuario, sin embargo, el terapeuta puede trabajar cualquier actividad en cualquier orden en función de las necesidades y capacidades de cada usuario. Los primeros cuatro niveles no requieren de previas configuraciones, basta con hacer clic con el puntero del ratón sobre la imagen de la actividad deseada y ésta se iniciará de manera inmediata, mientras que las actividades del nivel 5 requieren una mínima configuración de sexo y talla del usuario para optimizar su

funcionamiento

PreLingua incluye una sección denominada EVALUAR en la parte derecha de la pirámide, este conjunto de actividades permite evaluar el desempeño del niño en el control de la INTENSIDAD, SOPLO y la TONALIDAD, por medio juegos con un objetivo específico en el que se mide la diferencia entre patrones establecidos por el terapeuta y los generados por el usuario durante la sesión de trabajo. Para facilitar esta tarea, el sistema entrega un reporte de texto y una imagen con información estadística de cada sesión, de manera que el terapeuta puede hacer un fácil seguimiento de cada usuario, complementar la historia clínica y tomar decisiones sobre el tratamiento en función de los resultados obtenidos parcialmente.

Dentro de *PreLingua*, los algoritmos que hacen posible cada actividad se basan en las técnicas de procesado de voz de los Capítulos 5 y 6. El diagrama de bloques de la Figura 6.10 del Capítulo 6, se convierte ahora en el diagrama de bloques de la Figura 7.3 al que se le han adicionado los bloques de la parte de desarrollo.

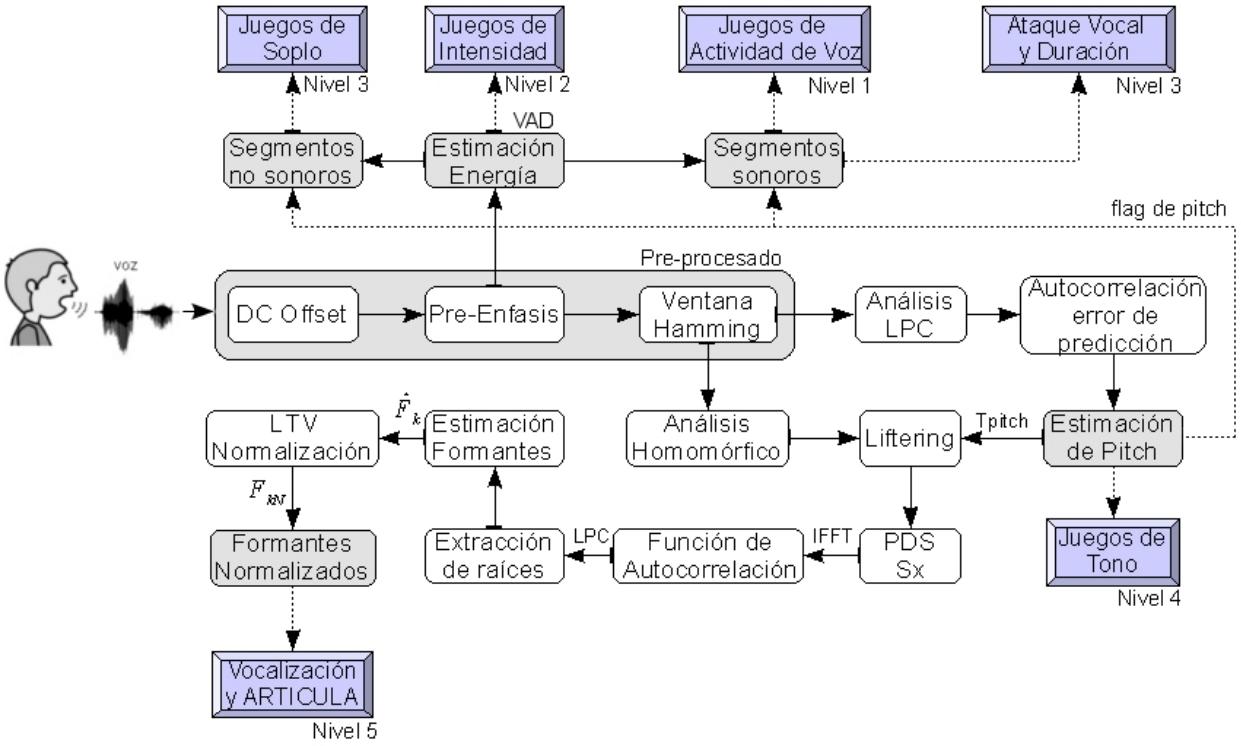


Figura 7.3: *Diagrama de bloques de PreLingua.*

EL nuevo diagrama de bloques muestra el origen de los cinco niveles de la pirámide partiendo básicamente de: la estimación de la energía de la señal tanto de segmentos sonoros como sordos, la estimación de la frecuencia de pitch y, la estimación robusta de formantes. En las sub-secciones siguientes, se explica en detalle como funciona cada nivel y se describen algunas de sus actividades.

7.1.1 DETECCIÓN DE VOZ

Un niño con problemas de comunicación a nivel pre-lingüístico o pre-lenguaje no diferencia los sonidos de su entorno de la voz humana, y por consiguiente no advierte que puede usar su propia voz para comunicarse. El nivel 1 de *PreLingua* permite trabajar esta área de la comunicación a través de seis actividades (Figura 7.4) muy simples que reaccionan ante la presencia de voz.

Figura 7.4: *Nivel 1 - DETECCIÓN DE VOZ.*

Todas las actividades hacen uso de un VAD de energía como el mostrado en la Figura 7.5, en el que si la energía estimada en el segmento sonoro supera el umbral pre-establecido, el sistema genera una señal cuadrada que vale 1 si hay actividad de voz y 0 en caso contrario, este umbral se establece por defecto al lanzar la aplicación donde se estima la energía de las primeras tramas de sonido que entran al sistema. Durante el funcionamiento de la aplicación, este umbral puede ser cambiado a criterio del terapeuta con el objetivo de aumentar o disminuir el nivel de exigencia en la sesión de trabajo.

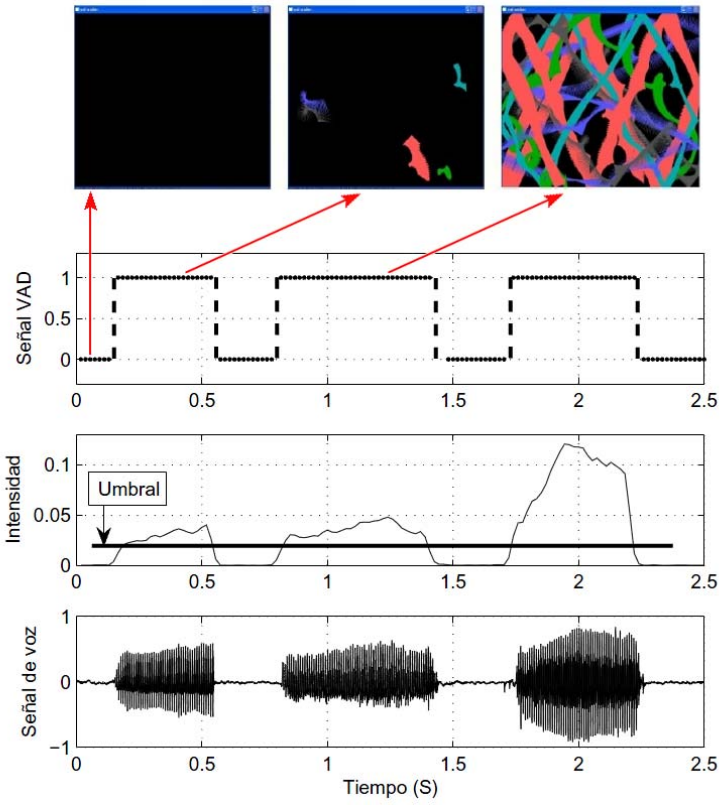


Figura 7.5: *VAD en la Activación de Imágenes.*

Como se aprecia en la Figura 7.5 en la parte superior, la pantalla inicial esta neutra y es cuando se establece el umbral de energía, a partir de allí a medida que el VAD entrega una señal de alto nivel es decir, aquellas tramas donde hay presencia de voz, el sistema dibuja en pantalla figuras geométricas de colores aleatorios que se desplazan en pantalla siempre y cuando exista voz.

Las distintas actividades de este nivel: *Aleatorio*, *Círculos*, *Coche*, *Dragón*, *Figuras* e *Imágenes*, tienen el mismo principio de funcionamiento cambiando solamente la interfaz gráfica. Este nivel tiene varias actividades ya que es la primera experiencia que tiene el niño con la herramienta en general, y entre más actividades se tengan mayor probabilidad de aceptación por parte del niño.

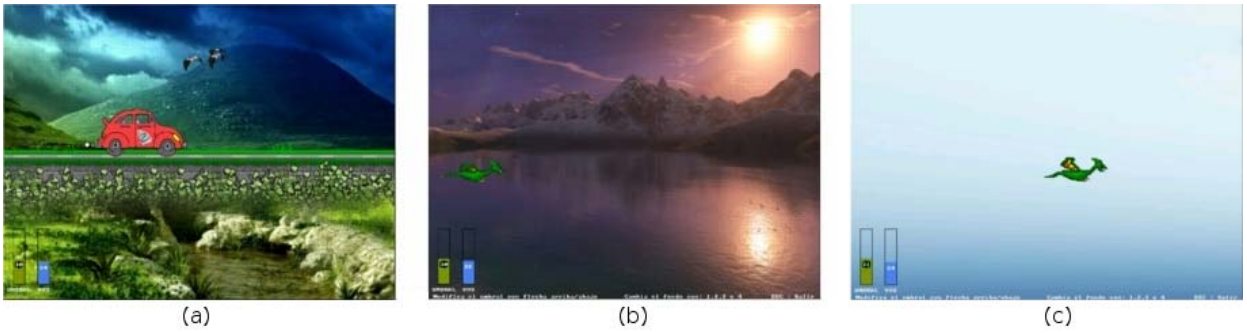


Figura 7.6: *Actividades de Coche (a) y Dragón en dos Escenarios (b) y (c).*

La Figura 7.6 muestra las actividades de *Coche* (a) y *Dragón* en dos escenarios (b) y (c), aquí la señal de activación del VAD se convierte en el movimiento horizontal de estos personajes. Para el caso del *Dragón* la misma aplicación permite cambiar la imagen del escenario a través del teclado numérico, esta opción se habilitó después de recibir observaciones de terapeutas manifestando la dificultad de algunos usuarios con deficiencia visual, para diferenciar el personaje principal que esta en movimiento como en la imagen (b), de manera que en su lugar pueden utilizar un escenario más adecuado como el mostrado en la imagen (c).

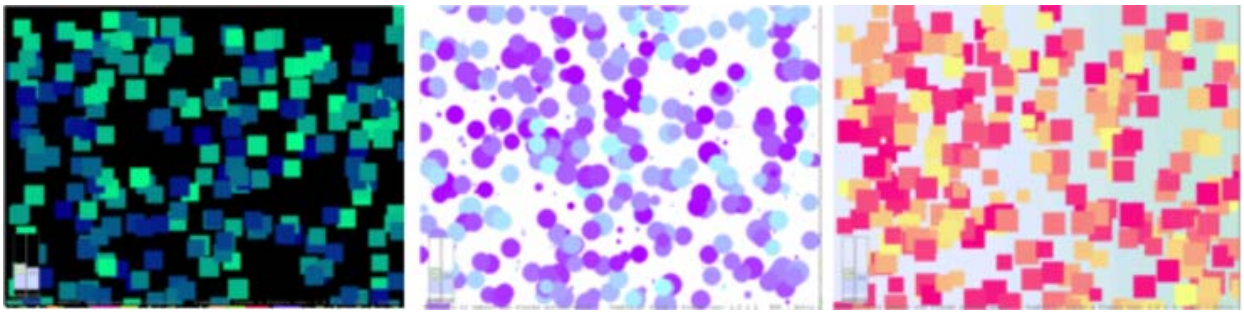


Figura 7.7: *Figuras Geométricas.*

La actividad de *Figuras* muestra en pantalla formas geométricas en presencia de voz como en la Figura 7.7, pero en ellas, se empieza a introducir el concepto de intensidad de la

voz ya que el valor de la energía de la señal de voz estimado, afecta directamente el tamaño de la figura geométrica, de manera que a mayor intensidad en la señal de voz mayor tamaño tendrán las figuras geométricas.

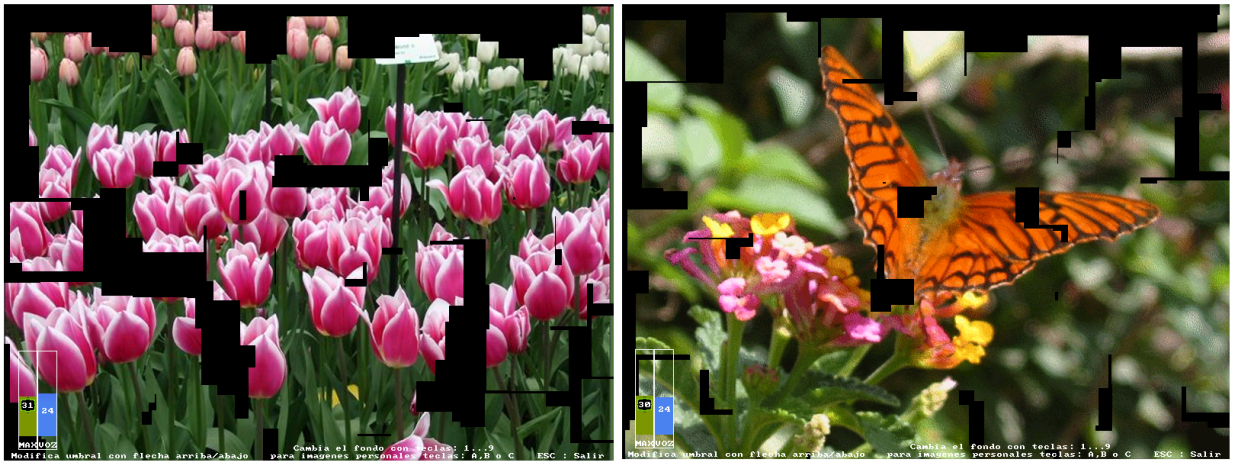


Figura 7.8: *Imágenes a Descubrir con la Voz.*

La actividad *Imágenes* consiste en ir descubriendo una imagen oculta con la voz como lo muestra la Figura 7.8, aquí la señal de activación del VAD va dibujando en pantalla una serie de imágenes ya integradas al sistema por medio de rectángulos que contienen dicha imagen. Esta actividad puede resultar muy motivadora para el niño o usuario, si la imagen a descubrir es la propia imagen de la persona o la de su personaje favorito, ya que existe la posibilidad de cargar imágenes personalizadas para tal fin.

En conjunto, las actividades del nivel 1 han sido muy bien valoradas por los usuarios que las han utilizado, e incluso, algunos manifiestan resultados positivos en niños que no eran candidatos iniciales para utilizar *PreLingua*, han visto buenos resultados utilizando la herramienta en áreas como la estimulación temprana y captura de atención en niños con profundas discapacidades cognitivas.

7.1.2 INTENSIDAD

Figura 7.9: *Nivel 2 - INTENSIDAD.*

Este nivel tiene seis actividades como se muestra en el Figura 7.9, en este nivel se espera que el niño ya tenga la habilidad de distinguir su propia voz para que ahora aprenda a

modular la intensidad de la misma.

Para conseguirlo, el sistema utiliza el valor de la estimación de la energía de la señal, y se lleva a un espacio gráfico de valores en píxeles para conseguir una proporcionalidad entre el valor de la intensidad y el movimiento de objetos en pantalla.

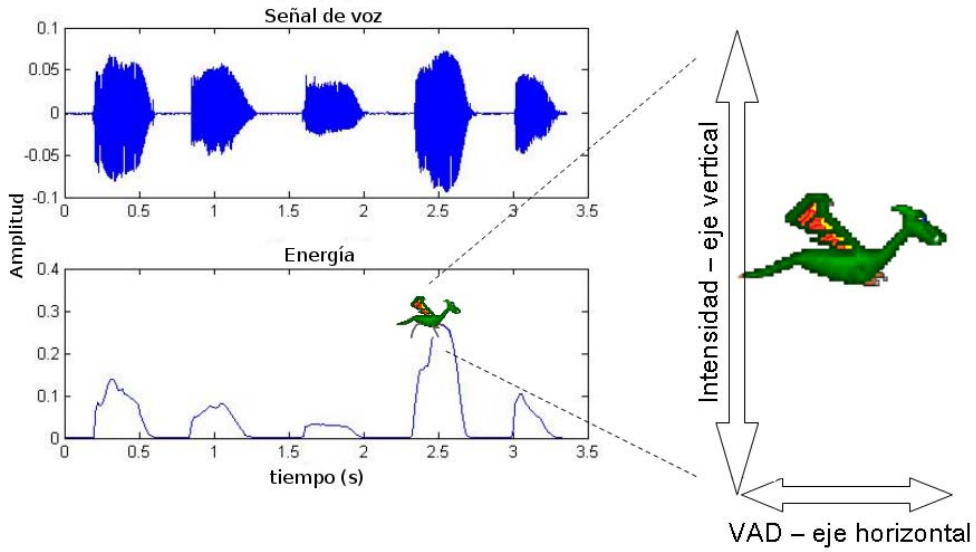


Figura 7.10: *Intensidad de la Voz a Posición Vertical.*

La figura 7.10 muestra en la parte superior una señal de voz y su respectiva estimación de la energía en la parte inferior, éste valor se convierte en la posición vertical del objeto animado de manera que un incremento en la intensidad de la voz, se convierte en un incremento de posición vertical y viceversa; el movimiento del eje horizontal es constante y activado por el VAD.

Figura 7.11: *Actividades de Coche1 (a) y Dragón2 (b).*

Para inducir al niño en la modulación de la intensidad de su voz, hay juegos como *Dragón*2 mostrado en la Figura 7.11(b) donde el objeto animado debe evadir obstáculos y

desplazarse a través de un laberinto para encontrar con su amada dragona, allí la trayectoria es única y el niño debe variar la intensidad de la voz para llegar al final del juego. La parte (a) de la misma figura muestra la actividad de *Coche*1, donde la intensidad de la voz se transforma en la velocidad horizontal del móvil, lo cual ayuda a asimilar muy bien el concepto de intensidad o fuerza en la producción sonora en el usuario. Cuando las actividades finalizan el sistema recompensa al niño con aplausos y fuegos artificiales ya que motivar el buen desempeño hecho por el niño es muy importante para obtener mejores resultados.

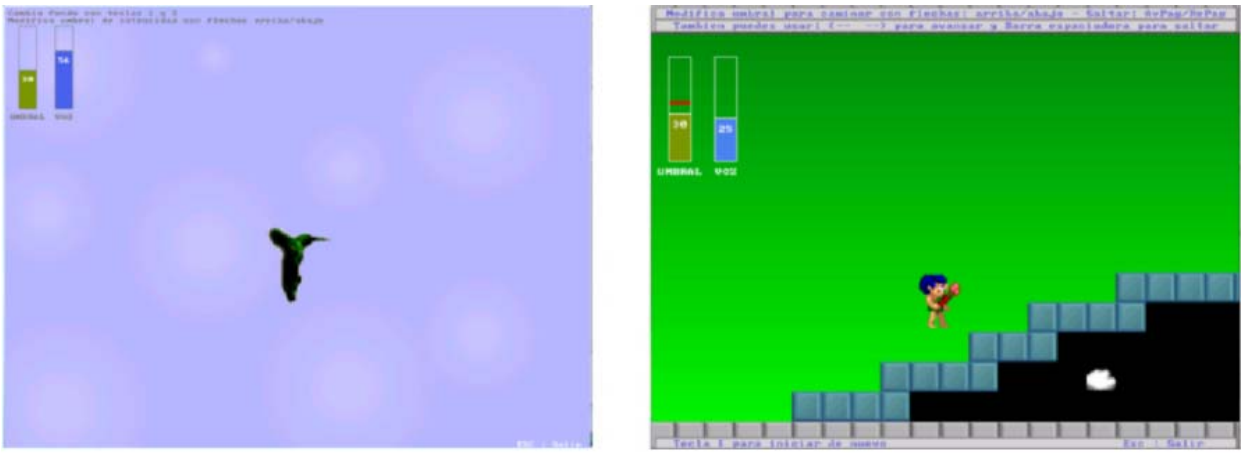


Figura 7.12: *Actividades de Colibrí y Saltar.*

La actividad de *Colibrí* de la Figura 7.12(a), funciona de manera más simple y esta diseñada para aquellos usuarios con deficiencias visuales y cognitivas más severas, el juego no tiene un objetivo específico, sencillamente el colibrí varia la posición vertical en función de la intensidad y el desplazamiento horizontal es constante. La actividad de la parte (b) de la misma figura llamada *Saltar*, es por el contrario la actividad más exigente y requiere un muy buen control de la intensidad de la voz. La actividad consiste en llevar al personaje hasta el final del tablero donde lo espera su amada, pero debe sortear una serie de obstáculos a manera de rampas en el camino. El personaje se controla con la voz y se manejan dos umbrales de intensidad, el primero permite que el personaje camine y el segundo que salte, de manera que controlando debidamente la intensidad de la voz es posible llevar al personaje hasta el final. Estos umbrales son modificables en cualquier momento para permitir variar el nivel de exigencia requerido.

7.1.3 SOPLO

El nivel 3 se compone de tres partes, la primera trabaja el SOPLO, la segunda el ATAQUE VOCAL y finalmente la tercera la DURACIÓN de los sonidos. SOPLO a su vez posee dos actividades *Molinos* y *Pipa de Soplar* como muestra la Figura 7.13. Hablar fluidamente requiere de una correcta respiración, y la modulación del soplo ayuda en tarea, en las actividades de SOPLO se busca que el niño aprenda a modular la intensidad de este, sin la generación de sonidos sonoros, es decir soplando hacia el micrófono.

Figura 7.13: *Nivel 3 - SOPLO.*

La Figura 7.14 muestra como en la actividad de *Molinos* la intensidad del soplo se transforma en la velocidad de rotación de las hélices, la figura muestra una señal de soplo seguida de la señal sonora generada al pronunciar la vocal /a/, ambas señales poseen energía pero la diferencia esta en que la señal sonora posee pitch mientras que la de soplo no, este es el flag de pitch presente en el diagrama de bloques de la Figura 7.3, de esta manera el sistema controla que el movimiento del objeto en pantalla dependa solo de la intensidad estimada en las tramas que no tienen pitch, si el niño grita (situación no deseada), el sistema lo detecta y detiene la rotación de los molinos.

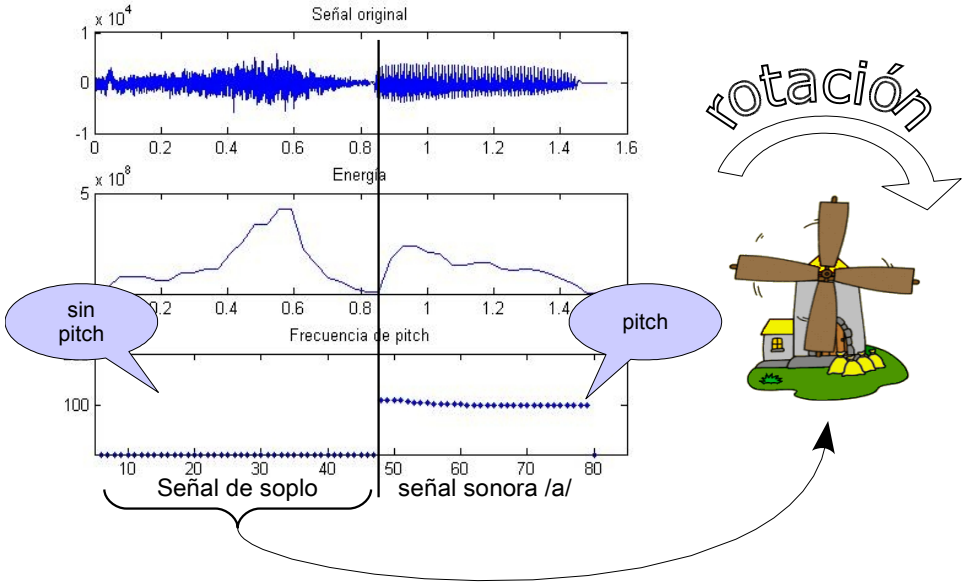


Figura 7.14: *Intensidad del Soplo a Rotación.*

La Figura 7.15(a) muestra la actividad de *Molinos*, la cual esta diseñada sencillamente como realimentación visual de la intensidad del soplo. En la parte (b) de la misma figura se muestra la actividad *Pipa de Soplar*, la cual simula la actividad de soplar a través de una pipa como ocurre en la realidad, para que una esfera se eleve dentro de un cilindro.

Esta última actividad es más exigente ya que requiere de un correcto control del soplo para lograr mantener la esfera a una altura determinada, un indicador situado en la pared izquierda del cilindro se ilumina cuando la esfera se encuentra a su mismo nivel, si el soplo no se modula adecuadamente y se sobrepasa el límite, la esfera acciona el mecanismo y un tomate caerá sobre el personaje de la derecha.

(a) (b)

Figura 7.15: *Actividad de Molinos (a) y Pipa de Soplar (b).*

7.1.4 ATAQUE VOCAL

Figura 7.16: *Nivel 3 - ATAQUE VOCAL Y DURACIÓN.*

ATAQUE VOCAL (Figura 7.16-izquierda) permite trabajar el inicio de la sonoridad controlando la apertura y cierre glóticos. La actividad denominada *Rana* consiste en hacer que esta salte con cada golpe de voz por una serie de bases a los largo de una trayectoria definida, hasta que llegue al final donde su compañera le esta esperando. Esta actividad mostrada en la Figura 7.17, resulta especialmente útil para trabajar trastornos del habla como la disfemia o tartamudez.

La aplicación toma la energía de los segmentos sonoros y permite la modificación del umbral de activación para su funcionamiento, permite también que el espacio entre las bases de la trayectoria cambie según las necesidades de trabajo de cada usuario. Este cambio se hace a través de los cuatro puntos de control ubicados a lo largo de la trayectoria, los puntos de control son círculos que pueden ser desplazados con el puntero del ratón en cualquier dirección modificando la exigencia en la generación, mantenimiento, y repetición de los sonidos sonoros.

Si la rana cae porque por no tener donde posarse, la actividad se reinicia y la rana vuelve al punto de partida, cuando la rana llega al final, el sistema muestra en pantalla el tiempo total de la sesión y el tiempo total de fonación, es decir la sumatoria en tiempo de la totalidad de las tramas sonoras, lo que en conjunto brinda información útil al terapeuta.

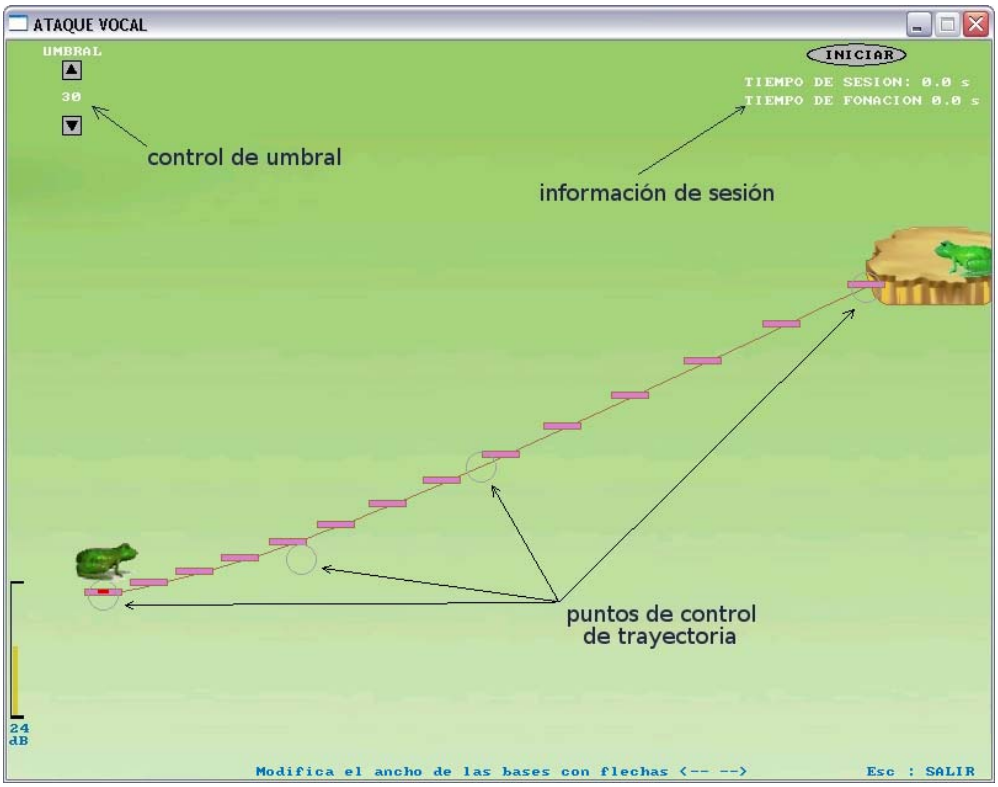


Figura 7.17: *Actividad Rana.*

7.1.5 DURACIÓN

La actividad de DURACIÓN a la que se accede haciendo click en el icono de la Figura 7.16-derecha, es una actividad que permite conocer el Tiempo Máximo de Espiración TME y el Tiempo Máximo de Fonación TMF, y el cociente entre ellos TME/TMF, ya que son de gran valor diagnóstico para el profesional de la voz pues ayuda a valora de la eficiencia del cierre glótico. En la consulta foniátrica, el profesional toma estos tiempo manualmente con un cronómetro y calcula el cociente para registrarlo en la historia clínica, es así como aprovechando los algoritmos ya existentes en *PreLingua* se diseño esta actividad para apoyar la labor del profesional.

Para el TME se puede utilizar la ese sorda /sss../ y para el TMF la zeta sonora /zzz../, también se pueden utilizar las fricativas labiodentales sorda como la /fff../ y sonora /vvv../ o los sonidos que el terapeuta crea convenientes en función de la capacidad de fonación del niño. La actividad consiste en motivar al niño a pronunciar la ese sorda /s/ el máximo tiempo posible después de una profunda inspiración, este hecho mueve un globo en pantalla como muestra la Figura 7.18 en dirección horizontal, mientras el sistema va registrando el tiempo consumido. Después de un breve descanso, se repite el procedimiento anterior pero pronunciando la zeta sonora /z/ el máximo tiempo posible, ya que este sonido tiene pitch el sistema lo identifica como sonoro y moverá el colibrí también en dirección horizontal. Cuando finaliza el procedimiento, el sistema muestra en pantalla el tiempo total en segundos de cada prueba y el cociente calculado, se consideran valores normales los cercanos a 1 y si es superior a 1.4 se considera un indicador clínico que requiere atención [Vila, 2009].

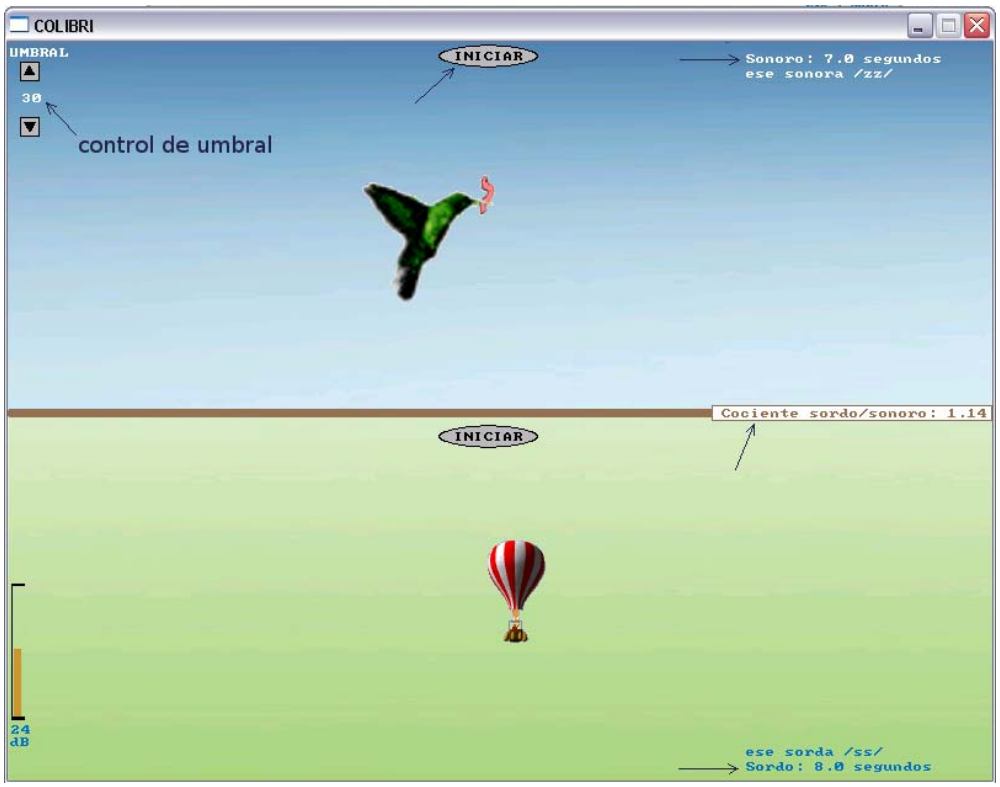


Figura 7.18: *Actividad Sordo/Sonoro.*

El buen funcionamiento de esta actividad depende de un umbral de intensidad adecuado, el cual puede ser modificado con el respectivo control en la parte superior izquierda de la actividad.

7.1.6 TONALIDAD

Figura 7.19: *Nivel 4 - TONALIDAD.*

Con una filosofía similar al control de la Intensidad, esta nivel busca que el niño aprenda a modular el tono de su voz con las actividades mostradas en la Figura 7.19.

Este nivel posee tres actividades: *Acuario*, *Bosque*, y *Submarino*, en las cuales los personajes a controlar con la voz son diferentes a los del nivel 2 para no mezclar los conceptos ya trabajados. La Figura 7.20 muestra como los tres personales de este nivel: el pez, la mariposa y el submarino, utilizan el valor de pitch estimado para variar la posición vertical del personaje respectivo, y la presencia de voz misma permite el movimiento horizontal.

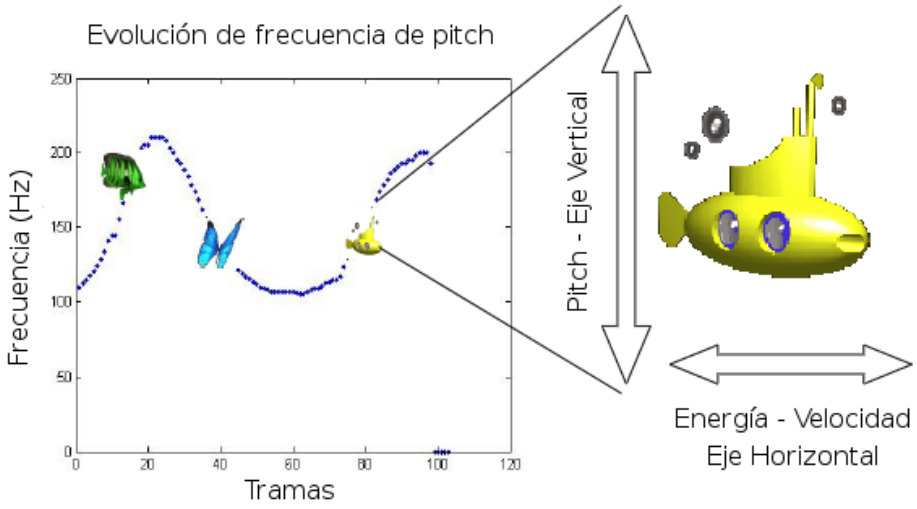


Figura 7.20: *Figuras controladas con el Tono.*

Como se analizó en el Capítulo 5, la frecuencia de pitch en la población infantil es muy superior a la del adulto y va disminuyendo a medida que el infante crece, de manera que el sistema trabaja en el rango de frecuencias entre 80 Hz y 420 Hz, una vez el sistema estima la frecuencia de pitch, este valor pasa al motor gráfico y modifica la posición vertical de los diferentes personajes.

Figura 7.21: *Actividad de Acuario (a) y Bosque (b).*

La Figura 7.21(a) muestra la actividad de *Acuario*, en la que un pez verde (dentro del círculo) varia su posición vertical en función del tono, y el movimiento horizontal es constante y activado por la presencia de voz. El objetivo es seguir a los demás animales como el pulpo y otros peces que inicialmente se encuentran estáticos, y al acercarse el pez del niño (pez verde) a los otros animales, estos se animan y se mueven en diferentes trayectorias, de manera que el niño debe seguirlos modificando la tonalidad de su voz.

En la actividad de *Bosque* (Figura 7.21(b)), una mariposa debe volar para descubrir

los otros animales del escenario. Los animales están inicialmente estáticos y al acercarse la mariposa a ellos, estos se animan (se mueven, saltan, oscilan), el niño podrá apreciar estos cambios solo si se acerca a los animales para lo cual tendrá que modular la tonalidad de su voz.

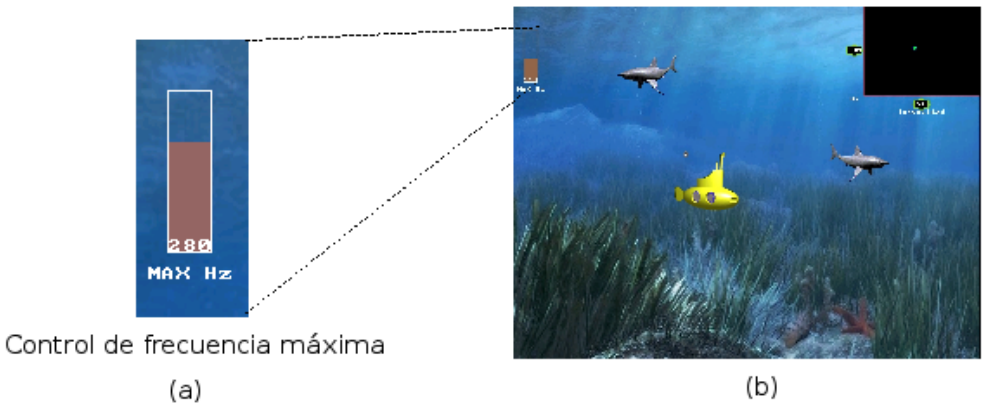


Figura 7.22: *Control de Frecuencia Máxima (a) y Actividad Submarino (b).*

La Figura 7.22(a) muestra el control que ajusta el límite máximo de operación, el cual representa la frecuencia de pitch máxima o altura máxima que alcanzará el personaje en el borde superior de la pantalla. La parte (b) de la misma figura muestra la actividad *Submarino*, la cual requiere de un buen control del tono y de intensidad en conjunto, ya que la tonalidad afecta la profundidad del submarino y la intensidad le da velocidad horizontalmente, la actividad consiste en controlar el submarino evadiendo los tiburones que aparecen aleatoriamente en pantalla. Las actividades de este nivel muestran en la parte superior derecha de la pantalla, un espacio con el trazado de pitch en función del tiempo para apreciar los patrones de entonación realizados por el usuario.

7.1.7 VOCALIZACIÓN

Figura 7.23: *Nivel 5 - VOCALIZACIÓN.*

La punta de la pirámide corresponde al nivel 5 de VOCALIZACIÓN como muestra la Figura 7.23. La actividad se denomina *Vocales* y se apoya también en la actividad ARTICULA la cual se describe en la Sección 7.2. *Vocales* se basa en el triángulo vocálico de la lengua española formado por los dos primeros formantes F1 y F2 que caracterizan dichas vocales, estos formantes dependen de la configuración geométrica del tracto vocal de cada niño y esta geometría cambia a su vez debido a varios factores como el sexo, edad, talla y raza entre otros. En esta actividad se aplicó la estimación robusta de formantes

$\tilde{F}_k$ del Capítulo 5, y la normalización de formantes F_{kN} propuesta en el Capítulo 6, con el objetivo de eliminar la influencia de la alta tonalidad presente en la voz infantil y reducir la variabilidad inter locutor.

La actividad *Vocales* se muestra en la Figura 7.24, allí puede apreciarse el panel de configuración en la parte superior, y en la parte derecha una región para la selección de vocales con una barra de acierto y otra de fallos, y la visualización del espectro de la señal de voz donde se resaltan los formantes estimados. En la parte inferior del espectro se ubican dos controles para iniciar o detener el barrido que se hace en tiempo real sobre la señal de entrada, facilitando la observación de la información gráfica ofrecida por el espectro. En la parte central, hay un tablero con una *diana* que cambia de lugar según la vocal seleccionada (en el caso de la figura la vocal /E/), y un cuadro con el puntaje de acierto y fallos. Ante la presencia de voz, el sistema dibuja los formantes normalizados F_{kN} estimados en la región central, y si se corresponden con la vocal seleccionada aparecerán en colores dentro de la región teórica para dicha vocal y el contador de aciertos se incrementa, en caso contrario, los formantes se dibujan en color gris y se incrementa el contador de fallos. Las centros teóricos pre-establecidos para las vocales, corresponden a la media obtenida en el corpus de voz infantil del Capítulo 4 clasificados por sexo y talla.

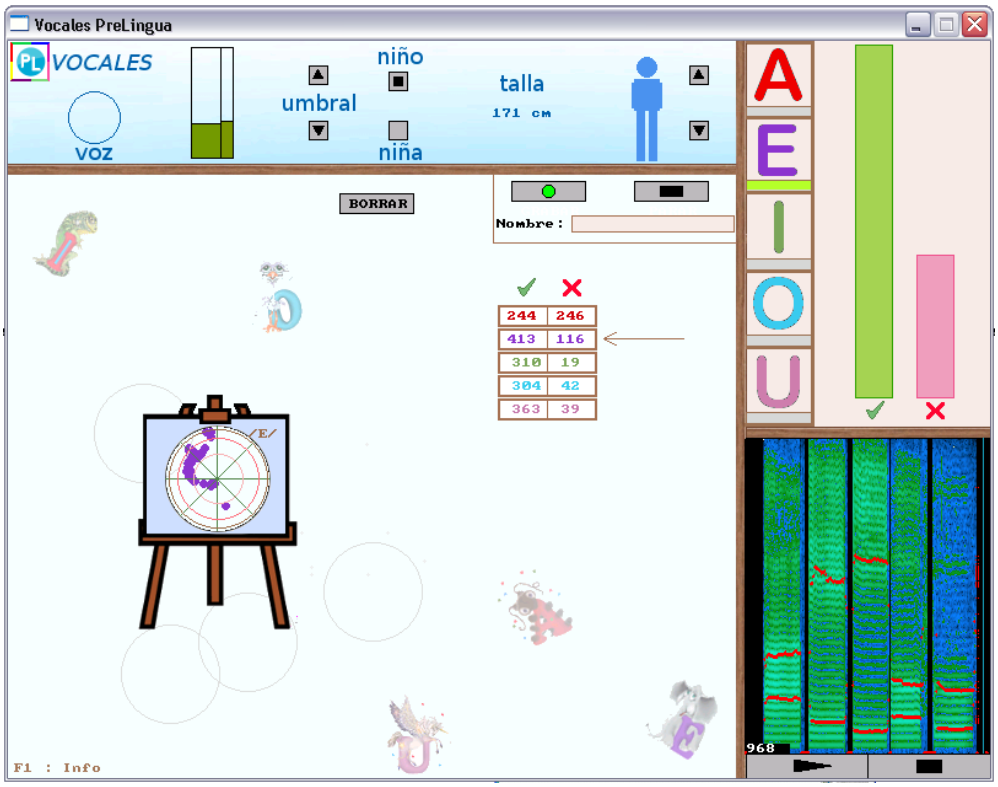


Figura 7.24: *Actividad Vocales.*

El panel superior mostrado en la Figura 7.25 permite configurar *Vocales* para su utilización, en la parte A se establece el umbral de detección de voz, en B se selecciona el sexo del niño y en C la talla aproximada. La selección de estos parámetros ubica las *dianas* en la posición media de los formantes para niños sanos con características semejantes, la

idea es que la emisión vocálica del niño en terapia se aproxime a la región mostrada sin la obligación de acertar en el centro, ya que es una tarea difícil y se trata de una aproximación estadística que busca mejorar la capacidad articulatoria del niño.

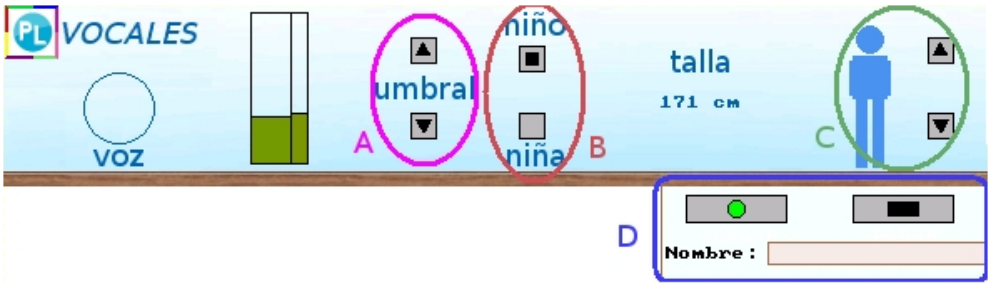


Figura 7.25: *Configuración de usuario.*

Esta actividad permite también la posibilidad de generar un archivo de texto y una imagen con información de la sesión, para ello, en D se puede introducir un nombre y presionar el botón de grabación al inicio y el de parar al final de la sesión. El reporte generado permitirá al terapeuta tener un registro del número de aciertos y fallos en cada sesión.

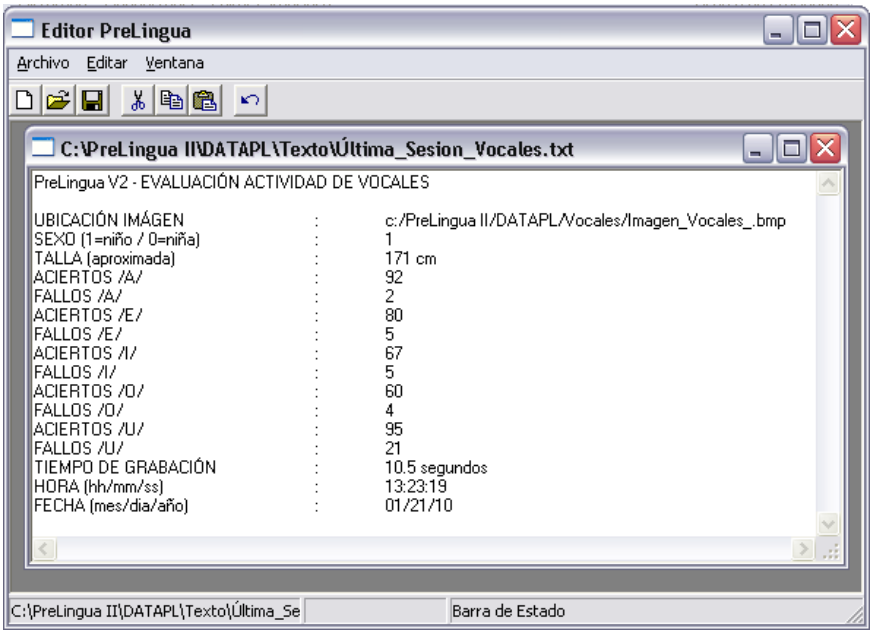


Figura 7.26: *Reporte Estadístico de Vocales.*

Dicho reporte se muestra en la Figura 7.26 el cual ofrece información general de la sesión como: Ubicación de la imagen, Sexo, Talla, puntuación otorgada por el sistema para Aciertos y Fallos en cada vocal, Tiempo de grabación, Hora y Fecha. La imagen generada es idéntica a la de la Figura 7.24 y es almacenada en formato .bmp, esta imagen junto al reporte son el soporte de la sesión para el terapeuta y ayudan a complementar la historia clínica.

7.1.8 Sección de Evaluación

Figura 7.27: *Sección EVALUAR.*

Esta sección permite evaluar la Intensidad, el Soplo y el Tono correspondientes a los niveles 2, 3 y 4 respectivamente de la pirámide como se muestra en la Figura 7.27. La evaluación se realiza midiendo el error generado entre los patrones definidos por el terapeuta para cada nivel y los patrones generados por el niño en sus respectivas sesiones al tratar de seguirlos, los datos generados se guardan en archivos de texto e imagen para ser utilizados posteriormente.

7.1.8.1 Evaluar Intensidad

Esta actividad permite trabajar y evaluar la intensidad de la voz de acuerdo a las necesidades específicas de cada usuario, el terapeuta tiene la libertad de configurar un patrón de intensidad: plano, ascendente, descendente, combinado, y que el niño deberá seguir para completar la actividad, en este caso la actividad usa un colibrí que debe llegar al nido. La Figura 7.28(a) muestra la actividad antes de ser utilizada y después de hacerlo en la parte (b). La trayectoria tiene siete puntos de control que permiten modificar el patrón con el puntero del ratón en cualquier dirección, y el nido también puede ser trasladado de la misma forma, el umbral de voz también puede modificarse y en la parte izquierda se muestra en todo momento la medición de la intensidad de entrada del micrófono y su valor en decibelios.

La aplicación puede utilizarse para practicar indefinidamente para que el niño comprenda lo que debe hacer: llevar el colibrí al nido modulando la intensidad de su voz. Una vez iniciada la actividad, a medida que el colibrí avanza va dejando tras de sí dibujada su trayectoria y se detiene al llegar al nido de forma automática, ó en cualquier momento presionando el botón PARAR, es entonces cuando el sistema calcula el error cuadrático medio entre la trayectoria definida previamente y la dejada por el colibrí. La parte (b) de

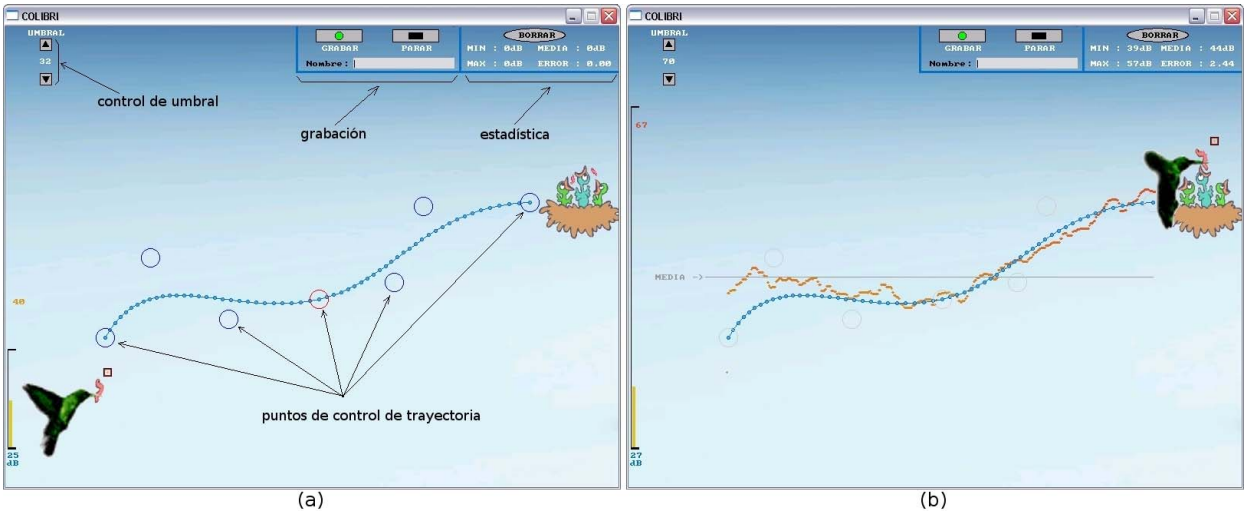


Figura 7.28: *Evaluación de INTENSIDAD.*

Figura 7.28 muestra el resultado de una grabación donde aparece una línea indicando la Media y los cálculos realizados como la Intensidad mínima, Intensidad máxima, Media de intensidad y el Error Cuadrático Medio en la esquina superior derecha de la pantalla.

Una vez realizada la grabación, el sistema genera un reporte en un archivo de texto y una imagen con la información de la sesión, el archivo puede ser abierto en el Editor de Estadística ubicado en el menú de la pantalla principal de *PreLingua* o en cualquier editor de texto. La Figura 7.31-(a) muestra dicho reporte con la siguiente información: Ubicación de la imagen, Intensidad mínima, Intensidad máxima, Rango dinámico, Media de Intensidad, Error Cuadrático Medio, Tiempo de grabación, Hora y Fecha. El Rango dinámico es un valor comúnmente utilizado por los terapeutas y es la diferencia entre el valor máximo y el mínimo registrados durante toda la sesión.

7.1.8.2 Evaluar Soplo

Con una filosofía similar a la actividad de evaluación de intensidad, también es posible trabajar y evaluar el soplo estableciendo diferentes patrones acordes a las necesidades del usuario. En este caso, la actividad consiste en llevar un globo que inicialmente esta suspendido en el aire y empieza a caer, hasta la base del otro lado de la pantalla por medio de la modulación del soplo y evitando que este caiga.

La Figura 7.29 muestra la actividad antes de ser utilizada en (a) y después de hacerlo en (b). La configuración del patrón se realiza desplazando los puntos de control con el puntero del ratón y la trayectoria final quedará descrita por la cadena de círculos amarillos, la plataforma de llegada también puede ser desplazada tanto horizontal como verticalmente con el puntero del ratón. Se busca que el niño sople y logre modular esta acción para mantener el globo en el aire y cerca de la trayectoria planteada, si el sistema detecta que el usuario emite sonidos sonoros, el globo deja de avanzar y empieza a caer lo que motiva al niño a volver a soplar como respuesta natural. Al completar la tarea, es decir cuando el globo llega a la plataforma, el sistema recompensa al usuario con eventos visuales y

sonoros. Cuando se graba la sesión, a medida que el globo avanza va dejando tras de sí

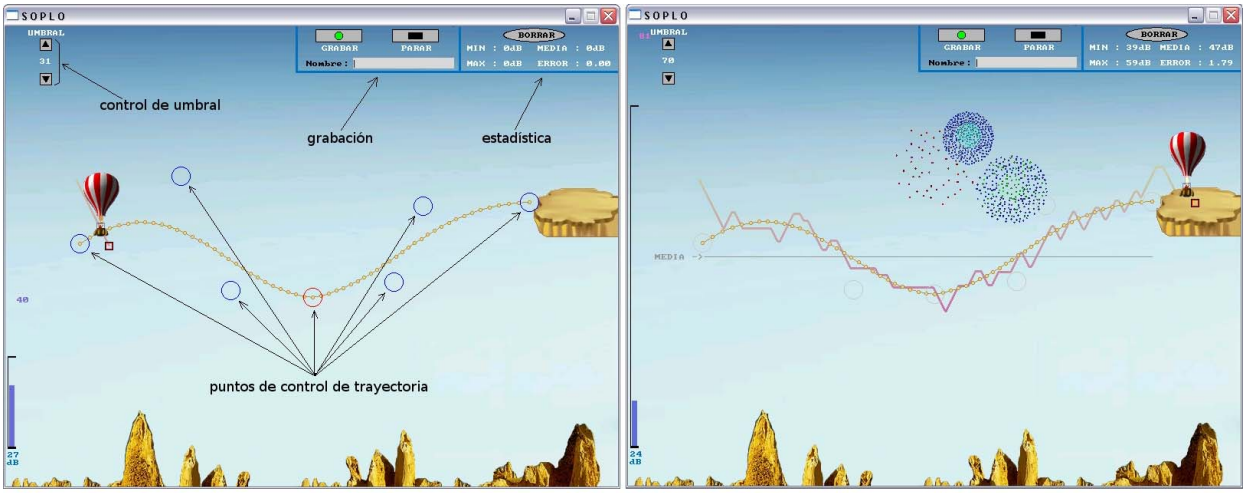


Figura 7.29: *Evaluación de SOPLO.*

dibujada su trayectoria hasta llegar a la plataforma y entonces se detiene la grabación, el sistema calcula el error cuadrático medio entre la trayectoria definida previamente y la dejada por el globo. Es entonces cuando en pantalla aparece una línea situada en el valor Medio y en la parte superior derecha los cálculos realizados, mostrando: la Intensidad mínima, Intensidad máxima, Media de intensidad y Error Cuadrático Medio. El reporte de texto generado (Figura 7.31-(b)) muestra información de la sesión como: Ubicación de la imagen, Intensidad mínima, Intensidad máxima, Rango dinámico, Media de Intensidad, Error cuadrático medio, Tiempo de grabación, Hora y Fecha.

7.1.8.3 Evaluar Tono

Figura 7.30: *Evaluación de TONO.*

Finalmente, esta sección permite evaluar la modulación del tono utilizando la misma filosofía de trabajo, es decir, midiendo el error entre la trayectoria definida por parte del

terapeuta y la trayectoria descrita por el usuario. La actividad de evaluación del tono utiliza un buzo que debe llegar a su submarino y solo es posible hacerlo modulando el tono de la voz. En esta caso las escalas de trabajo son los Hercios (Hz) y la configuración de la trayectoria y posición del submarino se configuran igualmente con el puntero del ratón. La Figura 7.30 muestra la actividad antes de ser utilizada en (a) y la misma aplicación después de ser utilizada en (b). En esta actividad el umbral de frecuencia máxima puede modificarse con el control respectivo en la esquina superior izquierda, y en la esquina inferior derecha se configura el umbral de voz.

Cuando inicia la grabación, a medida que el buzo avanza va dejando tras de sí dibujada su trayectoria en función de la frecuencia de pitch detectada, cuando el buzo llega al submarino o la grabación se detiene, el sistema calcula el error cuadrático medio entre la trayectoria definida previamente y la dejada por el buzo. Es entonces cuando en pantalla aparecen los cálculos realizados como: la Frecuencia mínima, Frecuencia máxima, Media de frecuencia y Error Cuadrático Medio.

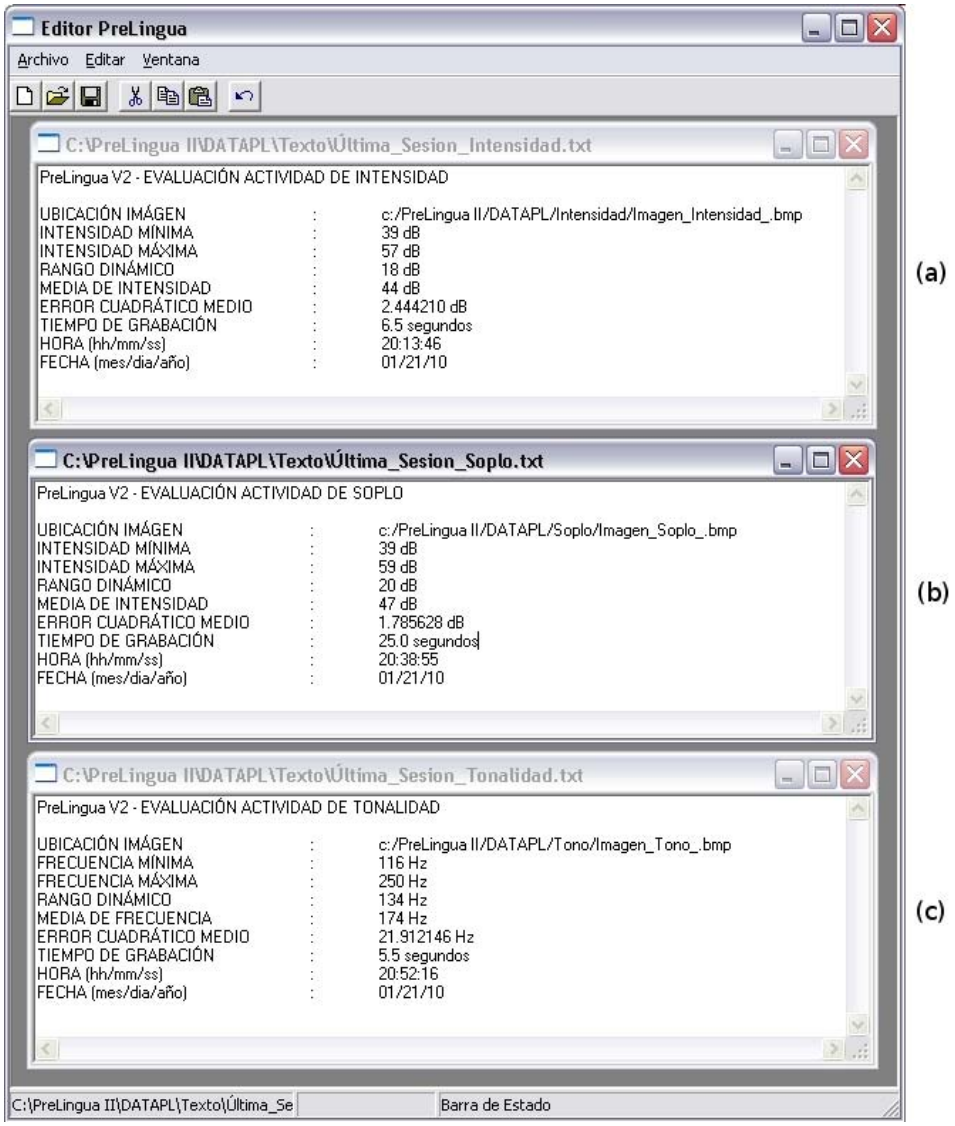


Figura 7.31: *Reportes Estadísticos de: Intensidad (a), Soplo (b) y Tono (c).*

El reporte de texto generado por esta actividad es mostrado en la Figura 7.31-(c), donde se muestra información de: Ubicación de la imagen, Frecuencia mínima, Frecuencia máxima, Rango dinámico, Media de Frecuencia, Error Cuadrático Medio, Tiempo de grabación, Hora y Fecha.

7.2 ARTICULA

Una de las necesidades más sentidas entre los logopedas a la hora de trabajar sonidos vocálicos en niños con discapacidad, es lograr que el niño entienda y visualice de manera gráfica como deben estar sus órganos articulatorios al momento de generar sonidos vocálicos. El problema comienza por la ausencia de herramientas que trabajen articulación vocálica en español y en tiempo real, además, las existentes que son de pago vienen para habla inglesa y muestran una información no útil para el niño a manera de curvas y gráficos técnicos como se trato en la Sección 2.5.

Como fruto de esta investigación se desarrolló una herramienta totalmente didáctica para los niños, que muestra en tiempo real una aproximación de la posición de los órganos articulatorios al momento de generar sonidos vocálicos. La herramienta permite la comparación entre la pronunciación vocálica del niño, y un patrón teórico para dicha vocal por medio de un avatar, de esta manera el niño podrá asimilar de una manera más gráfica y natural, el proceso de articulación vocálica en español.

Figura 7.32: *Nivel 5 - ARTICULA.*

Esta herramienta apoya el nivel 5 de VOCALIZACIÓN de *PreLingua* y se accede a ella haciendo click en icono mostrado en la Figura 7.32. Con esta aplicación podemos observar en tiempo real la posición aproximada de la lengua, labios y mandíbula inferior durante la articulación vocálica, como lo muestra la Figura 7.33. Allí se observa un avatar masculino en este caso compuesto por un cráneo como parte estática, y lengua, mandíbula y labios como partes dinámicas cuyo funcionamiento se describe en la Sección 7.2.1, para complementar el modelo, en él se visualizan también las cuerdas vocales las cuales oscilan en función de la frecuencia de pitch.

En la región **1** de la figura se establece el umbral de detección de voz, en **2** se configura el sexo, talla y esta el campo para introducir un nombre si se quiere al grabar la sesión, LTV muestra una aproximación en centímetros de la longitud del tracto vocal según la talla y sexo establecidos. En **3** se visualiza la señal de voz con la evolución de la intensidad en dB, en **4** se encuentra la evolución de pitch en Hercios (Hz), en **5** se observa la evolución de los formantes estimados $\tilde{F}_1$ y $\tilde{F}_2$ Hz y en **6**, el espectro de la señal de voz con los mismos formantes resaltados en color rojo. El cuadro de la región **7** muestra la estimación del error cuadrático medio existente entre el modelo y la pronunciación realizada por el niño.

Los campos **3**, **4**, **5** y **6** se encuentran alineados en tiempo lo que resulta muy útil en el análisis simultaneo de parámetros de la voz, también es posible detener el barrido en tiempo presionando el respectivo control en la esquina inferior derecha, así como su activación en la misma zona. Otra utilidad de esta herramienta es la posibilidad de conocer el valor numérico de cualquier parámetro (intensidad, pitch o formantes) simplemente pasando el puntero del ratón por estos campos.

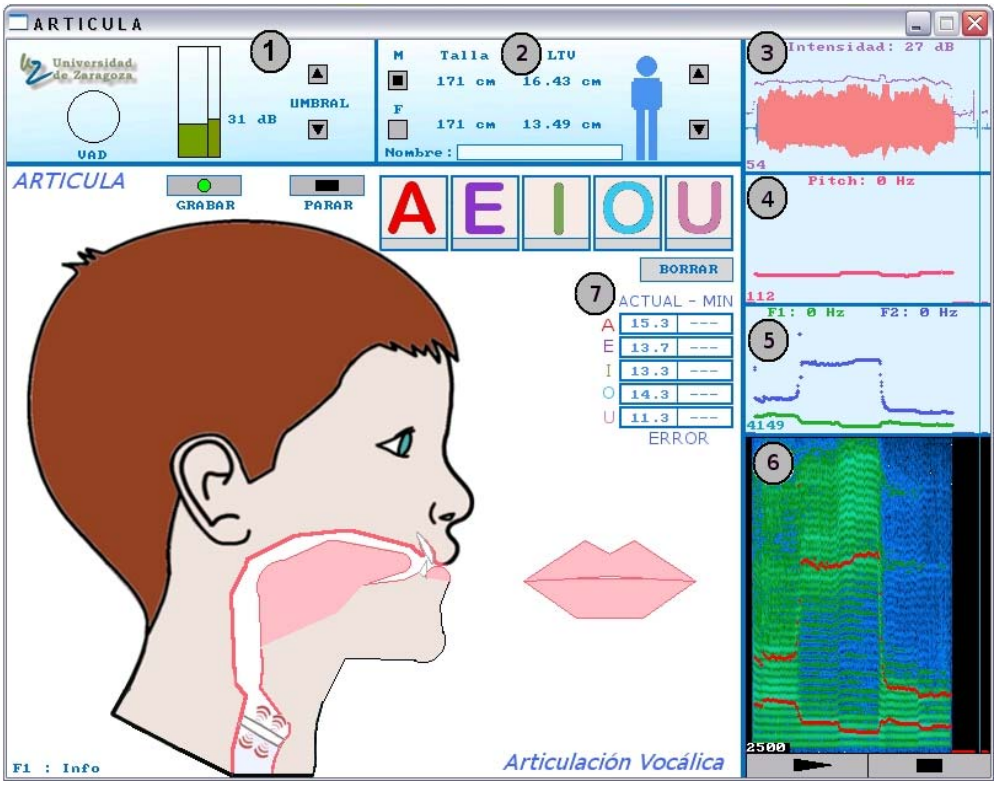


Figura 7.33: *ARTICULA. 1-Umbral de voz, 2-Selección de género y talla, 3-Señal de voz y trazado de intensidad, 4-Evolución de pitch, 5-Formantes $\tilde{F}_1$ y $\tilde{F}_2$, 6-Espectro de voz y formantes, 7-Tabla de errores calculados.*

Para utilizar la herramienta se selecciona primero sexo y talla para ajustar la normalización, la cual se realiza interpolando sobre los valores de las Figuras 6.6 y 6.7, a continuación con el puntero del ratón se selecciona la vocal deseada y aparecerá en pantalla un patrón lineal en color azul que indica la forma y posición aproximadas que toma la lengua en dicha vocal, como lo muestra la Figura 7.36 para la vocal /e/. Luego se motiva al niño a pronunciar dicha vocal de manera sostenida e intentar imitar el patrón azul mostrado en pantalla, es cuando el sistema utiliza los formantes estimados para mover en tiempo real de la lengua, mandíbula, labios y cuerdas vocales del avatar.

7.2.1 Diseño Interno

Es bien sabido que los sonidos vocálicos están determinados principalmente por la posición de la lengua, el grado de constricción de la luz vocal, y la forma de los labios. Desde un punto de vista acústico, las vocales pueden ser identificadas por sus dos primeros formantes

$F1$ y $F2$, y por fortuna, las vocales del español están relativamente separadas en el triángulo vocálico, situación que ayuda a su diferenciación. La Figura 7.34 muestra la posición de la lengua para las cinco vocales del español, y la representación del triángulo vocálico que permite explicar dichas posiciones, es decir un triángulo en el que se cambian los ejes respecto a la manera tradicional para ubicarlo en la cavidad bucal.

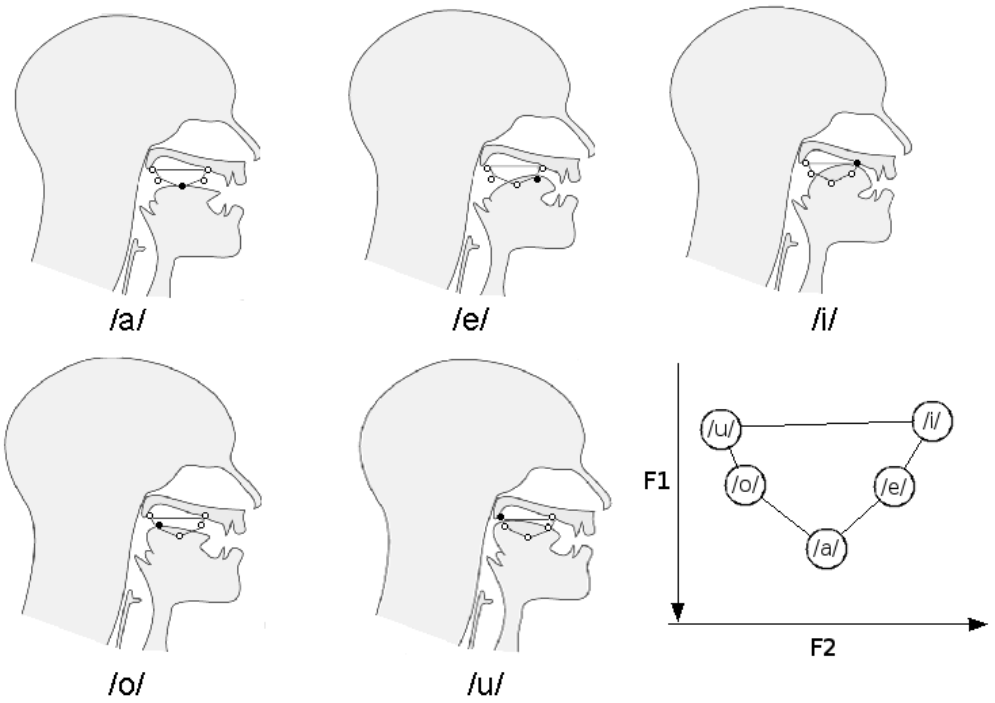


Figura 7.34: *Posición de la lengua en la producción vocálica.*

Esta representación indica que $F1$ esta correlado con la altura de la lengua dentro de la cavidad bucal, mientras que $F2$ esta correlada con la posición horizontal de la misma [Watt and Fabricius, 2002]. Estas son las premisas en las que se basa el diseño de ARTICULA para conseguir que la interfaz final fuera lo más natural posible. Aprovechando los avances logrados en la estimación robusta de formantes del Capítulo 5 y su normalización como en el Capítulo 6, ARTICULA utiliza los primeros dos formantes normalizados F_{1N} y F_{2N} para animar el avatar de un niño o niña según el caso, el avatar ha sido desarrollado con un cráneo como parte estática y tres partes dinámicas compuestas por la lengua, la mandíbula y los labios. Los formantes normalizados F_{kN} modifican la posición horizontal y vertical de los componentes dinámicos pero en diferentes proporciones para cada uno.

La Figura 7.35 muestra los tres componentes de manera independiente con sus respectivos grados de libertad representados por las flechas de doble punta, la lengua tienen dos grados de libertad mientras que la mandíbula inferior tiene solo uno y sus coordenadas cartesianas están descritas por las expresiones:

$$lengua(x_l + \alpha F_{2N}, y_l + \beta F_{1N}) \tag{7.1}$$

$$mand\acute{i}bula(x_m, y_m + \gamma F_{1N}) \tag{7.2}$$

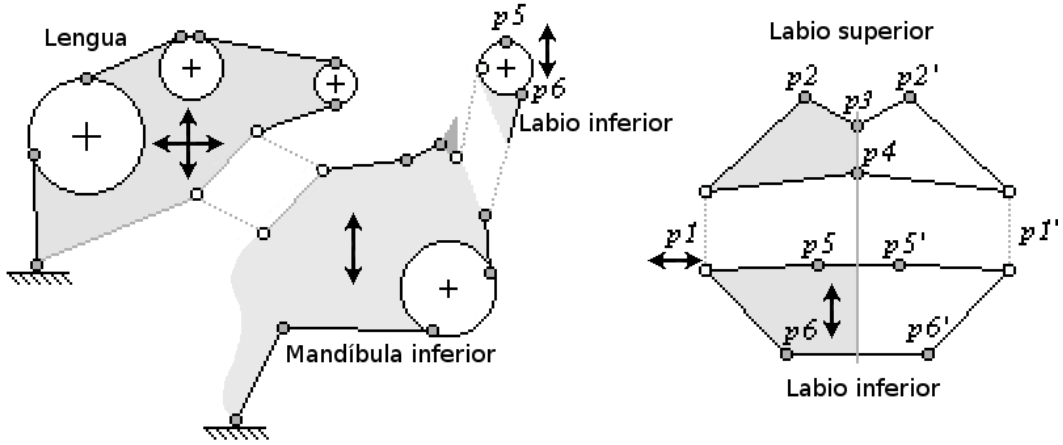


Figura 7.35: *Componentes dinámicos: Lengua, Mandíbula inferior y Labios.*

donde x_l, y_l y x_m, y_m son las coordenadas en píxeles de la posición en pantalla de la lengua y la mandíbula respectivamente en estado de reposo, y α, β y γ son los factores de escala para convertir las unidades de Hercios (Hz) en píxeles. Experimentalmente estos valores se han establecido en $\alpha = 0.022$, $\beta = 0.063$ y $\gamma = 0.03$.

La parte derecha de Figura 7.35, muestra los labios modelados con dos grados de libertad: uno en la dirección horizontal representado por $p1$ y localizado en la comisura de los labios, y otro en el labio inferior que afecta a los puntos $p5$ y $p6$. Los puntos con la notación px' significa que tienen el mismo comportamiento que los puntos px pero en la otra mitad de la boca.

El comportamiento de los puntos $p1, ..., p6$ están gobernados por las expresiones:

$$p1 = (x_1 + \Delta x, y_1) \tag{7.3}$$

$$p2 = (x_2, y_2) \tag{7.4}$$

$$p3 = (x_3, y_3) \tag{7.5}$$

$$p4 = (x_4, y_4) \tag{7.6}$$

$$p5 = (x_5, y_5 + \Delta y) \tag{7.7}$$

$$p6 = (x_6, y_6 + \Delta y) \tag{7.8}$$

donde: x_i, y_i con $i = 1, ..., 6$ son las coordenadas de la posición en pantalla en píxeles para cada punto, y Δx y Δy son los factores que mueven los labios propiamente y están definidos por:

$$\Delta x = k_1 \delta \tag{7.9}$$

$$\Delta y = 0.85 \gamma F_{1N} \tag{7.10}$$

donde δ, es una distancia obtenida directamente de los formantes por la expresión:

$$\delta = \sqrt{F_{1N}^2 + F_{2N}^2} \tag{7.11}$$

la distancia δ provee una proporción de la distancia física entre los ángulos de la boca, es decir, las vocales cerradas como la /o/ y la /u/ tienen valores de formantes bajos y por ende δ es bajo, mientras que vocales abiertas como la /a/ y la /e/ tienen valores más altos de formantes y por lo tanto δ será mayor. k_1 es el factor de escala para ajustar los valores de la distancia δ a las coordenadas en pantalla en píxeles, en este caso $k_1 = 0.016$, y finalmente, Δy es la componente vertical del labio inferior cuyo valor es una proporción de la componente vertical de la mandíbula.

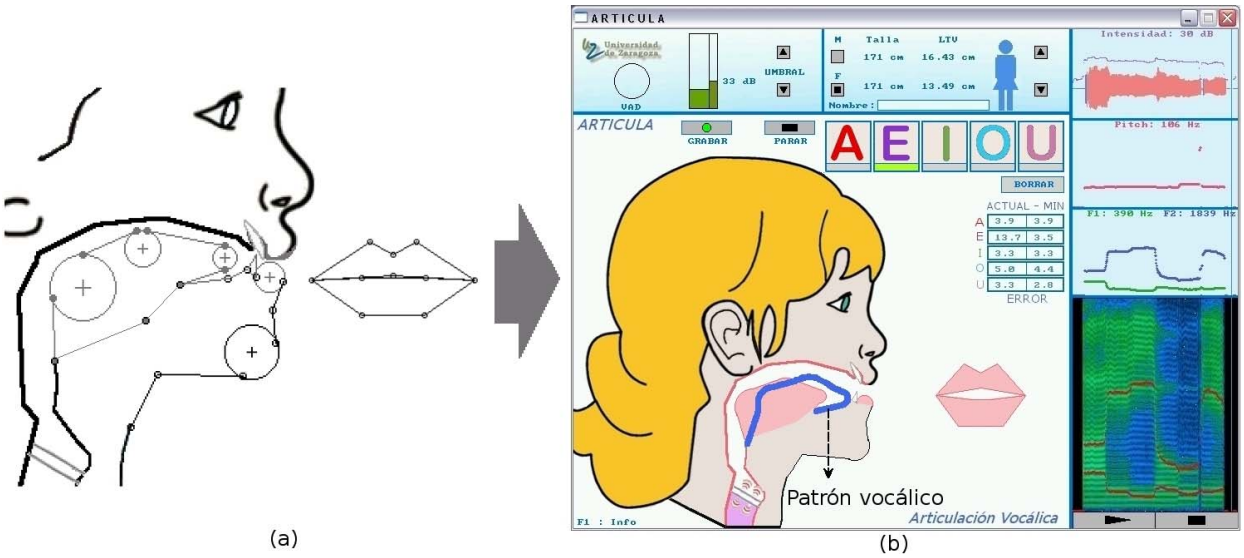


Figura 7.36: *Unión de componentes estático y dinámicos en el avatar (a) y Aplicación final de usuario (b).*

Integrando estos componentes dinámicos con la parte estática, se completa el modelo del avatar como muestra la Figura 7.36(a), y la parte (b) muestra la aplicación final con el avatar femenino en este caso. Para utilizar la herramienta, el terapeuta debe seleccionar la talla y el sexo del niño para establecer la normalización que aplicará el sistema, seguidamente se selecciona la vocal a trabajar y es cuando aparece en pantalla el patrón teórico de dicha vocal representado por una silueta de la lengua en color azul. La silueta del patrón mostrado y el de la lengua del avatar han sido creados a partir de figuras geométricas simples como círculos y líneas, y basados en imágenes de resonancia magnética (MRI) de estudios como en: [J. Gurlekian and Eleta, 2000].

7.2.2 Evaluación de la Articulación Vocálica

ARTICULA también ofrece la posibilidad de evaluar al usuario en cada sesión entregando los resultados en un reporte estadístico e información gráfica. En este caso para medir la habilidad del usuario en la articulación vocálica, se mide el error cuadrático medio mínimo obtenido entre el patrón vocálico del sistema (silueta azul), y la silueta de la lengua del avatar controlada por la emisión vocálica del usuario.

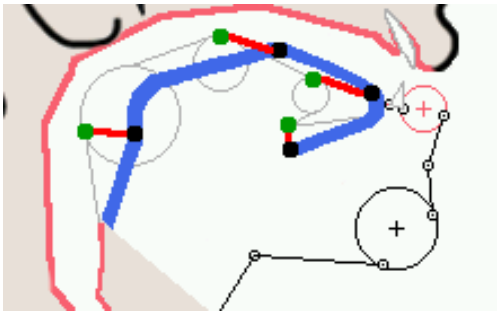


Figura 7.37: *Error entre patrones vocálicos.*

La Figura 7.37 muestra de donde se obtiene el error cuadrático medio (líneas rojas) entre cuatro puntos de referencia del patrón teórico (puntos negros), y cuatro puntos de referencia en la lengua del avatar (puntos verdes). En general la actividad se trata de una aproximación, ya que los formantes se ven afectados por muchos factores inherentes a cada usuario y más aun, si el usuario padece de alguna malformación física en su tracto vocal. De manera que el objetivo no es que los patrones coincidan perfectamente sino acercarse al máximo para ganar control en la articulación. El reporte estadístico generado es mostrado

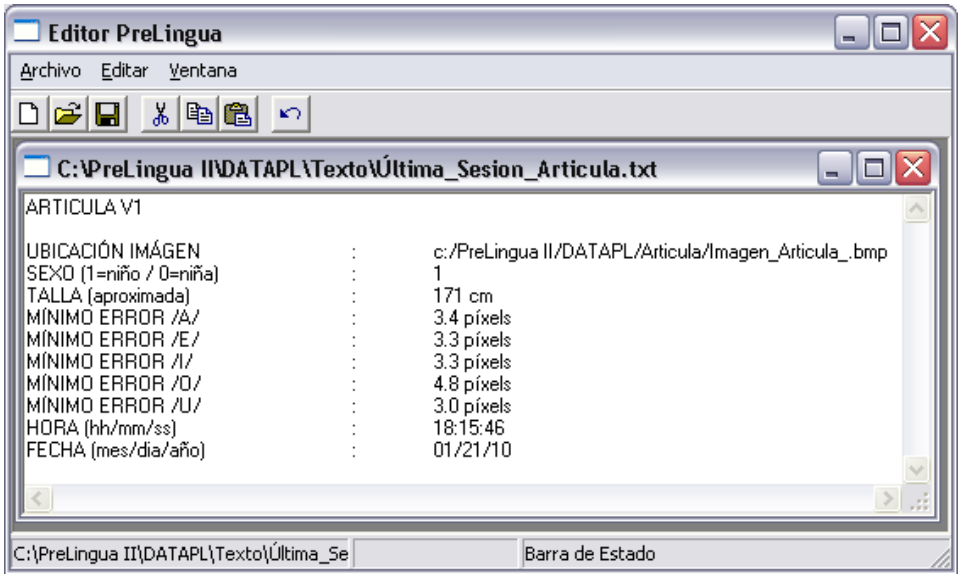


Figura 7.38: *Reporte estadístico de ARTICULA.*

en la Figura 7.38, el cual brinda información de la sesión como: Ubicación de la imagen, Sexo, Talla, Errores Cuadrático Medio mínimos para cada vocal durante la sesión, Hora y Fecha.

7.3 ViVo

El estudio del triángulo vocálico del español esta documentado en su mayoría para adultos y de manera muy escasa para niños, limitándose la información existente a tablas y datos muy generales. Derivado del trabajo de normalización de formantes en función de las características de cada individuo, se diseñó una herramienta de carácter más académico

que permita la visualización en tiempo real de los parámetros más relevantes de la voz, sin que la variabilidad del usuario afecte esta apreciación.

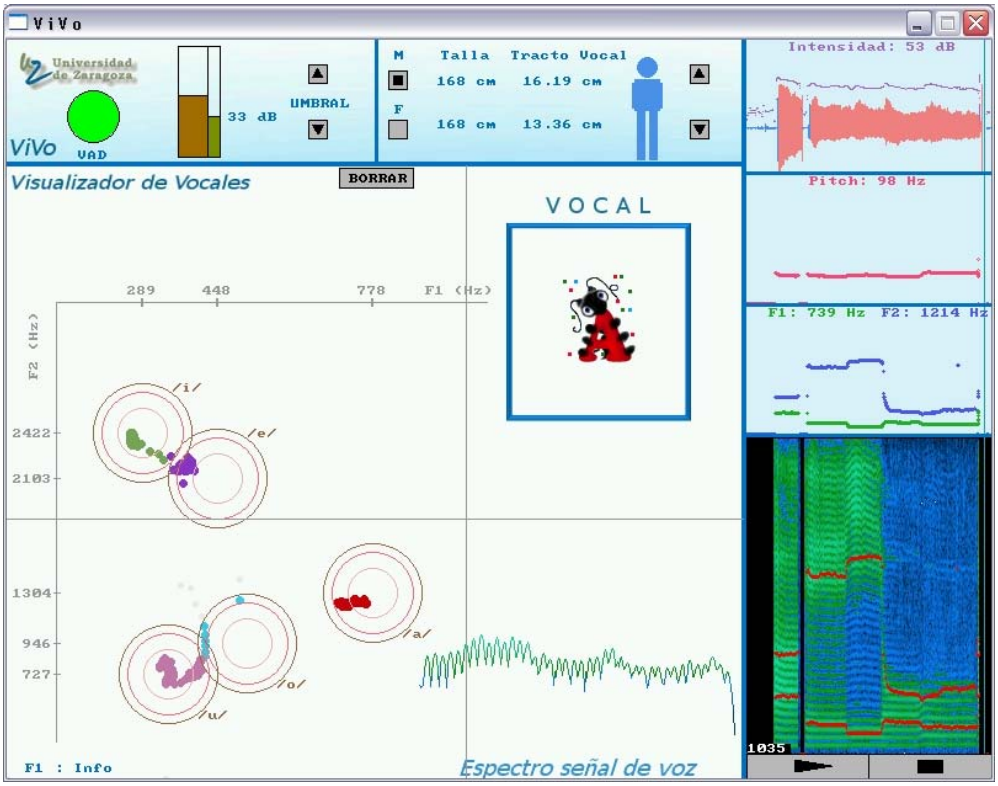


Figura 7.39: *Visualizador de Vocales ViVo.*

La aplicación denominada Visualizador de Vocales ViVo se muestra en la Figura 7.39, esta herramienta esta basada en la actividad de *Vocales* del nivel 5 de *PreLingua* con algunas modificaciones, ViVo muestra el triángulo vocálico completo, y el espectro de la señal de voz obtenido de la transformada de Fourier localizada, complementando la información acústica ya mostrada, también hace identificación vocálica según los formantes detectados (/a/ en el caso de la figura).

La herramienta resulta de gran utilidad para estudios mismos de voz, fonética y lingüística, igualmente en la práctica clínica y en la academia es usual el contraste de información de la voz en el mismo instante de tiempo, ojalá sin necesidad de utilizar varias herramientas simultáneamente.

7.4 VocalCLICK

Las herramientas como *PreLingua* y ARTICULA buscan ayudar a personas con alteraciones en su voz, pero existen personas cuya discapacidad no altera su voz sino que por el contrario presentan serias limitaciones físicas. La anterior situación puede generar la exclusión de estas personas en el acceso a las nuevas tecnologías y el uso de ordenadores. Existen diversas

ayudas técnicas como: pulsadores, ratones o joystick especialmente adaptados que intentan disminuir esta exclusión [Sánchez, 2002], pero infortunadamente se repite la situación de que generalmente son de elevado costo de adquisición.

Con los avances obtenidos en esta tesis, se busca expandir el potencial de aplicación de las Tecnologías de Habla permitiendo que este tipo de población pueda acceder a ordenadores por medio de su voz, la herramienta presentada aquí ofrece una alternativa para que personas con discapacidad física accedan al control del puntero del ratón por medio de emisiones vocálicas. La herramienta desarrollada denominada $VocalCLICK$, permite emular los movimientos del ratón con sonidos vocálicos cuyos formantes derivan de la estimación robusta y normalización seguida en esta tesis. Partiendo de los triángulos vocálicos normalizados de la Figura 6.9 del Capítulo 6, se han definido cuatro regiones cuyos espacios corresponden a las cuatro direcciones del puntero del ratón: derecha, arriba, izquierda y abajo.

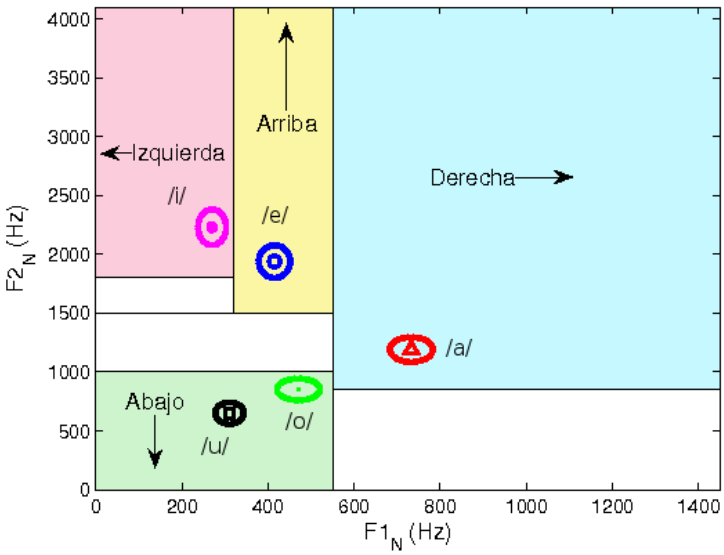


Figura 7.40: *División en regiones del triángulo vocálico.*

La Figura 7.40 muestra el triángulo vocálico y las cuatro regiones definidas para cada dirección, los límites establecidos responden a que las zonas entre ellos son las que mejor diferencian cuatro zonas principales, una para cada dirección. De manera que el usuario al pronunciar la vocal /a/, sus formantes normalizados por el sistema hacen que el puntero del ratón se desplace hacia la derecha, si pronuncia la /e/ el puntero irá hacia arriba y al pronunciar la vocal /i/ el puntero irá hacia la izquierda, finalmente, y debido a que las vocales cerradas /o/ y /u/ tienen formantes muy próximos, se estableció una sola zona conjunta con ellas para definir la dirección vertical.

La Figura 7.41 muestra la herramienta $VocalCLICK$. Se compone de una sección de configuración compuesta por: Usuario, Umbral y Ventana de VOZ, y una zona que visualiza los parámetros de la voz estimados durante su funcionamiento. La configuración de Usuario permite configurar la herramienta especificando características del usuario como sexo y talla para que el sistema aplique la normalización, y en la sección de Umbral, se establece el nivel

de energía a partir del cual la trama analizada se considera sonora.

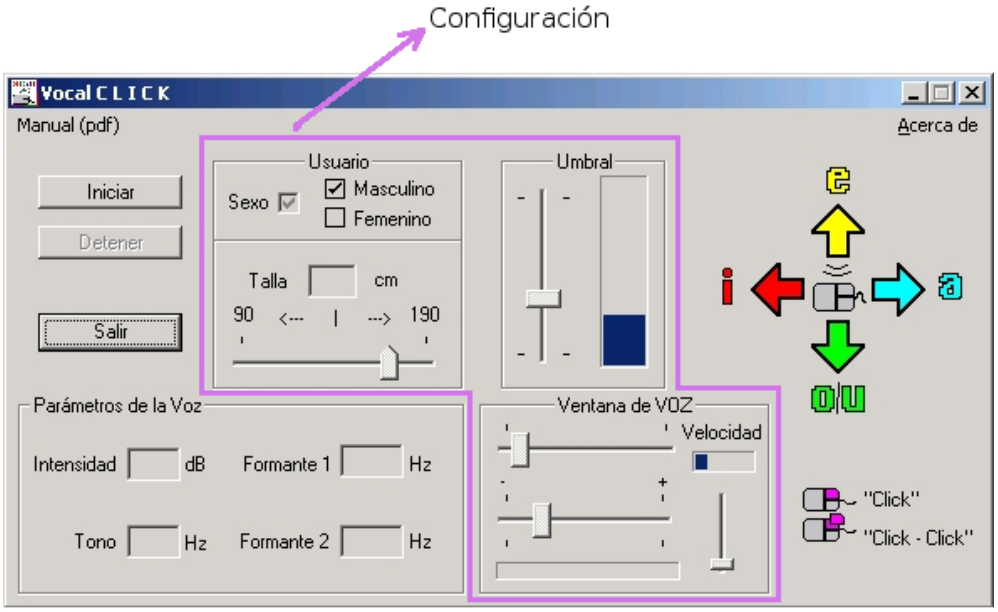


Figura 7.41: *VocalCLICK.*

En la sección de Ventana de VOZ, se configuran los eventos de click izquierdo del ratón, y la velocidad de desplazamiento del puntero. Como muestra la Figura 7.42, hay dos controles horizontales y un indicador de avance que muestra la acumulación de tramas sonoras, la zona comprendida entre el control superior y el inferior (en rojo) corresponde a la zona de click, lo que significa que si la acumulación de tramas sonoras llega a esta zona y se interrumpe, el sistema convierte este evento en un click izquierdo de ratón, si el usuario puede repetir dos veces este proceso en un intervalo corto de tiempo, el sistema lo interpreta como un doble click izquierdo. Si la emisión sonora llega hasta la zona de movimiento después del control inferior sin interrupción (en verde), el sistema desplazará el puntero del ratón según la dirección correspondiente a la vocal pronunciada.

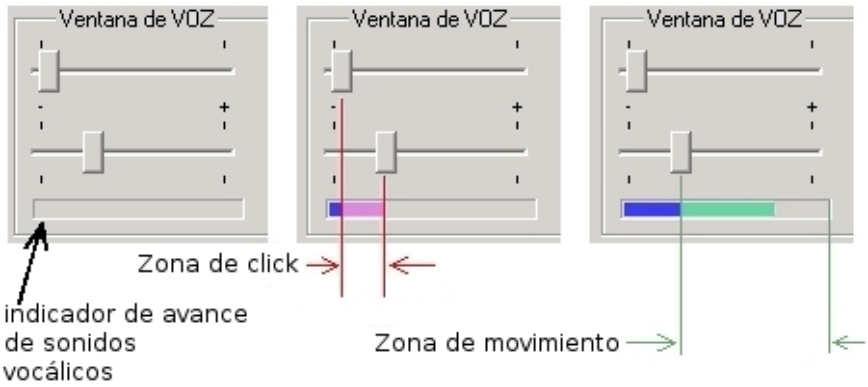


Figura 7.42: *Control Ventana de Voz.*

En la misma zona de Ventana de Voz existe un control vertical que establece la velocidad

de desplazamiento del puntero, sin embargo, la velocidad del puntero depende también de la intensidad de la señal de voz, lo que significa que a mayor volumen en la emisión vocálica el puntero se desplazará a mayor velocidad por la pantalla, y cuando la intensidad es baja la velocidad del puntero disminuirá. Para los eventos de click doble y sencillo, se puede utilizar cualquier emisión sonora como /*click*/ o /*ki*/, o cualquiera sonido sonoro que el usuario pueda producir con facilidad, el único requisito es que en tiempo la emisión sonora llegue hasta la zona de click previamente establecida y se interrumpa dentro de ella. En $VocalCLICK$, una vez establecidos los niveles de trabajo de la herramienta y las características del usuario, estos valores son guardados por el sistema para su posterior uso sin necesidad de repetir todo el proceso al iniciar la aplicación, aunque la configuración puede cambiarse en cualquier momento y el sistema guardará los últimos valores establecidos.

Este capítulo ha presentado el conjunto de herramientas desarrolladas para logopedia y educación especial derivadas del tema de investigación de la presente tesis. Las herramientas están disponibles de manera gratuita y en versiones de prueba en Internet, y se reciben continuamente experiencias de uso y observaciones de quienes las han usado, lo que ha permitido mejorarlas aumentando su funcionalidad y robustez. El siguiente capítulo presenta un estudio real realizado en dos colegios de educación especial, aplicando *PreLingua* con el objetivo de poner a prueba la herramienta y obtener resultados cualitativos y cuantitativos de dicha aplicación.

Capítulo 8

Aplicación en Reconocimiento Automático del Habla

Este capítulo presenta los resultados de aplicar la estimación de la longitud del tracto vocal descrita en el Capítulo 6, en la tarea de normalización del tracto vocal en RAH. Es bien sabido que la normalización del tracto vocal es un proceso utilizado con éxito hasta la fecha para mejorar las prestaciones de los sistemas de RAH con el objetivo de reducir la variabilidad inter-locutor. Las técnicas para Normalización de la Longitud del Tracto Vocal o Vocal Tract Length Normalization (VTLN), suelen requerir de varios pasos para estimar el mejor factor de deformación de los ejes de frecuencia para un locutor dado, ya sea por el método de máxima verosimilitud o Maximum Likelihood (ML), o por medio de la estimación de características acústicas del locutor como los formantes, lo que implica un elevado costo computacional y obliga a que el proceso sea Off-line.

Ya que la presente investigación plantea una manera robusta de estimar la longitud del tracto vocal, y éste es un parámetro que caracteriza muy bien al locutor, en ésta sección se propone utilizar la LTV estimada para obtener el factor de deformación de los ejes de frecuencia y lograr mejorar los resultados obtenidos en reconocimiento utilizando la base de datos TIDigits. El sistema propuesto utiliza una función de actualización del factor de deformación aprovechando la LTV estimada trama a trama, consiguiendo que el sistema funcione de manera On-line y cuyos resultados se muestran muy semejantes a las técnicas Off-line.

La Sección 8.1 presenta una breve introducción a las técnicas de VTLN, la Sección 8.2 muestra como se estima el factor de deformación y como éste se actualiza dependiendo de la longitud del tracto estimada. Finalmente, la Sección 8.3 describe el marco de experimentación y los resultados obtenidos.

8.1 Técnicas de VTLN en RAH

En un sistema de RAH, las grandes diferencias que pueden existir entre el conjunto de locutores utilizado para entrenar los modelos acústicos y el conjunto de locutores utilizados para reconocer, hace que los resultados del reconocimiento se degraden considerablemente. Una fuente bien conocida de éstas diferencias es la gran variación anatómica de los tractos

vocales entre locutores, lo que se traduce en una alta variabilidad espectral entre las señales de voz. Ésta situación se acentúa aun más si entre los locutores hay hombres o mujeres, adultos o niños, es decir, que un sistema de RAH entrenado con adultos puede tener un mal desempeño para reconocer locutores infantiles y viceversa.

Diferentes alternativas han surgido para reducir estas diferencias entre los datos de entrenamiento y los datos de reconocimiento. Algunas de ellas requieren del re-entrenamiento de los modelos acústicos, como en las técnicas de adaptación al locutor Máximo A Posteriori (MAP) [Gauvain and Lee, 1994] o la de Regresión Lineal de Máxima Verosimilitud (Maximum Likelihood Linear Regression (MLLR)) [Legetter and Woodland, 1995]; en tanto que otras actúan directamente sobre la señal de voz y no modifican los modelos. Por ejemplo, la técnica de VTLN es bien conocida por reducir estas diferencias sin modificar los modelos acústicos iniciales [Lee and Rose, 1998], [Gouvea and Stern, 1997], [Molau et al., 2000], ésta técnica considera que la principal diferencia entre dos locutores está en el cambio del eje frecuencial entre ellos debido a la diferencia entre las longitudes de sus tractos vocales.

Sin embargo, las técnicas de VTLN usualmente demandan un alto costo computacional para procesar los datos de la señal de voz, y encontrar previamente al reconocimiento final la mejor transformación del eje frecuencial de un locutor dado, a un eje frecuencial de un locutor objetivo, situación que dificulta la utilización de ésta técnica en aplicaciones en tiempo real. Así pues, la técnica de VTLN tiene por tarea proporcionar una función de deformación que transforme el eje frecuencial de un locutor dado f, a el eje frecuencial de un locutor objetivo f'. Diversas opciones han sido investigadas para obtener esta función de deformación, desde aproximaciones secuenciales lineales hasta funciones exponenciales. Todas ellas dependen del factor de deformación α como en la ecuación 8.1, la cual expande o contrae el espectro de la señal de voz según se desee [Lee and Rose, 1998].

$$S_{deformado}(f) = S_{no-deformado}(f'(\alpha, f)) \tag{8.1}$$

Un factor de deformación que contrae el eje frecuencial se utiliza para transformar locutores con un tracto vocal corto (como en los niños o mujeres), en locutores con un tracto vocal más largo (como en hombres), y un factor de deformación que expande éste eje frecuencial se utiliza para transformar un tracto vocal largo en un tracto vocal corto. Un ejemplo de una función de transformación exponencial se puede apreciar en la Figura 8.2. Una técnica más eficiente consiste en transformar y deformar la escala del banco de filtros Mel cuando se calculan los coeficientes cepstrales MFCC o Mel Frequency Cepstrum Coeficients, en lugar de deformar todas las tramas de voz a la entrada del reconocedor. Funciona de manera contraria en el sentido de que contraer la escala Mel equivale a expandir el espectro, y expandir la escala Mel equivale a contrael el espectro.

La alternativa derivada de ésta investigación y que aquí se propone, consiste en estimar el factor de deformación del eje frecuencial a partir de la estimación de la longitud del tracto vocal propuesta en el Capítulo 6, a la que se le ha adicionado una función de actualización que solo depende de la longitud estimada en la trama actual y en la anterior para conseguir que su funcionamiento sea en tiempo real.

8.2 Estimación y Actualización del Factor de Deformación α

La estimación del factor de deformación α suele ser la parte más delicada en las técnicas de VTLN. Un valor inadecuado puede reducir la potencial mejora ofrecida por la técnica o incluso reducir considerablemente el desempeño de todo el sistema. Dos fuertes tendencias para la estimación del factor α se encuentran en la literatura: por una parte, la basada en la Máxima verosimilitud o Maximum Likelihood (ML), la cual selecciona el factor de deformación que mejor verosimilitud obtenga, entre varias versiones de la señal de entrada deformadas por varios factores a un determinado modelo acústico [Lee and Rose, 1998]; por otro lado, técnicas basadas en características acústicas del locutor como los formantes o una combinación de ellos para estimar el factor de transformación, ya que las frecuencias de resonancia están correladas con el tracto vocal [Gouvea and Stern, 1997].

8.2.1 Técnicas Basadas en Modelos

Dentro de las técnicas basadas en modelos encontramos la basada en máxima verosimilitud ML-VTLN la cual ofrece muy buenos resultados pero trabaja de manera Off-line. La Figura 8.1(a) [Lee and Rose, 1998], muestra el diagrama en el que se basa ésta técnica, en donde el sistema hace una transcripción inicial de la frase pronunciada, con esta transcripción, un conjunto de n codificadores Viterbi aplica diferentes factores de deformación $\{\alpha_1 ... \alpha_n\}$, para decidir que factor α_i tiene mayor probabilidad de acuerdo a la puntuación obtenida por cada decodificador, finalmente, el factor de deformación seleccionado, es utilizado en una segunda etapa de reconocimiento haciendo uso del VTLN para mejorar la estimación de la frase reconocida y dar un resultado final optimo. La implementación que se usará en este trabajo utiliza 11 factores de deformación en la fase de decodificación Viterbi, desde 0.9 a 1.1 en intervalos de 0.02.

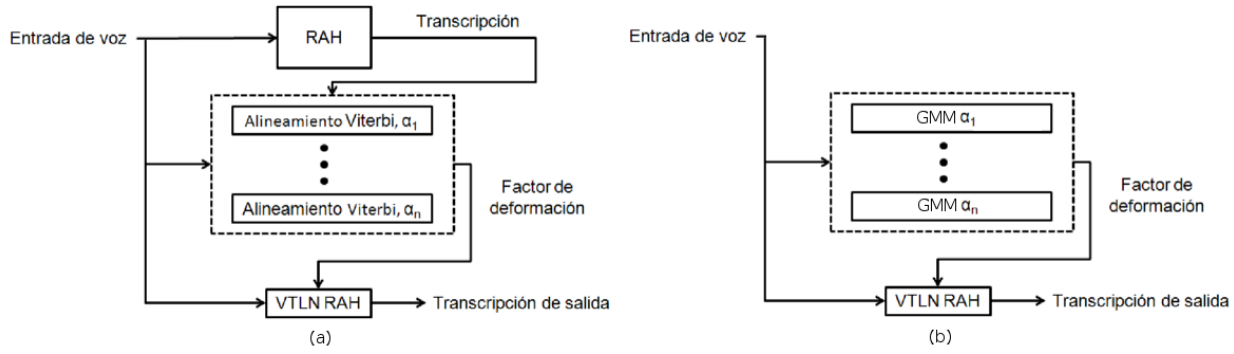


Figura 8.1: *Diagramas de las técnicas basadas en ML-VTLN y en ML-GMMs*

Otra técnica basada en modelos para estimar el mejor factor de deformación, es la basada en Modelos de Mezclas de Gaussianas (GMMs) como en [Molau et al., 2000], donde se calcula una GMM para cada factor de deformación con todos los datos de entrenamiento como se muestra en la Figura 8.1(b), por lo tanto, cada modelo ahora representa la distribución de un factor de deformación específico en el espacio de características y se

adicionan la varianza en todos los modelos. Durante el reconocimiento, los vectores acústicos sin normalizar se testan con todas las GMMs para encontrar el mejor factor de deformación según la probabilidad a posteriori.

8.2.2 Técnicas Basadas en Características

Considerando la técnica LTV presentada en ésta tesis en el Capítulo 6, el sistema tiene en cuenta la información formántica de cada locutor, pero para estimar la longitud del tracto vocal, y a partir de ésta estima el factor de deformación. En la técnica propuesta, el sistema hace una estimación de la longitud del tracto vocal siempre y cuando el segmento analizado sea sonoro, es decir que exista estimación formántica; mientras que en los segmentos no sonoros o de silencio el sistema entrega una salida vacía sin estimación numérica. Para todos los valores de longitud estimados para un locutor, se calcula la media de la longitud del tracto para este locutor (LTV_{loc}) y se obtiene el factor de deformación con la ecuación 8.2, donde $\overline{LTV}_{modelo}$ es la media del tracto vocal calculada para todos los locutores utilizados en la fase de entrenamiento del modelo acústico, lo que se hace de manera Off-line en una etapa previa. El factor λ se utiliza para moderar la cantidad de deformación aplicada y se estableció en $\lambda = 0.5$, después de algunas pruebas iniciales en bases de datos pequeñas.

$$\alpha = 1 + \lambda \frac{\overline{LTV}_{modelo} - LTV_{loc}}{\overline{LTV}_{modelo}} \tag{8.2}$$

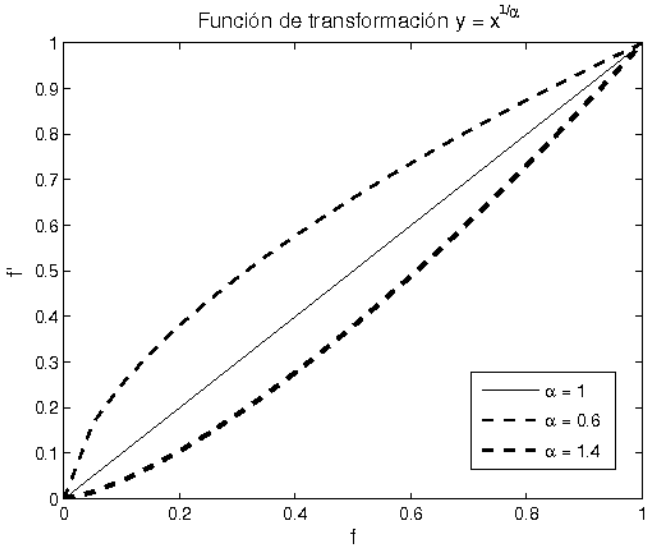


Figura 8.2: *Función de transformación exponencial.*

Este factor de deformación α es aplicado a la función de transformación exponencial $y = x^{\frac{1}{\alpha}}$ para hacer la transformación frecuencial sobre el banco de filtros como muestra la Figura 8.2, en donde se aprecia la transformación para $\alpha = 1$, $\alpha = 0.6$ y $\alpha = 1.4$. El diagrama de la técnica propuesta para estimar el factor de deformación α se aprecian en la Figura 8.3, en donde el sistema hace una transcripción de la frase pronunciada directamente sin pasos previos, trabajando directamente sobre la señal y pasándola por un solo modelo.

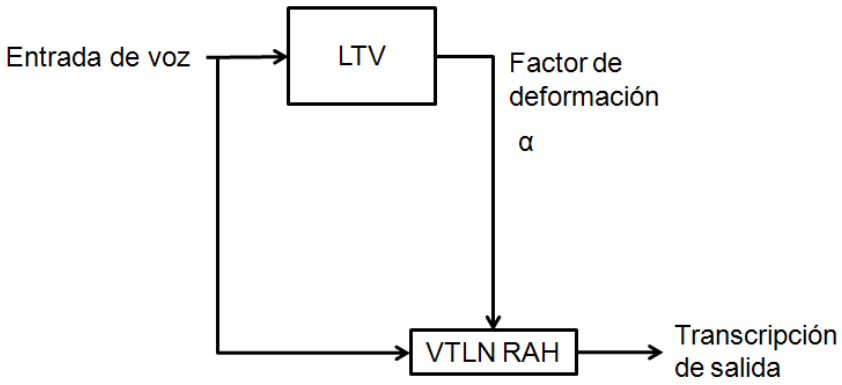


Figura 8.3: *Diagrama de la técnica LTV*

Con el objetivo de poder hacer la estimación en tiempo real, y aprovechando que la estimación de la longitud del tracto se hace trama a trama, se hace una re-estimación de la longitud como muestra la ecuación 8.3 añadiendo un factor β denominado factor de memoria del sistema. De manera que el valor de longitud estimado para una trama dada i, depende del valor estimado en la trama anterior $i-1$ y el valor estimado en la trama actual i. El factor de memoria utilizado para la experimentación fue de $\beta = 0.99$, con este factor de memoria se evitan las variaciones locales en las estimaciones de la longitud del tracto, con una tendencia hacia la media del locutor cuando se analizan suficientes tramas.

$$LTV_{loc}(i) = \beta * LTV_{loc}(i-1) + (1-\beta) * LTV(i) \tag{8.3}$$

De ésta manera, cuando un locutor accede por primera vez a un sistema de RAH, la longitud del tracto vocal es inicializada con la longitud del tracto vocal del modelo como indica la ecuación 8.4; como por cada nueva trama sonora el sistema entrega el valor de la longitud del tracto estimado, la longitud del tracto vocal del locutor es actualizada de acuerdo a la ecuación 8.3 (con $\beta = 0.99$), y el factor de deformación para ésta trama se calcula finalmente con la ecuación 8.5 (con $\lambda = 0.5$).

$$LTV_{loc}(0) = \overline{LTV}_{modelo} \tag{8.4}$$

$$\alpha(i) = 1 + \lambda \frac{\overline{LTV}_{modelo} - LTV_{loc}(i)}{\overline{LTV}_{modelo}} \tag{8.5}$$

Otra técnica basada en características acústicas es la que tiene en cuenta el tercer formante propuesta por [Eide and Gish, 1996], y que posteriormente fue utilizada por [Gouvea and Stern, 1997] pero utilizando como característica acústica la media de los primeros tres formantes para estimar el mejor factor de deformación. Los resultados de este último estudio mostraron, que las tasas de error son ligeramente menores que otras técnicas y que hay una reducción en el costo computacional. Con el objetivo de aplicar esta técnica y comparar sus resultados con la técnica propuesta en este estudio, la estimación del factor de deformación a partir de F3 se implementó como muestra la Ecuación 8.6, donde $F3_{loc}$ es la media del tercer formante para un locutor, y $\overline{F3}$ es la media de F3 calculada para

todos los locutores utilizados en la fase de entrenamiento del modelo acústico, y se utilizó el mismo factor de $\lambda = 0.5$.

$$\alpha = 1 + \lambda \frac{F3_{loc} - \overline{F3}}{\overline{F3}} \tag{8.6}$$

8.3 Marco Experimental y Resultados

La evaluación de las técnicas aquí propuestas se han hecho sobre la base de datos TIDigits [Leonard, 1984]. Este corpus consta de 25 niños, 26 niñas, 55 hombres y 57 mujeres para el entrenamiento de modelos; 25 niños, 25 niñas, 56 hombres y 57 mujeres para la evaluación de reconocimiento. Se han diseñado siete condiciones con siete modelos acústicos diferentes entrenados para cada condición: *Niño*, *Niña*, *Hombre*, *Mujer*, *Adultos* (hombre y mujer), *Niños* (niño y niña) y *Todos* los locutores. Finalmente, el reconocimiento se realizó con todos los 163 locutores disponibles para evaluación.

Aplicando la técnica de estimación de la longitud del tracto vocal a los locutores de la base de datos TIDigits, tanto a los locutores de entrenamiento como a los de evaluación, se han calculado las medias y desviaciones estándar las cuales se muestran en la Tabla 8.1. Los valores confirman las grandes diferencias entre los grupos, especialmente en los hombres respecto a los demás grupos; situación que sustenta la necesidad de aplicar las técnicas de normalización de locutores y así disminuir la variabilidad.

Tabla 8.1: *Media de la longitud del tracto vocal (cm) y desviación estándar estimadas para los grupos de locutores en la base de datos TIDigits.*

Locutores de entrenamiento				
	Niño	*Niña*	*Hombre*	*Mujer*
VTL	16.0±0.64	15.5±0.65	18.8±0.67	16.6±0.64
Locutores de evaluación				
	Niño	Niña	Hombre	Mujer
VTL	15.9±0.74	15.4±0.58	18.8±0.71	16.6±0.63

En la tarea de reconocimiento, un conjunto de 11 palabras (one, two, three, four, five, six, seven, eight, nine, zero, oh) o modelos ocultos de Markov (HMM) representando dígitos en ingles fueron entrenados para cada condición. Un parametrizador tipo ETSI se utilizó para extraer el vector de características MFCC de cada señal, utilizando los primeros parámetros estáticos ($c1$-$c12$), más el logaritmo de la energía, y sus primeras y segundas derivadas constituyendo un vector final de 39 dimensiones. El sistema de RAH utilizado para los experimentos fue el de decodificación Viterbi.

Los resultados del *baseline* se muestran en la Tabla 8.2 en términos de Word Error Rate (WER) definida en la Ecuación 8.7, en donde se han entrenado los modelos acústicos para cada grupo y reconocido con cada grupo de evaluación de manera independiente. Como es de esperarse, los errores mínimos se encuentran cuando se entrenan los modelos acústicos y se reconoce bajo el mismo tipo de grupo, por ejemplo: entrenando con el grupo *Niña* y evaluando también con el grupo "Niña", la tasa de error es de 0.55%, mientras que

Tabla 8.2: *Resultados del baseline en WER para la base de datos TIDigits.*

Grupos de entrenamiento	Grupos de evaluación						
	Niño	Niña	Hombre	Mujer	Adultos	Niños	Todos
Niño	1.32%	1.11%	18.78%	1.55%	10.08%	1.21%	7.36%
Niña	1.55%	0.55%	50.84%	4.06%	27.23%	1.05%	19.20%
Hombre	50.94%	65.91%	0.66%	20.1%	10.47%	58.41%	25.16%
Mujer	4.27%	5.29%	10.88%	0.34%	5.56%	4.77%	5.32%
Adultos	4.6%	6.1%	0.67%	0.4%	0.53%	5.34%	2.01%
Niños	0.82%	0.51%	21.5%	1.74%	11.52%	0.66%	8.19%
Todos	0.93%	0.76%	0.77%	0.35%	0.55%	0.84%	0.65%

entrenando con el mismo grupo pero reconociendo con el grupo "Adultos", la tasa de error se eleva al 27.23%.

$$WER = \frac{Inserciones + Sustituciones + Borrados}{Numerodepalabrasareconocer} \tag{8.7}$$

Tomando como referencia los resultados de reconocimiento obtenidos con el grupo de evaluación "Todos", los peores resultados se presentan con los modelos de voz de *Hombre*, ya que ellos presentan los tractos vocales más largos quedando bastante separados del resto de locutores. En el otro extremo, en el grupo *Niña*, quienes tienen los tractos vocales más cortos tampoco presentaron buenos resultados en el reconocimiento, mientras que los modelos entrenados con el grupo *Todos* tienen el mejor resultado logrando un 0.65% de WER.

Para comparar las técnicas de normalización del tracto vocal basadas en modelos como la de máxima verosimilitud ML-VTLN y la de ML-GMMs, y las basadas en características como la de F3 y la propuesta en esta investigación basada en la longitud del tracto LTV, se han realizado seis experimentos y obtenido los resultados de reconocimiento para cada uno. El primer experimento es utilizando la técnica Off-line de ML-VTLN, el segundo experimento utiliza la técnica LTV sin aplicar el liftado de la Sección 5.2 y también de manera Off-line, es decir utilizando el α de la Ecuación 8.2, el tercero utiliza la misma técnica LTV pero aplicando el liftado, el cuarto experimento utiliza la técnica LTV en la versión On-line propuesta, la cual utiliza la Ecuación 8.5 para obtener el factor de deformación α, el quinto experimento está basado en el F3, y finalmente el sexo basado en la técnica de GMMs. Reconociendo sobre el grupo de evaluación "Todos", los resultados del *baseline* y de los seis experimentos mencionados se recopilan en la Tabla 8.3.

Los resultados obtenidos en las dos primeras técnicas Off-line de la tabla (la basada en ML-VTLN en la fila 2, y la basada en LTV fila 3), indican que la técnica LTV tiene un desempeño muy similar a la primera técnica, considerada como de referencia en el estado del arte lo que muestra un buen punto de partida. Por otra parte, la utilización de la técnica de liftado de la Sección 5.2 (fila 4) produjo una reducción del WER incluso mayor que la técnica ML-VTLN para los modelos acústicos entrenados con: *Niño*, *Niña* y *Niños*, lo que confirma que la estimación robusta de formantes en voz infantil repercute positivamente en las estimaciones de las longitudes, y por ende, en la normalización del tracto vocal de estos locutores; mientras que la técnica ML-VTLN muestra mejores resultados con los

Tabla 8.3: *Resultados en WER para la base de datos TIDigits en: baseline, tres técnicas Off-line y una On-line.*

	Niño	*Niña*	*Hombre*	*Mujer*	*Adultos*	*Niños*	*Todos*
Baseline	7.37%	19.21%	25.17%	5.32%	2.01%	8.20%	0.65%
Off-line ML-VTLN	2.47%	5.26%	8.58%	1.28%	1.05%	2.40%	0.57%
Off-line LTV	2.84%	5.37%	11.25%	1.94%	1.19%	2.81%	0.66%
Off-line LTV-liftado	2.35%	3.92%	10.15%	1.57%	1.07%	2.18%	0.65%
On-line LTV	2.61%	4.78%	10.48%	1.82%	1.18%	2.49%	0.65%
F3	4.34%	12.24%	14.27%	2.83%	1.47%	4.76%	0.62%
ML-GMMs	2.31%	5.06%	8.54%	1.22%	1.07%	2.64%	0.68%

modelos entrenados en *Hombre*, *Mujer* y *Adulto*. Los resultados obtenidos con los modelos entrenados con *Todos*, son similares en las técnicas LTV Off-line y On-line y mejores en la técnica ML-VTLN. La técnica de F3 muestra tasas elevadas de error con respecto a la mayoría de las técnicas propuestas, y una tasa ligeramente inferior para los modelos entrenados con *Todos*. Finalmente, la técnica ML-GMMs demuestra ser también una buena técnica basada en modelos ya que presenta resultados semejantes a la técnica ML-VTLN, y ligeramente mejores en los modelos entrenados con *Niño*, *Hombre*, y *Mujer* respecto a todas las técnicas.

El principal resultado de este capítulo es el desarrollo de un método para la normalización On-line del tracto vocal de locutores con aplicación en RAH. Este método se basa en la estimación robusta de la longitud del tracto vocal del locutor a partir de los formantes presentes en las tramas sonoras, lo que permite estimar un factor de deformación que puede ser actualizado y mejorado entre más información formántica se tenga del locutor. Con este método se supera el inconveniente de las técnicas tradicionales que requieren de varias etapas de análisis para estima el mejor factor de deformación, impidiendo su aplicación en tiempo real.

Aplicando el método propuesto en la base de datos TIDigit, los mejores resultados se obtuvieron con los modelos entrenados para *Niña* y *Niños* con respecto a todas las técnicas analizadas, también se obtuvieron resultados muy semejantes a técnicas de comprobada eficiencia como las basadas en modelos ML-VTLN y ML-GMMs, lo que confirma que, en general, se puede aplicar el método propuesto para normalizar el tracto vocal de locutores adultos o niños en sistemas de RAH en tiempo real.

Parte IV

Estudio Experimental y Resultados

Capítulo 9

Estudio Experimental

Cualquier trabajo que se plantea unos objetivos como los de esta tesis, en donde el producto de la investigación es el desarrollo de herramientas para trabajar con población con discapacidad, hace necesario poner a prueba la tecnología propuesta en casos reales para ver hasta que punto esta es realmente útil, e identificar la mejor manera de acercar la tecnología a esta población tan específica. La gran diversidad de discapacidades y las múltiples maneras de manifestarse en los individuos hace que se requieran herramientas fácilmente adaptables, y es cuando toma gran importancia el concepto profesional de los terapeutas que interactúan directamente con ellos, un estudio que reúna entonces pacientes, la parte técnica y la parte terapéutica, tiene más posibilidades de obtener buenos resultados.

En el transcurso de la investigación se establecieron importantes convenios con instituciones de educación especial, que posibilitaron la realización de un estudio aplicando la herramienta *PreLingua* a un grupo de personas con diferentes discapacidades con el objetivo de evaluar la tecnología propuesta y obtener resultados cualitativos y cuantitativos. Este capítulo describe entonces el estudio realizado, las características de la población participante y las dificultades encontradas, así como la metodología seguida para obtener los resultados.

9.1 Entidades Participantes

Los criterios para seleccionar las instituciones participantes del estudio fueron, por una parte que tuvieran alumnos cuyos padres o tutores autorizaran la participación de estos en el estudio por medio de un permiso debidamente diligenciado y firmado, también que los profesionales de dichas instituciones tuvieran la disposición y el tiempo de participar en el estudio ya que debía adecuarse a sus rutinas habituales de trabajo. Otro aspecto importante fue poder contar con una institución de algún país latinoamericano ya que el número de descargas en esta parte del mundo es muy numerosa, y las necesidades y recursos tecnológicos varían respecto a las presentes en España.

Teniendo en cuenta lo anterior, el estudio se realizó con el Colegio Público de Educación Especial (CPEE) "Alborada[1]" en Zaragoza España, quienes ya habían participado en la

[1] http://centros6.pntic.mec.es/cpee.alborada

adquisición del corpus descrito en el Capítulo 4, y cuyo compromiso y colaboración con la investigación fue siempre en todo momento inmejorable. Allí se contó con el continuo apoyo de dos de sus logopedas y una psicóloga, inicialmente participaron en el estudio 9 de sus alumnos y lo concluyeron 6 de ellos cuyas características se describen en la Sección 9.3. La otra institución participante fue la Fundación "Centro de Educación Especial del Niño Diferente CEDESNID[2]" en Bogotá Colombia, esta entidad apoyó el estudio con tres fonoaudiólogas y dos psicólogos, quienes participaron de manera incondicional y muy activa en todo momento. Esta institución participo inicialmente con 33 de sus alumnos pero que debido a diferentes dificultades finalmente participaron 21 alumnos.

Los entornos de trabajo se muestran en la Figura 9.1, los cuales constan sencillamente de un aula equipada con un ordenador con la herramienta *PreLingua* instalada, y un micrófono tradicional de escritorio.

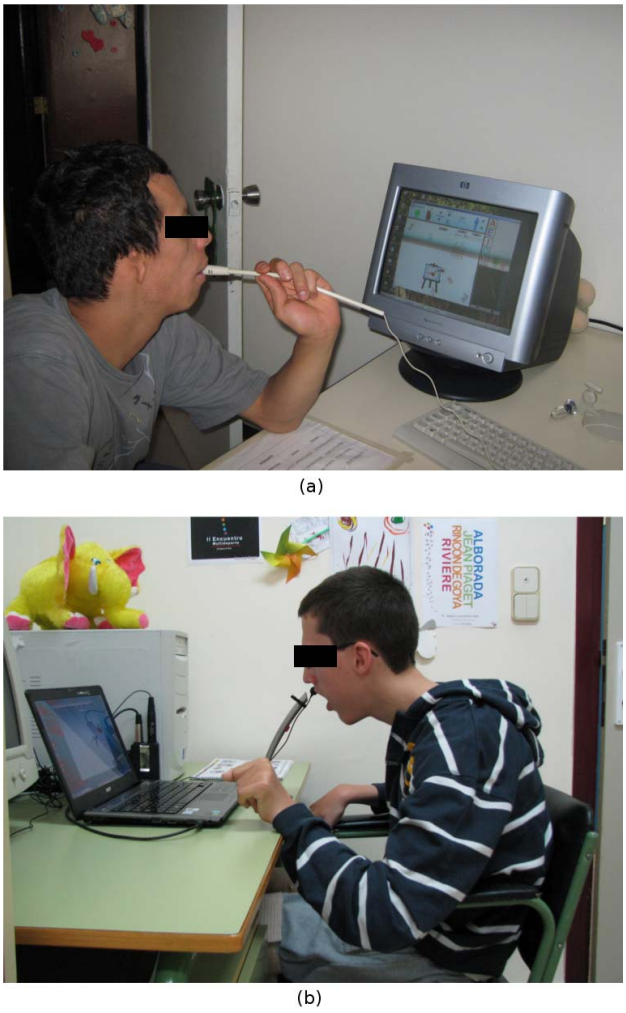


Figura 9.1: *Entorno de trabajo en Colombia (a) y España (b).*

[2]http://www.cedesnid.org

9.2 Dificultades del estudio

Como se mencionó en la sección anterior, dos instituciones de educación especial en Colombia y España participaron en el estudio de aplicación de *PreLingua*, con un total de 39 alumnos de los cuales 27 finalizaron el estudio. Lamentablemente los 12 alumnos restantes fueron descartados debido a varios factores que influirían notoriamente en los resultados del estudio, entre ellos se destaca:

- La causa más acusada fueron las frecuentes ausencias de algunos alumnos a las sesiones de trabajo por motivos como enfermedad, citas médicas, actividades no programadas previamente, o porque el día de la sesión de trabajo el alumno sencillamente no quiso colaborar.
- En algunos casos debido a la condición discapacitante del alumno como disfonías severas o malformación en el paladar, esta condición representaba serias dificultades para trabajar con la herramienta pese a que el alumno entendía el objetivo a cumplir, hasta el punto de que el alumno se estresaba o molestaba y por supuesto no era la finalidad del estudio.
- Algunos de los alumnos con diagnóstico de retraso mental que inicialmente se consideraron como candidatos para el estudio, en el transcurso de algunas sesiones de trabajo se comprobó que ellos no comprendían realmente los objetivos a cumplir en las actividades y que su participación obedecía al simple hecho de jugar como los demás alumnos participantes.

También se presentaron algunas dificultades menores relacionadas con el ruido del entorno de trabajo y los micrófonos utilizados, pero que por su naturaleza fueron fácilmente superadas sin mayores contratiempos.

9.3 Población Participante

La población participante del estudio consta de 27 casos (21 de Colombia y 6 de España), de los cuales 20 corresponden al sexo masculino (74%) y 7 al sexo femenino (26%). Sus edades oscilan entre los 11 y los 34 años de edad, siendo la población más numerosa y con más edad la perteneciente a la institución colombiana. Pese a que la herramienta se diseño para población infantil, en la institución CEDESNID quienes también manejan población adulta, consideró oportuno incluir alumnos adultos ya que según su criterio profesional ellos se verían beneficiados igualmente con su participación en el estudio.

Los siguientes son los criterios utilizados para la selección final de los casos de estudio:

- Que el alumno a pesar de su discapacidad comprendiera el objetivo de las actividades.
- Que tuviera alguna alteración en su voz o en su capacidades de comunicación en donde *PreLingua* ofreciera una ayuda.
- Que asistiera a un mínimo del 50% de las sesiones.
- Que el padre o tutor autorizara por escrito su participación en el estudio.

La Tabla 9.1 muestra los 27 casos de estudio con sus características como: Sexo, Edad, Ubicación y Diagnóstico, en el estudio no se utilizaron los datos personales de los participantes en su lugar se utilizó la notación Caso 1 al Caso 27.

Tabla 9.1: *Características de la población.*

Caso	Sexo	Edad	Ubicación	Diagnóstico
1	masculino	14	Colombia	Retraso mental moderado, Parálisis cerebral
2	masculino	18	Colombia	Retraso mental leve, Parálisis cerebral
3	masculino	13	Colombia	Desorden de comunicación
4	masculino	17	Colombia	Retraso mental moderado, Parálisis cerebral
5	masculino	22	Colombia	Retraso mental moderado, Desorden de comunicación
6	masculino	18	Colombia	Retraso mental moderado, Parálisis cerebral
7	masculino	20	Colombia	Retraso mental moderado, Síndrome convulsivo.
8	masculino	34	Colombia	Retraso mental moderado, Síndrome de Down
9	masculino	18	Colombia	Retraso mental moderado, Síndrome convulsivo.
10	masculino	24	Colombia	Desorden de comunicación
11	masculino	23	Colombia	Retraso mental severo
12	femenino	18	Colombia	Retraso mental moderado
13	masculino	18	Colombia	Retraso mental severo
14	masculino	17	Colombia	Retraso mental moderado
15	femenino	34	Colombia	Retraso mental moderado
16	masculino	13	Colombia	Retraso mental severo
17	masculino	14	Colombia	Retraso mental moderado
18	masculino	17	Colombia	Retraso mental moderado
19	masculino	21	Colombia	Retraso mental moderado
20	femenino	21	Colombia	Retraso mental moderado
21	masculino	12	Colombia	Retraso mental moderado
22	masculino	11	España	Retraso mental moderado, Hipertonía
23	femenino	16	España	Retraso mental moderado, Hipotonía
24	masculino	15	España	Retraso mental moderado, Tetraplegia
25	femenino	14	España	Retraso mental moderado
26	femenino	14	España	Retraso mental moderado
27	femenino	16	España	Retraso mental moderado, Síndrome de Down

9.4 Estudio

Teniendo en cuenta los recursos humanos y físicos disponibles en las instituciones y que el tiempo a invertir en el estudio no podía representar una carga excesiva de trabajo, se diseño un estudio de 12 semanas de duración en donde los terapeutas utilizarían la herramienta semanalmente en sus sesiones habituales de trabajo, y una vez por semana grabarían dichas sesiones con las utilidades que ofrece la herramienta, para enviar los datos generados al laboratorio de investigación en Zaragoza para su análisis. Una versión especial de *PreLingua* se desarrolló para que permitiera grabar la señal de voz en un archivo junto con los reportes estadísticos generados por la herramienta descritos en la Sección 7.1.8. Esta

versión especial de *PreLingua* se instaló en las dos instituciones participantes y se realizaron pruebas de funcionamiento durante dos semanas previas al inicio del mismo. Las actividades incluidas en le estudio fueron: la Intensidad, el Soplo, y el Tono, donde participaron los 27 sujetos, y finalmente la Articulación de vocales con la participación de 24 sujetos.

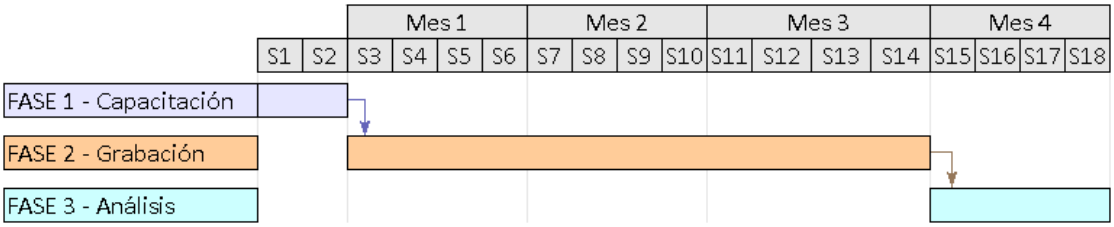


Figura 9.2: *Diagrama de Grantt del estudio.*

La Figura 9.2 muestra el diagrama de Grantt del estudio distribuido en tres fases que abarcan 18 semanas de trabajo. La FASE 1 es una etapa de capacitación de dos semanas de duración en las cuales las instituciones instalaron la herramienta y las terapeutas adquirieron destreza en el manejo de la misma, una vez finalizada esta etapa inicia la FASE 2 de grabación, la cual tiene una duración de 12 semanas y en cada una de ellas las terapeutas utilizaron la herramienta en sus clases normales con los alumnos, y la última sesión de la semana fue grabada por el sistema, finalmente los datos generados en dicha sesión por cada alumno se enviaron por vía electrónica al laboratorio en Zaragoza para su registro y procesamiento. La FASE 3 de análisis busca principalmente la generación de resultados cuantitativos a partir de los reportes generados por la herramienta, como muestra la Figura 9.3, se registraron los datos semanalmente para cada caso de estudio considerando: los valores mínimos y máximos, el rango dinámico, la media, el error cuadrático medio y el tiempo e sesión para las actividades de: Intensidad, Soplo, Tono y la Articulación vocálica. Como muestra la misma figura, el caso 16 presento cuatro ausencias a las terapias y la primera semana no colaboró en la actividad de Intensidad, revelando la dificultad más frecuente del estudio.

9.4.1 Evaluación logopédica

Antes de iniciar la FASE 2 y justo después de finalizada, el terapeuta realizó a cada alumno una evaluación logopédica para conocer las características y alteraciones de la voz del paciente antes y después del estudio, esto permitirá comparar de manera cualitativa si el alumno presento alguna modificación o mejora en su voz después de haber trabajado con la herramienta.

La evaluación logopédica de voz fue creada por profesionales del área de audición y lenguaje de la "Junta de Andalucía"[3], de la comunidad autónoma de Andalucía en España, su uso fue recomendado por los profesionales del colegio público de educación especial "Alborada" en Zaragoza donde se realizó el estudio. Esta evaluación que puede observase

[3]http://usuarios.multimania.es/maestrosayl/evaluacion-lenguaje.htm

en el Apéndice B, incluye aspectos de la adquisición prelingüística como los citados en la Sección 9.4.1 y evalúa también aspectos propios de la voz. En resumen, evalúa:

- Aspectos previos al lenguaje como: la capacidad de atención, la percepción visual, la percepción auditiva, y la imitación de sonidos.
- Cualidades de la voz como: el tipo de voz, la entonación y el ritmo.
- Una evaluación anatómica de: paladar, lengua, velo de paladar, frenillo, úvula, labios, dientes y amígdalas.
- La capacidad de relajación, y todo lo relativo a la respiración.
- La imitación de expresiones faciales y práxias de: lengua, labios, mejillas, y maxilares.
- La movilidad del velo del paladar.

Esta evaluación se basa en escalas subjetivas, en donde la experiencia del terapeuta es fundamental para una correcta valoración de los pacientes.

9.4.2 Evaluación objetiva

	Caso	**16**											
	Sexo	M											
	Edad	13											
	Talla	156											
	Diagnóstico	Discapacidad Cognitiva Moderada-Severa											
	SEMANAS	SEM 1	SEM 2	SEM 3	SEM 4	SEM 5	SEM 6	SEM 7	SEM 8	SEM 9	SEM 10	SEM 11	SEM 12
INTENSIDAD	Intensidad MIN (dB)	0	0	49	32	61	41	0	38	0	61	0	41
	Intensidad MAX (dB)	0	0	73	70	69	64	0	59	0	69	0	72
	Rango Dinámico	0	0	24	38	8	23	0	21	0	8	0	31
	Media (dB)	0	0	67	64	66	60	0	55	0	65	0	68
	ECM Intensidad	0	0	21,39	19,07	20,94	16,69	0	10,22	0	8,32	0	9,33
	Tiempo de sesión	0	0	14,5	11,5	11	12	0	11,5	0	12	0	13,5
SOPLO	Intensidad MIN	54	0	50	57	49	57	0	44	0	53	0	39
	Intensidad MAX	70	0	74	70	70	70	0	59	0	70	0	70
	Rango Dinámico	16	0	16	13	21	13	0	15	0	17	0	31
	Media	69	0	69	68	68	67	0	54	0	63	0	58
	ECM Soplo	22	0	22,2	20,86	21,54	19,18	0	8,17	0	9,86	0	7,11
	Tiempo de sesión	63,5	0	58	57	61	57	0	50	0	54	0	54,5
TONALIDAD	Frecuencia MIN (Hz)	160	0	157	148	151	140	0	163	0	133	0	160
	Frecuencia MAX (Hz)	200	0	218	242	235	222	0	182	0	195	0	200
	Rango Dinámico	40	0	61	94	84	82	0	19	0	62	0	40
	Media (Hz)	190	0	196	198	205	172	0	170	0	166	0	174
	ECM Tono	38,39	0	43,3	48,16	44,34	34,4	0	11,71	0	7,78	0	10,55
	Tiempo de sesión	11	0	9	9,5	10	10,5	0	12,5	0	11	0	11,5
ARTICULA	ECM min /a/	3,7	0	4,2	4,4	3,1	3,8	0	4,4	0	4	0	3,9
	ECM min /e/	3,6	0	3,3	4	3,1	3,2	0	3	0	3,5	0	3,8
	ECM min /i/	7,9	0	7,2	7,7	6,5	3,5	0	6,1	0	5,5	0	5,1
	ECM min /o/	4	0	3,3	3,6	3,4	3,8	0	3,3	0	4,3	0	3,8
	ECM min /u/	5,5	0	5,5	5,6	5,1	7,3	0	2,8	0	3,4	0	2,7

Figura 9.3: *Registro de datos semanal.*

La evaluación objetiva se realizó con los datos generados por la herramienta en las sesiones grabadas semanalmente, se consideraron los datos registrados de las primeras tres sesiones del estudio y las últimas tres sesiones. En cada sesión, los usuarios recibieron

las instrucciones necesarias para realizar las actividades de Intensidad, Soplo, Tono, y ARTICULA, y los datos generados por la herramienta con los 27 casos de estudio se almacenaron para su posterior análisis. La evaluación consistió en buscar diferencias entre los resultados de las sesiones iniciales y finales detectando cambios significativos en las habilidades de cada usuario en cada actividad, para conseguirlo, se consideró el Error Cuadrático Medio (ECM) existente entre el patrón establecido por el terapeuta en las diferentes actividades y el patrón descrito por el usuario con su voz; también se calculó la media y la desviación estándar para las sesiones iniciales y finales de manera independiente.

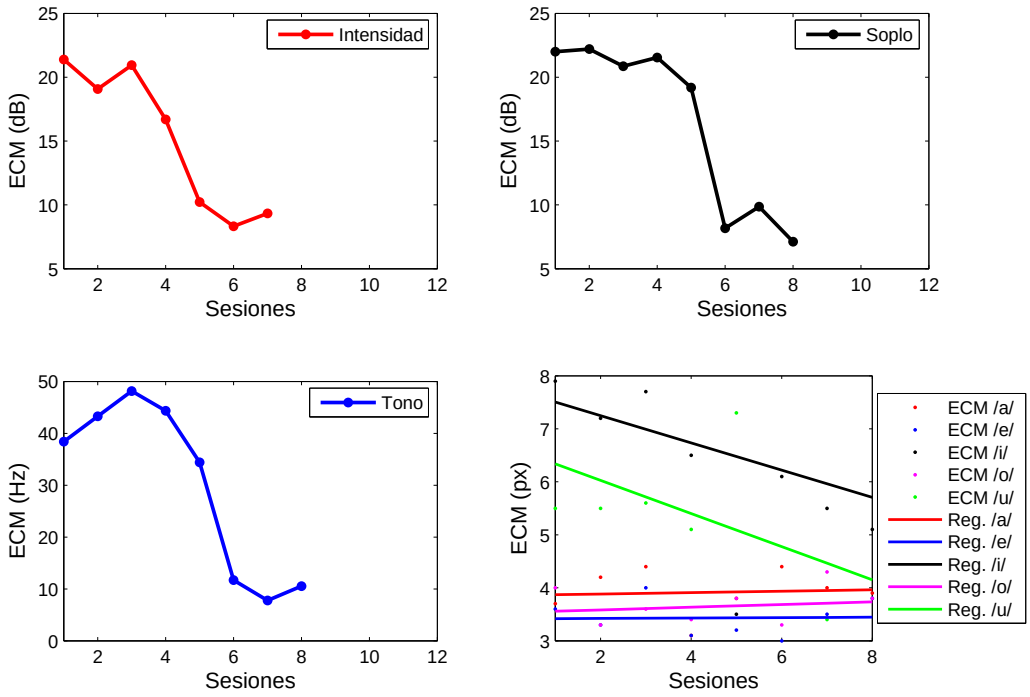


Figura 9.4: *Registros de Intensidad, Soplo, Tono, y Vocales para el caso 16.*

Retomando el caso 16 de la Figura 9.3, la evolución de los errores cuadrático medios registrados durante el estudio para intensidad, soplo, tono y articulación vocálica, se muestran en la Figura 9.4. La curva de Intensidad muestra un total de 7 sesiones y una disminución del ECM de 10dB aproximadamente, en la actividad de soplo, se registran 8 sesiones de trabajo con una disminución de 13dB en el ECM entre las sesiones iniciales y finales. El tono muestra 8 sesiones de las cuales las tres primeras muestran un incremento en el error estimado, pero luego presenta una disminución de hasta 35Hz con respecto a las sesiones finales, lo que evidencia un mejor manejo sobre las cuerdas vocales. Finalmente, en la articulación vocálica se muestra una disminución en la línea de tendencia del ECM en las vocales /i/ y /u/, mientras que en las vocales /a/, /e/ y /o/ se muestra por el contrario un aumento en el error a lo largo de las sesiones de trabajo.

El caso presentado aquí es representativo en el sentido de que presenta mejoras en todas las actividades evaluadas y se cita aquí, para clarificar la metodología del estudio en la

evaluación objetiva. Infortunadamente no todos los casos de estudio presentaron mejoras como se mostrará en los resultados finales del estudio en el Capítulo 10.

Capítulo 10

Resultados

Los resultados aquí presentados se dividen en resultados cuantitativos en términos de significancia estadística, y en resultados cualitativos derivados de las evaluaciones logopédicas realizadas por los terapeutas a cada usuario. También se han tenido en cuenta las experiencias obtenidas a lo largo de la investigación por parte de quienes de alguna manera han participado en el proyecto, y las aportaciones recibidas de quienes simplemente han descargado y utilizado la herramienta.

10.1 Resultados Cuantitativos

La Tabla 10.1 muestra los resultados de como los diferentes casos de estudio alcanzaron o no una mejora en sus habilidades de voz considerando: la Intensidad, el Soplo, el Tono y la Articulación vocálica, de acuerdo a los valores de ECM registrados en el estudio. La mejora (Mej. S=Si, N=No) se consideró como existente (S) en aquellos casos con una reducción del ECM entre la media de las tres sesiones iniciales y la media de las tres sesiones finales, con una significancia estadística por encima del 50%, y se consideraron sin mejora (N) aquellos casos sin reducción del ECM entre las sesiones iniciales y finales, o con una significancia estadística menor al 50%. Como se mencionó en la Sección 7.1.8, el ECM mide la distancia entre el patrón establecido por el terapeuta para: Intensidad, Soplo, y Tono, y el patrón descrito por la emisión sonora del usuario, de igual manera en la articulación vocálica el ECM es el existente entre la lengua patrón mostrada por el sistema, y la lengua del avatar movida por la emisión sonora del usuario.

Con base a los datos obtenidos por el sistema, una prueba t-test de significancia estadística fue aplicada para cada usuario con el fin de establecer si las mejoras obtenidas en las habilidades de voz fueron realmente significativas o si fueron producto del azar. Los valores de ECM obtenidos al principio y al final del estudio (X_1 y X_2) fueron caracterizados con sus medias y desviaciones estándar ($\overline{X_1}$, $\overline{X_2}$, σ_1 y σ_2) y el número de muestras en cada caso (n_1 y n_2) fue de 3, ya que se analizaron las tres sesiones iniciales y las tres sesiones finales del estudio. Como la varianza de los datos a analizar era diferente, una adaptación de la prueba t Student llamada test de Welch fue utilizada, en esta prueba el estadístico t utilizado para verificar si las medias eran significativamente diferentes fue calculado con la Equación 10.1, con $S_{\overline{X_1}-\overline{X_2}}$ como el estimador insesgado de las varianzas tal y como se define en la Ecuación 10.2.

$$t = \frac{\overline{X_1} - \overline{X_2}}{S_{\overline{X_1} - \overline{X_2}}} \tag{10.1}$$

$$S_{\overline{X_1} - \overline{X_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{10.2}$$

Para el test de significancia, la distribución de la estadística se aproximó a una distribución t Student de doble cola de datos no apareados, con grados de libertad obtenidos por la ecuación de Welch-Satterthwait. Los resultados de significancia obtenidos que se muestran en la Tabla 10.1, fueron muy variables para cada actividad y cada caso de estudio, ya que la voz alterada y la discapacidad misma hacen que aumenten las variables que pueden afectan los resultados. Considerando un nivel de significancia ≥99%, el estudio muestra mejoras en 4 sujetos (14.8% del total) en la actividad de Intensidad, 5 sujetos (18.5%) mostraron mejoras en Soplo, un solo sujeto mostró mejoras en la actividad de Tono (3.7%) y, 2 sujetos (8.3%) en al menos una vocal. Con un umbral más bajo pero conservando una alta significancia (≥95%), el estudio mostró mejoras en 8 sujetos (29.6% del total) en Intensidad, 7 sujetos (25.9%) en Soplo, 6 sujetos (22.2%) en Tono y, 5 sujetos (20.8%) en al menos una vocal. Finalmente, con un nivel de significancia ≥80% se registraron mejoras en 15 sujetos (55.6%) en Intensidad, 15 sujetos (55.6%) en Soplo, 8 sujetos (29.6%) en Tono y, 8 sujetos (33.3%) en al menos una vocal. Cabe recordar que en las actividades de Intensidad, Soplo y Tono, el total de casos de estudio fue de 27 mientras que en articulación vocálica fueron solo de 24 casos.

En general, Intensidad y Soplo fueron las actividades donde más sujetos alcanzaron mejoras significativas en todos los niveles de significancia descritos. Estos resultados son especialmente buenos en la actividad de Soplo, ya que esta actividad requiere un alto nivel de concentración comparada con la actividad de Intensidad que es considerada la más fácil de realizar. Un número menor de usuarios alcanzaron mejoras significativas en la actividad de Tono, posiblemente influenciada por la corta duración del estudio y el alto nivel de exigencia requerido en esta actividad a nivel de conciencia, y control/modulación sobre las cuerdas vocales.

Respecto a la actividad de articulación vocálica, es bien sabido que el proceso de articulación esta afectado por las condiciones geométricas de la cavidad vocal del usuario, algunos de los usuarios presentaban mal formaciones en el paladar blando o duro, dientes torcidos, y/o hipotonía o hipertonia, de manera que los resultados obtenidos en esta actividad no fueron tan relevantes como en las otras actividades, y se espera sean mejores en la medida en que se utilice más la herramienta y se hagan más sesiones de trabajo. Comparando las diferentes vocales, un pequeño número de usuarios presentó mejoras en la articulación de la vocal /o/ (solo 2 casos significativos ≥90%) y de la vocal /a/ (solo 1 caso significativo ≥77%), cuyos primeros formantes $F1$ son mayores que en las otras vocales y esta correlado con la altura de la lengua, situación que resalta la dificultad por parte de los usuarios en la apertura de la boca. Por otro lado, las mejoras en la articulación de las vocales /e/, /i/ y /u/ fueron más numerosas con 8 casos con un nivel de significancia ≥68% y 4 casos con un nivel de ≥90%, estas vocales al ser altas requieren de un menor esfuerzo en la apertura de la boca.

Tabla 10.1: *Resultados Cuantitativos para: Intensidad, Soplo, Tono, y Articulación, para cada caso de estudio. S (Si): Mejora o reducción del ECM entre las sesiones iniciales y finales, N(No): No hay mejora o reducción del ECM.*

Caso	Intensidad		Soplo		Tono		Articulación	
No.	*Mej.*	Sig.%	*Mej.*	Sig.%	*Mej.*	Sig.%	*Mej.* Vocal	Sig.%
1	S	76.6	S	84.9	N	-	/u/	59.0
2	S	99.6	N	-	S	80.0	N	-
3	S	92.3	N	-	N	-	/e/	71.9
4	S	66.7	N	-	N	-	N/A	N/A
5	N	-	N	-	S	78.2	N/A	N/A
6	N	-	N	-	S	78.8	/i/ /o/	80.9 62.2
7	S	75.3	N	-	N	-	N/A	N/A
8	S	91.4	N	-	N	-	N	-
9	S	86.4	S	80.1	N	-	N	-
10	S	79.4	S	83.7	S	98.0	N	-
11	S	91.3	S	74.5	S	84.6	N	-
12	S	98.3	N	-	N	-	N	-
13	S	98.6	S	68.1	S	96.4	N	-
14	S	97.1	N	-	S	65.5	N	-
15	S	99.9	S	99.9	N	-	/o/	90.0
16	S	99.9	S	99.5	S	99.2	/i/ /u/	98.9 99.3
17	S	94.1	S	99.4	S	98.5	/e/	60.9
18	S	90.4	S	99.2	S	96.4	/a/	77.5
19	S	93.0	S	86.3	N	-	/e/	74.0
20	S	97.1	S	99.6	N	-	N	-
21	S	99.9	S	95.6	S	97.0	N	-
22	N	-	S	90.1	N	-	/u/	68.8
23	S	78.3	S	86.1	N	-	/a/ /i/	67.4 91.2
24	N	-	N	-	N	-	/a/ /e/ /u/	58.2 73.6 97.2
25	S	68.3	S	98.6	N	-	/o/ /u/	99.5 67.7
26	N	-	S	86.9	N	-	/i/	96.8
27	N	-	S	88.6	N	-	N	-

10.2 Resultados Cualitativos

De acuerdo a las observaciones realizadas por los terapeutas, en la Tabla 10.2 se han resumido los 27 casos de estudio con la información cualitativa de las evaluaciones logopédicas realizadas antes y después de aplicar *PreLingua*. Los tópicos descritos en la tabla son: Intensidad, Duración del Soplo, Tono, Práxias de lengua, Ritmo y finalmente una columna con aquellos tópicos resaltados por los terapeutas como Habilidades Adicionales Observadas (HAO) en los usuarios al finalizar el estudio.

Tabla 10.2: *Resultados Cualitativos. Evaluaciones logopédicas antes y después del estudio.* **A***: Astenica,* **AL***: Alterada,* **DS***: Dirección de Soplo,* **BR***: Bradilalia,* **NP***: No Puede,* **D***: Disminución,* **SI***: Seguimiento de Instrucciones,* **Ent***: Entrecortado,* **Au***: Aumento,* **AtA***: Aumento del Tiempo de Atención,* **M***: Monótono,* **N***: Normal,* **R***: Robótico,* **Hab***: Habilidad,* **HS***: Habilidades de Socialización,* **AS***: Áspera,* **TL***: Taquilalia,* **CD***: Con Dificultad,* **CE***: Con Esfuerzo,* **H.A.O.***: Habilidades Adicionales Observadas.*

Caso	Intensidad		Duración Soplo		Tono		Práxias Lengua		Ritmo		H.A.O.
No.	Antes	Después	Antes	Después	Antes	Después	Antes	Después	Antes	Después	++
1	N	N	N	N, Au-Hab	N	N	NP	Au-Hab	Ent, CD	Au-Hab	AtA, SI, HS
2	N	N	N	N, Au-Hab	N	N	NP	Au-Hab	N, CD	Au-Hab	HS
3	N	N	N	N, Au-Hab	R	N	N	N	N, Ent	Au-Hab	AtA, SI, HS
4	A	N	N	N, Au-Hab	N	N	-	-	Ent, CD	Ent, CD	AtA
5	A	A	N	N	M	M-N	-	-	BR	N	-
6	A	N	AL	AL, N-CE	M	N	N	N	N, CD	N	AtA, SI, HS
7	N	N	N	N, Au-Hab	N	N	-	-	Ent	N	AtA
8	A	N	AL	AL	N	N	N	N	N	N	-
9	A	A	N	N, Au-Hab	M	M	CD	CD	Ent, CD	N	SI, HS
10	N	N	N	N, Au-Hab	N	N	N	N	N, CD	N	-
11	N	N	AL	AL	M	M	CD	CD	Ent	Ent	AtA, HS
12	AS	N-AS	AL	AL	M	M	N	N	N	N	HS
13	AL	AL	AL	AL	M	M	N	N	BR	BR	SI, AtA
14	AS	N-CE	N	N	M	N-CE	N	N	TL	TL	-
15	AS	N-CE	AL	N-CD	N	N	N	N	TL	TL	HS
16	AS	AS	AL	AL	M	M	N	N, Au-Hab	TL	TL	HS
17	N	N	AL	AL, Au-Hab	N	N, Au-Hab	N	N	N	N	DS
18	AL	AL	AL	AL	M	M	N	N	TL	TL	SI
19	AS	AS	N	N	N	N	N	N	TL	TL, N-CE	DS
20	AL	AL, N-CD	AL	AL, N-CE	M	M	N	N	TL	TL	-
21	A	N-CE	AL	AL, N-CE	N	N	N	N	N	N	DS, HS
22	AS	N-CE	N	N, Au-Hab	N	N	CD	N, CE	TL	TL	AtA, HS
23	A	N-CE	N	N, Au-Hab	M, R	M, N-CE	CD	CD, Au-Hab	Ent	Ent	SI, AtA
24	N	N	AL	AL, Au-Hab	N	N	CD	CD, Au-Hab	N	N	AtA, HS
25	A	A, Au-Hab	AL	AL, Au-Hab	M	M, N-CE	N	N, Au-Hab	N, Ent	N	HS
26	A	A, Au-Hab	AL	AL, Au-Hab	M	M, N-CE	N	N, Au-Hab	Ent	Ent, N-CE	-
27	N	N	N	N, Au-Hab	N	N	N	N	N	N	SI, DS

La evaluación logopédica de voz utilizada y que se muestra en el Apéndice B, incluye varios aspectos de la adquisición prelingüística y de la voz misma. Todos estos aspectos fueron evaluados en escalas subjetivas utilizadas por los terapeutas de acuerdo al tópico evaluado, por ejemplo, el ritmo fue evaluado como: normal, con taquilalia, entre-cortado o con bradilalia; o por ejemplo el tono fue evaluado como: normal, monótono o robótico. En la Tabla 10.2, cada columna muestra la valoración hecha por el terapeuta antes y después del estudio según las escalas establecidas a manera de siglas las cuales están explicadas en la cabecera de la tabla.

Duración del Soplo fue la actividad donde un mayor número de usuarios mostraron mejoría según las evaluaciones de los terapeutas, seguido de la intensidad y el ritmo. 12 sujetos (44.4%) mostraron cambios positivos al finalizar el estudio en la actividad de Intensidad, por ejemplo, cambios de una voz áspera a una voz normal aunque con algunas dificultades, o aumento en la habilidad vocal en voces asténicas. 18 sujetos (66.6%) mejoraron sus habilidades en la duración del soplo como en los casos 1,2,3,4, y 7 entre otros, quienes evidencian un mejor control sobre los pliegues vocales, labios y demás estructuras que regulan la salida del aire. En el Tono, 8 sujetos (29.6%) presentaron una evolución positiva en este parámetro de la voz ya que pasaron de una entonación monótona o robótica a una entonación normal o normal con dificultades. Práxias de Lengua es la actividad en donde *Articula* interviene directamente, 8 casos de estudio (33.3%) mostraron una mejoría en esta habilidad según los terapeutas, y en 7 de ellos aumentó la habilidad para los movimientos de la lengua. El ritmo se apoya en actividades como Ataque Vocal y Duración, en donde 11 casos de estudio (40.7%) mostraron mejoras en este parámetro de la voz según las evaluaciones, mostraron un aumento en la habilidad de control sobre el ritmo y en algunos casos alcanzaron un ritmo de voz normal.

Finalmente, la última columna de la Tabla 10.2 reúne las habilidades adicionales observadas por los terapeutas al finalizar el estudio y que al inicio de este no se esperaban obtener. De los 27 casos de estudio, 21 de ellos mostraron alguna habilidad adicional como por ejemplo: Aumento del tiempo de atención (AtA), seguimiento de instrucciones (SI), habilidades de socialización (HS) y, direccionalidad del soplo (DS).

Al final del estudio se les pidió a los terapeutas evaluar el trabajo realizado con la herramienta, ellos consideran que la herramienta es fácil de usar y muy atractiva para los usuarios finales, resaltan las mejoras observadas al finalizar el estudio en algunos casos de estudio como la habilidad para sostener el soplo en lugar de presentar patrones interrumpidos, también las mejoras en los seguimientos de patrones continuos planos y ondulantes en las emisiones sonoras. Los terapeutas citan también la posibilidad de aplicar la herramienta en diferentes áreas relacionadas con la educación especial como en niños sordos, casos de mutismo, autismo y apráxias, y también la posibilidad de aplicarla en casos de accidentes cerebro vasculares en adultos donde se ve afectada la voz y el habla.

Otras observaciones mencionadas por usuarios de la herramienta en hispanoamérica son la mejora en la captura de atención, mejores niveles de concentración y memorización, y una alta motivación por parte del usuario final. Desde un punto de vista senso-perceptual, algunos usuarios finales muestran una mejor coordinación y localización espacial de elementos en pantalla y también una mejora en la percepción visual y auditiva.

Mencionan también que en las habilidades de comunicación algunos sujetos muestran un incremento de emisiones sonoras e inteligibles, que ocurren en situaciones y momentos cotidianos donde no se esta utilizando la herramienta. También observan un aumento en las habilidades de socialización entre los niños, mencionan actitudes positivas como juego en equipo, respetar los turnos para jugar en clase, la ayuda entre ellos, sana competencia, y autoexigencia en algunos casos. Respecto a *Articula*, los terapeutas evalúan la aplicación como amigable y fácil de entender, resaltan la apropiada y comprensible interface de la herramienta para trabajar articulación vocálica en tiempo real, lo que a criterio de todos garantiza una adecuada motivación y entendimiento en las sesiones de trabajo.

Parte V

Discusión y Conclusiones

Capítulo 11

Discusión

Los principales puntos de discusión que surgen al concluir esta investigación y después de analizar los resultados obtenidos pueden tratarse bajo tres planteamientos. Por una parte ver hasta que punto *PreLingua* puede considerarse como una herramienta para terapia y evaluación de voz, otro aspecto de importante discusión y muy relevante debido a que ésta es una tesis de aplicación, es el impacto que la herramienta ha tenido en la comunidad terapéutica de voz y de educación especial. Finalmente, el impacto de la aplicación de la tecnología propuesta en otras áreas de la discapacidad, así como en el reconocimiento automático de habla.

11.1 *PreLingua* como Herramienta para Terapia y Evaluación de Voz

Después de analizar los resultados del Capítulo 10 es importante determinar si *PreLingua* puede ser considerada como una herramienta adecuada para mejorar las habilidades en el manejo de la voz de pacientes que tienen alteraciones en su voz, y ver si *PreLingua* y sus actividades de evaluación pueden también servir para evaluar estas habilidades en diferentes usuarios a lo largo del tiempo. Es importante resaltar que estos resultados provienen del trabajo con población con discapacidad lo que dificulta en gran medida cualquier terapia con ellos, aplicar la tecnología propuesta en población con alteraciones en su voz pero sin discapacidad, seguramente tendrá mejores resultados pero serán necesarios entonces los estudios de rigor para corroborarlo.

En el estudio, los resultados cualitativos que se resumen en la Figura 11.1, mostraron un número de sujetos que usando *PreLingua* durante 12 semanas, ciertamente mejoraron sus habilidades de voz en aspectos como el soplo, la intensidad, y el ritmo, en un 66.6%, 44.4% y 40.7% de los sujetos respectivamente al finalizar el estudio. Habilidades como la entonación y praxias de lengua las cuales requieren de un mejor control de los músculos y demás estructuras anatómicas, mostraron resultados menos relevantes con un 29.6% y 33.3% respectivamente, convirtiéndosen en los aspectos de la voz más difíciles de trabajar por parte de los usuarios. Debido al alto nivel de entendimiento y concentración requeridos, se necesitaría de un estudio más amplio (quizá de seis meses a un año) y con más casos de estudio que refleje mayores variaciones de estos parámetros en el tiempo, para saber si

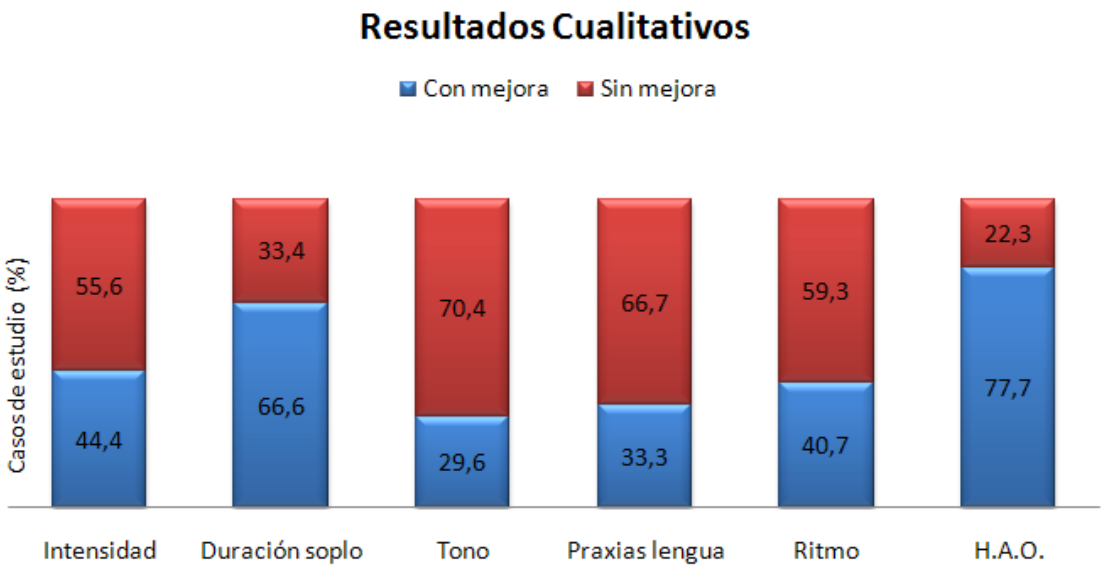


Figura 11.1: *Resumen Resultados Cualitativos.*

PreLingua tiene las mismas posibilidades de mejorar estos aspectos de la voz en sujetos con voz alterada, como las mejoras obtenidas en aspectos como la intensidad, el soplo y el ritmo. Una parte muy positiva de los resultados cualitativos son las habilidades adicionales observadas (H.A.O.) que, aunque no se esperaban al inicio del estudio, éstas se presentaron en el 77.7% de los casos de estudio convirtiéndosen en los mejores resultados cualitativos obtenidos. Por su parte, los resultados cuantitativos resumidos en la Figura 11.2 muestran, que objetivamente la intensidad fue la actividad con mejores resultados en todos los niveles de significancia con respecto a las demás, seguida del soplo, la articulación vocálica y finalmente el tono.

Una parte muy interesante del estudio aparece cuando se comparan los casos de estudio que obtuvieron mejoras según las evaluaciones logopédicas, con los casos de estudio que obtuvieron mejoras según las actividades de evaluación de la herramienta en los resultados cuantitativos, dicha comparación puede apreciarse en la Figura 11.3. De un total de 27 casos de estudio, 9 de 12 casos de estudio (casos: 4,8,12,14,15,20,21,23,25) que presentaron mejoras en la intensidad según las evaluaciones logopédicas, también presentaron mejoras en la evaluación objetiva con diferentes niveles de significancia. De igual manera, 12 de 18 casos de estudio (casos: 1,9,10,15,17,20,21,22,23,25,26,27) que presentaron mejoras en la actividad de soplo en la evaluación logopédica, presentaron también mejoras en la evaluación objetiva con diferentes niveles de significancia. El tono por su parte, fue la actividad con el menor número de coincidencias entre los resultados con 4 casos de estudio (casos: 5,6,14,17) de 8 posibles según la evaluación logopédica y 11 posibles según los resultados cuantitativos y con diferentes niveles de significancia. De los 24 casos de estudio que participaron en actividad de articulación vocálica, 13 de ellos mostraron mejoras según la evaluación objetiva en al menos una vocal y con diferentes niveles de significancia, y 7 de ellos (casos: 1,16,22,23,24,25,26) fueron evaluados también positivamente en las evaluaciones logopédicas de un total de 8 casos, de manera que la actividad de articulación fue la que mayor número de coincidencias presentó.

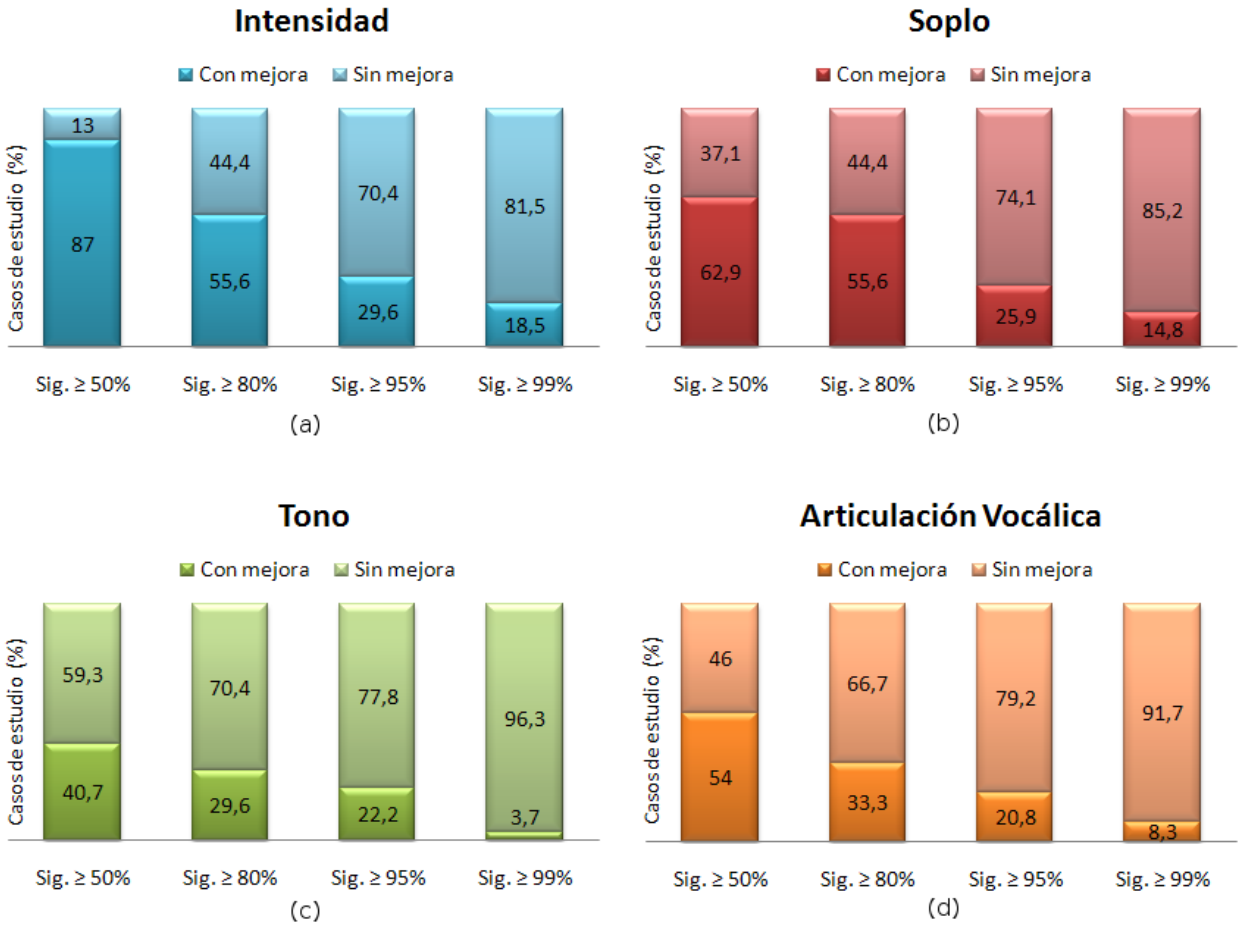


Figura 11.2: *Resumen Resultados Cuantitativos.*

En general, hubo coincidencias entre los terapeutas y la herramienta en establecer si un sujeto ha mejorado sus habilidades de voz, aunque hubo un número de sujetos con mejoras significativas en la valoración objetiva pero que no fue percibido por los terapeutas, situación que podría interpretarse como que el usuario sencillamente aprendió a jugar con la herramienta pero que no mejoró sus habilidades a criterio de los terapeutas, de manera que la comparabilidad entre la evaluación objetiva de la herramienta en términos de ECM entre el patrón establecido por el terapeuta y la producción oral del sujeto, no puede ser considerada como completa, única, o de referencia. Cuando un terapeuta evalúa la intensidad de la voz de un sujeto, son considerados muchos elementos como: fuerza, rasposidad, modo respiratorio, posición del tórax, etc. que pueden no afectar la capacidad del sujeto para seguir un patrón establecido. Una situación similar ocurrió con la actividad de tono, donde no todos los elementos para una correcta entonación pueden conocerse y tratarse con la actividad de tono propuesta por *PreLingua*, además, cuando un terapeuta evalúa soplo o vocales, ellos están evaluando la duración del soplo, direccionalidad, grado y modo de salivación, y todos los mecanismos implicados en la fonación vocálica así como las praxias de lengua, en donde se evalúa la habilidad del sujeto para seguir movimientos o trayectorias específicas de la lengua con o sin fonación; aunque estas dos actividades (soplo y vocales) se pueden trabajar con *PreLingua*, la herramienta no tiene en cuenta la misma cantidad de variables que utiliza un terapeuta experimentado y que razonablemente no son fácilmente medibles por medio de tecnologías del habla.

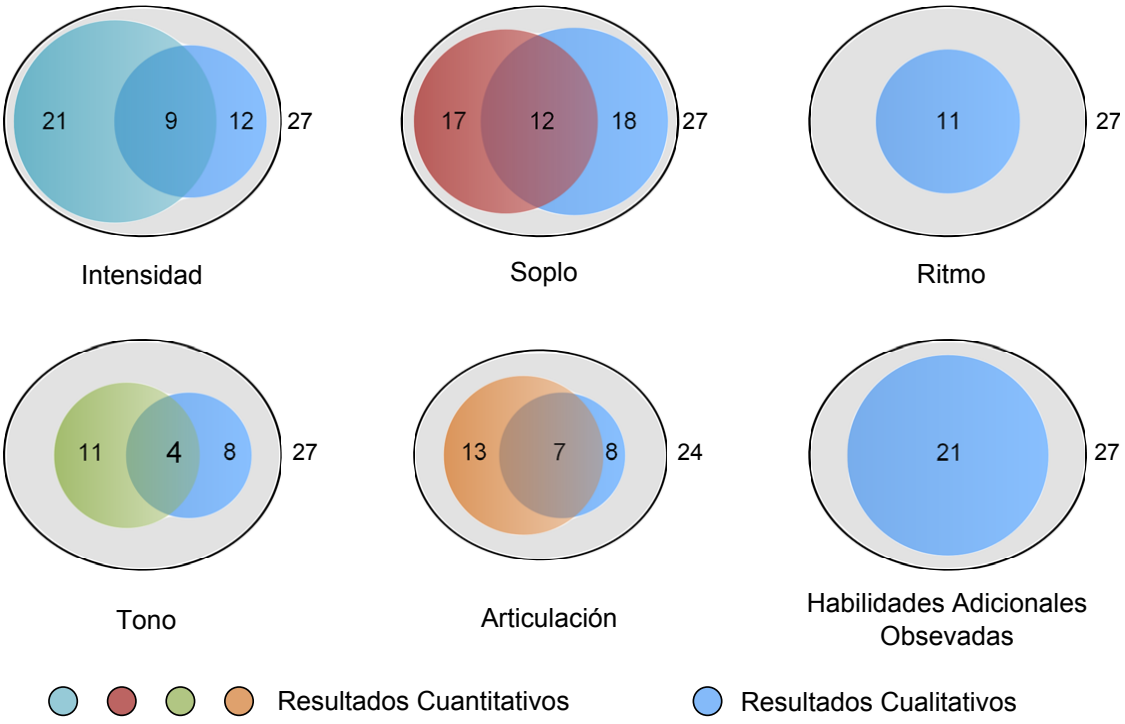


Figura 11.3: *Coincidencias en los resultados.*

Las variables medidas y trabajadas por *PreLingua* son aquellas en las que las tecnologías del habla brindan un apoyo científico y técnico, variables como la intensidad de la voz o el soplo medida en decibelios, el tiempo en segundos, o la frecuencia fundamental y formantes en Hercios, son tan solo una parte de toda la información necesaria por un terapeuta para evaluar y tratar los problemas de voz en un sujeto, *PreLingua*, puede ser considerada entonces como una herramienta de apoyo para la actividad diaria del terapeuta tanto para la terapia misma como para la evaluación, teniendo en cuenta que sus variables medidas apoyan el diagnóstico y permiten conocer como evoluciona el tratamiento sin cubrir por supuesto la totalidad de criterios necesarios en la historia clínica. Los resultados cualitativos del estudio indican que en las actividades como la intensidad, soplo y ritmo, la herramienta ofreció buenos resultados en un número apreciable de usuarios, mientras que para las actividades de tono y articulación, se obtuvieron resultados menos relevantes con respecto a la totalidad de la población siendo necesarias más sesiones de trabajo y estudios más extensos para lograr resultados más relevantes en un mayor número de usuarios.

Los terapeutas observaron que *PreLingua* tuvo un alto poder de motivación y de captura de atención en los sujetos que participaron del estudio (niños y adultos con diversas discapacidades). La especial interface de *PreLingua* diseñada con un entorno amigable, permite trabajar con diferentes usuarios independiente de sus edades, sin embargo, y teniendo en cuenta las continuas sugerencias de los terapeutas para la creación de una versión especial de la herramienta para adultos, quienes han perdido su habla debido a un traumatismo como los accidentes cerebro vasculares, habría que determinar si en estos adultos se obtendrían resultados satisfactorios al utilizar la herramienta, y se necesitaría de un re-diseño en las actividades para que la interface sea más adecuada al entorno adulto y

no precisamente al entorno infantil.

La herramienta ofrece grandes aportes e innovaciones respecto a herramientas existentes libres y de pago, por ejemplo, la normalización de formantes en función de las características del usuario, reduce la variabilidad inter locutor y posibilita trabajar con formantes infantiles durante su etapa de crecimiento. Otro valor agregado muy significativo de cara al usuario final es, la utilización de un avatar para trabajar la articulación vocálica en tiempo real convirtiéndose en una interface muy natural, motivante y adecuada para este tipo de población. Finalmente, y teniendo en cuenta la reducida investigación y desarrollo de este tipo de herramientas en español, *PreLingua* se esta convertido en una herramienta de referencia en el mundo hispanohablante no solo por sus innovaciones sino por estar diseñada para español y ser de libre distribución.

11.2 Impacto en la Comunidad Terapéutica

El proyecto COMUNICA [Rodríguez et al., 2008] del que hace parte *PreLingua* y otras herramientas libres para terapia del habla como Vocaliza y Cuéntame, distribuye sus herramientas a través del dominio www.vocaliza.es desde inicios de 2008, en este portal web, la única condición para descargar las herramientas es registrarse con una cuenta de correo válida. Desde su creación, se han registrado hasta Septiembre de 2010 un total de 7133 usuarios distribuidos principalmente en España y latinoamérica.

Figura 11.4: *Primeros 500 Usuarios Registrados.*

La Figura 11.4 muestra la distribución geográfica de los primeros 500 usuarios registrados donde se observa una mayor densidad en España y en menor medida en países latinos, pero con una tendencia a aumentar día a día gracias a los canales de difusión como conferencias, congresos, o la recomendación misma de quienes ya han utilizado la herramienta. Por otro

lado, debido a la gran demanda de soporte, en Febrero de 2009 se creó un canal en YouTube con vídeo-tutoriales de todas las herramientas del proyecto COMUNICA, para el caso del vídeo-tutorial de *PreLingua*, se han registrado 2834 reproducciones hasta Septiembre 15 de 2010.

La curva de reproducciones totales en el tiempo y su popularidad puede apreciarse en la Figura 11.5, allí se observa que incluso países cuya lengua materna no es el español, se han interesado por conocer la herramienta y posiblemente probarla. También se cuenta con la difusión en diferentes blogs dedicados a la logopedia y educación especial, hecha por quienes consideran la herramienta útil y con valor suficiente para difundirla.

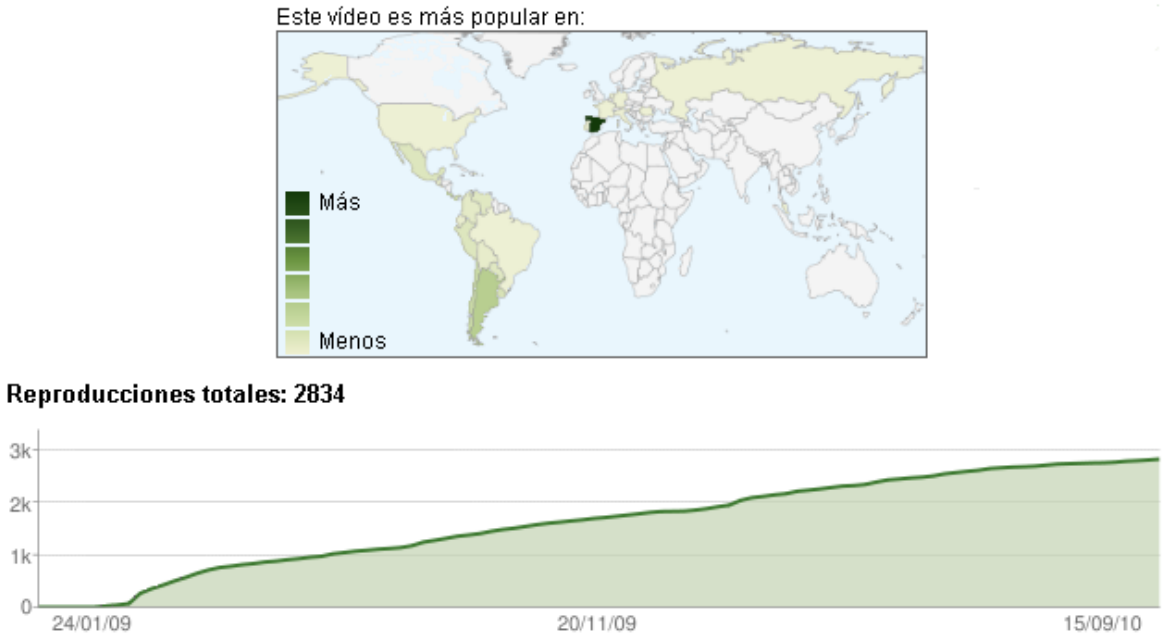


Figura 11.5: *Reproducciones y popularidad de PreLingua en YouTube.*

Las anteriores imágenes y cifras junto con diversas entrevistas en medios de comunicación, son indicadores del buen impacto que ha tenido la herramienta, y de que ésta responde en lo posible a las necesidades de la comunidad terapéutica. En el transcurso de la investigación y en especial en el último año, se han recibo numerosas aportaciones directas de logopedas, fonoaudiólogos, psicólogos y educadores especiales, con críticas constructivas y casos de éxito en la aplicación de la herramienta. Cabe destacar un caso de desmutización exitosa en un niño con craneosinostosis en Valencia (España), solución a una dislalia de la /s/ con la actividad de Soplo en Coruña, y una notoria mejoría en la movilidad de los órganos fonoarticulatorios, en el modo respiratorio, direccionalidad y fuerza del soplo, en un paciente de 14 años con discapacidad cognitiva, trastorno de comunicación y parálisis cerebral en Bogotá.

11.3 Otras Aplicaciones de la Tecnología

Con respecto a otras aplicaciones de la tecnología propuesta, VocalClick al ser el último desarrollo tiene por delante toda la fase de experimentación sobre casos reales y mejoras derivadas del mismo, las pocas experiencias de uso anticipan una gran potencial de beneficio para quienes padecen de impedimentos físicos, pero de igual manera, sugieren mejoras a nivel de adaptación y personalización para que aquellos usuarios que tienen serios problemas de articulación vocálica, puedan controlar la herramienta con los sonidos sonoros que ellos puedan emitir. Acerca del visualizador de vocales ViVo, su difusión ha sido mucho menor pero satisfactoria en el sentido de que ofrece información acústica de la voz en tiempo real, algunos de los usuarios han sugerido ampliar la herramienta para que permita grabar y cargar archivos de audio para su posterior análisis.

Acerca de la aplicación en RAH, el método propuesto permite la estimación robusta de la longitud del tracto vocal de un locutor dado trama a trama, lo que permite aplicar un factor de deformación que se actualiza y mejora entre más datos se obtengan. Esto supera el inconveniente de las técnicas tradicionales de VTLN en donde la misma señal de voz se procesa en varias etapas para obtener el mejor reconocimiento posible, además, el método propuesto trabaja sobre las características acústicas propias del locutor y no en aproximaciones estadísticas. Lo anterior hace que el método propuesto sea adecuado para utilizarlo en tareas de reconocimiento con diferentes tipos de locutores como adultos o niños, sin que la alta tonalidad en estos últimos afecte las estimaciones acústicas, lo que también representa una mejora frente a técnicas tradicionales. Profundizar en su robustez y capacidad de personalización haría viable su utilización en aplicaciones reales, en sistemas de reconocimiento automático del habla de pequeño y gran vocabulario.

Capítulo 12

Conclusiones y Líneas Futuras

Este último capítulo reúne las ideas y conclusiones derivadas de ésta investigación durante los cuatro años de trabajo, y plantea algunas líneas de actuación futuras sobre las que seguir trabajando y desarrollando, de cara siempre a dar robustez a las herramientas, ampliar su funcionalidad, y en último término, intentar mejorar la calidad de vida de las personas con discapacidad. La Sección 12.1 presenta un breve resumen de la investigación donde se recopilan los pasos realizados durante la misma, en la Sección 12.2 se describen los aportes de esta tesis y como se fueron cumpliendo los objetivos científicos y de desarrollo planteados inicialmente. La Sección 12.3 describe las potenciales líneas de trabajo futuras que toman como base las metas alcanzadas en ésta tesis y que definitivamente contribuirán a la mejora y ampliación de la tecnología propuesta, finalmente, en la a Sección 12.4 se hace una breve reseña de los méritos alcanzados en diferentes publicaciones, eventos y medios de comunicación.

12.1 Breve Resumen del Trabajo Realizado

Esta sección sintetiza el trabajo realizado durante la investigación resaltando las actividades o procesos relevantes que permitieron alcanzar la consecución de los objetivos propuestos, estos cuatro ítemes se relaciona con las primeras cuatro partes en la que esta divida la tesis.

- La primera parte de la investigación consistió en una contextualización de campo en áreas como la logopedia, la educación especial, y alteraciones de la voz infantil, también se conoció de cerca las necesidades y herramientas utilizadas por este sector profesional para tratar la voz, esto benefició en gran medida el inicio de la investigación y permitió trazar el rumbo de la misma. De igual manera se analizaron técnicas en procesado de señal de voz que servirían como punto de partida en la investigación.

- Con conceptos más claros sobre la voz infantil alterada y el mundo de la educación especial, se realizaron alianzas estratégicas para realizar la grabación de un corpus de voz infantil no alterada, y así disponer de una base de datos para aplicar las técnicas de procesado de señal de voz tradicionales y hacer estimaciones de los parámetros acústicos de este tipo de voz. La aplicación de dichas técnicas mostraron su debilidad

ante la voz infantil en donde la alta tonalidad afecta la estimación fiable de formantes, fue entonces necesario la aplicación de otras técnicas como el análisis homomórfico y el liftado para poder eliminar esta influencia y así poder estimar formantes de una manera más robusta. Una vez obtenidos formantes fiables en la voz infantil, se trato el problema de como reducir la alta variabilidad formántica entre diferentes locutores (niños, niñas, mujeres, hombres) debido a las diferentes condiciones geométricas del tracto vocal, una manera de abordar esta dificultad fue, la normalización de dichos formantes por medio de la longitud del tracto vocal obtenida en función de la talla y sexo del usuario, esta longitud se obtuvo de los formantes estimados de manera robusta en el mismo corpus y modelando el tracto vocal como un tubo homogéneo. Una vez obtenido este modelo fue posible normalizar los formantes estimados y reducir la alta variabilidad interlocutor lo que permitió iniciar el desarrollo de herramientas y las pruebas de las mismas.

- En la parte de aplicación y desarrollo, se dio comienzo a la creación de herramientas libres para logopedia y educación especial donde se aplicara la tecnología propuesta para tratar de manera robusta la voz infantil. Se desarrollo la herramienta *PreLingua* la cual permite trabajar alteraciones de la voz y aspectos de la comunicación prelingüística, por medio de juegos interactivos en tiempo real y con una interface adecuada para este tipo de población. Posteriormente a esta herramienta se integro ARTICULA, una aplicación para trabajar articulación vocálica en español en tiempo real con una interface muy natural, atractiva y entendible para los usuarios finales. La misma tecnología propuesta permitió la creación de ViVo y VocalClick, en donde el primero permite conocer los parámetros acústicos de la voz durante la emisión de vocales y, el segundo, emula los movimientos del ratón y los eventos de click únicamente utilizando los sonidos vocálicos del español.

  Otra aplicación de la tecnología fue en reconocimiento automático del habla, aquí, la estimación robusta de la longitud del tracto vocal del locutor a partir de los formantes presentes en las tramas sonoras, permitió estimar un factor de deformación que puede ser actualizado y mejorado entre más información formántica se tenga del locutor, con este método se supera el inconveniente de las técnicas tradicionales en donde se requiere de varias etapas de análisis para estimar el mejor factor de deformación frecuencial.

- Una vez desarrolladas las herramientas la siguiente etapa en la investigación y tal vez la más relevante en función de los objetivos trazados, fue la realización de un estudio aplicando *PreLingua* en casos reales de niños con alteraciones en su voz y con discapacidad. En este estudio participaron 27 niños de dos instituciones de educación especial en Colombia y España, se realizó durante 12 semanas en las cuales se registraron los valores de intensidad, tono, soplo, y articulación vocálica entregados por el sistema para ser evaluados objetivamente, y se realizó también una evaluación logopédica al inicio y al final del estudio para establecer de manera cualitativa si hubo diferencias en las habilidades de voz del usuario al final del estudio con respecto a las evaluadas al principio del mismo.

Los resultados cuantitativos obtenidos mostraron que el 29.6%, 25.9% y 22.2% de la población presento mejoras en las actividades de intensidad, soplo, y tono respectivamente (sig.≥95%), reduciendo el error entre los patrones de trabajo establecidos por el terapeuta y los patrones descritos por la voz del usuario entre las sesiones iniciales y finales. En la articulación vocálica, el 20.8% de la población presento una mejora en al menos una vocal y con el mismo nivel de significancia. Los resultados cualitativos mostraron por su parte, que el 44.4% de la población mostró mejoría en el control de la intensidad, el 66.6% mostraron una mejora en la duración del soplo, el 29.6% evolucionaron positivamente en el control del tono, y el 33.3% mejoraron en las práxias y habilidad para mover la lengua. El ritmo también se vio beneficiado con la aplicación de la herramienta en donde el 40.7% de la población mostró un mejor control sobre éste.

Como valor agregado del estudio, los terapeutas observaron ciertas habilidades adicionales que no se esperaban al comienzo del estudio, habilidades como: el aumento del tiempo de atención, el seguimiento de instrucciones, la direccionalidad del soplo, y algunas habilidades de socialización como la sana competencia, el respeto de turnos, y la autoexigencia también fueron observados hasta en un 77% de la población.

12.2 Aportes y Cumplimiento de Objetivos

Son varios los aportes generados por la investigación al término de estos cuatro años. Por una parte, en el conocimiento mismo de la voz infantil y de sus parámetros acústicos, ya que ésta información es escasa y muy poco documentada para idioma español, el conocer como cambian las frecuencias de resonancia para las vocales del español en niños y niñas en función de su crecimiento, es información útil que apoyará la continuación de ésta y otras investigaciones relacionada con la voz infantil. Por otra parte, y teniendo en cuenta que aunque esta investigación no propone formalmente técnicas nuevas en procesado de señal de voz, si propone como utilizarlas para realizar una estimación robusta de formantes en voz infantil eliminado la influencia de su alta tonalidad. También aporta una manera de reducir la alta variabilidad formántica entre locutores utilizando una estimación de la longitud del tracto vocal para su normalización, y cuya aplicación en reconocimiento automático del habla permite, obtener un factor de deformación apropiado para cada locutor en función de una medida acústica propia de éste y que se actualiza en tiempo real.

Tal vez el campo en donde la presente tesis ha aportado más cosas significativas es en el campo terapéutico de la logopedia y la educación especial, ya que como fruto de la misma se han creado herramientas libres en español y para español, para trabajar la comunicación pre-lingüística y tratar problemas de la voz alterada, también, la posibilidad de trabajar articulación vocálica en tiempo real con una interface adecuada y fácilmente entendible por la población infantil. Lo anterior en su conjunto, no existía en el entorno terapéutico hispanohablante hace cuatro años con éstas características y su buena aceptación por parte de éste gremio profesional, se refleja en el creciente número de usuarios y las continuas sugerencias y aportes de quienes las han utilizado.

Ahora, retomando los objetivos propuestos en la Sección 1.3, estos se plantearon como objetivos científicos y de desarrollo, los cuales se fueron cumpliendo en el transcurso de la

investigación tal y como se describe a continuación.

12.2.1 Cumplimiento de Objetivos Científicos

Los tres objetivos científicos planteados en esta tesis se han cumplido en diferentes grados:

- Desde una óptica muy técnica como la ingeniería, el acercamiento al mundo terapéutico de la logopedia y la educación especial fue vital para orientar adecuadamente ésta investigación y conocer las reales necesidades y dificultades de este gremio de cara a trabajar con población infantil con discapacidad. Así mismo, la adquisición del corpus de voz no alterada fue la materia prima inicial e imprescindible, para conocer y afrontar las dificultades técnicas encontradas en este tipo de voz.

- Una idea final muy clara surgida de ésta investigación es, que tratar la voz infantil es una tarea técnicamente difícil y que requiere de más investigación para la optimización de las técnicas existentes. Una vez analizado el corpus de voz, la mayor dificultad se encontró en la estimación de formantes fiables en voces con presencia de alta tonalidad, debido al solapamiento de la señal de pitch y sus armónicos con los formantes, situación que puede ser abordada utilizando técnicas como el análisis LPC y homomórfico para eliminar esta influencia. Una vez estimados los formantes de manera robusta, se pudo establecer como cambian estos en un individuo en función de su crecimiento y sexo, y hacer también una estimación de la longitud aproximada del tracto vocal para correlarla con su talla. Con esta información, se puede entonces predecir el comportamiento de los formantes de otros locutores con características físicas semejantes. La información generada en esta parte de la investigación, no solo ayudo de manera muy relevante a la concepción y diseño de las herramientas planteadas, sino que de hecho contribuye sustancialmente al conocimiento del comportamiento y evolución de la voz infantil. Por otra parte, la estimación de la longitud del tracto vocal a partir de los formantes estimados de manera robusta, permitió la creación de un método para estimar el factor de deformación de los ejes frecuenciales en sistemas de reconocimiento automático del habla de manera OnLine, y con resultados comparables a las técnicas de normalización del tracto vocal VTLN de manera OffLine.

- Finalmente, y aprovechando el conocimiento generado respecto a como varia la longitud del tracto vocal en un individuo en función de la talla y sexo, fue posible reducir la alta variabilidad formántica entre locutores aplicando una normalización con la longitud del tracto estimada, dejando los nuevos formantes normalizados en un espacio más homogéneo de trabajo. Esta normalización permitió el desarrollo de herramientas para articulación vocálica y emulación de eventos del ratón en tiempo real, teniendo en cuenta características propias de cada usuario dando a la herramienta un buen nivel de personalización.

12.2.2 Cumplimiento de Objetivos de Desarrollo

Respecto a los objetivos de desarrollo planteados, estos se cumplieron a cabalidad y posiblemente ya se han superado.

- Como fruto de ésta tesis se crearon varias aplicaciones para terapia de voz integradas en la herramienta *PreLingua*, la cual permite trabajar alteraciones de la voz y articulación vocálica en población infantil. La herramienta esta diseñada en español y ha sido descargada libremente de www.vocaliza.es por más de 7800 usuarios en España y Latinoamérica. También se desarrollaron herramientas como VocalClick para que usuarios con impedimentos físicos puedan controlar el puntero del ratón con sonidos vocálicos, y ViVo, la cual permite estudiar las características acústicas de la voz en la emisión de sonidos vocálicos sostenidos.

- Actualmente *PreLingua* se ha convertido en una buena alternativa para apoyar la labor diaria de terapeutas y profesionales de la voz y la educación especial en hispanoamérica, la herramienta cuenta con un buen grado de aceptación y difusión entre éstos profesionales al punto de ser utilizada no solo en terapia de voz sino también en otros campos relacionados con la educación especial y la logopedia, haciendo un poco más fácil su trabajo y favoreciendo la inclusión de esta población a los avances tecnológicos. En éste sentido, pequeñas contribuciones en la mejora de las competencias de comunicación de éstas personas es definitivamente mejorar su calidad de vida, y ayudar a que puedan comunicarse de una manera más eficiente.

12.3 Líneas Futuras

Durante algo más de dos años de los cuatro que han sido necesarios para concluir ésta tesis, se han publicado y presentado avances de la misma en diversos congresos, charlas, revistas, y entrevistas en medios de comunicación, lo que posiblemente no queda completamente reflejado en ésta memoria pero si muestra el potencial que ha tenido la investigación. Continuar investigando en ésta línea y abordando otros frentes de trabajo, contribuirá notoriamente en el impacto que las tecnologías del habla puedan tener sobre la discapacidad. A continuación, se citan cuatro posibles frentes de trabajo sin que ello signifique la no existencia de otras líneas de investigación y aplicaciones.

- Teniendo en cuenta la gran aceptación de la herramienta *PreLingua* y las continuas sugerencias por parte de los terapeutas respecto a su ampliación, será beneficioso para ellos y en especial para la población con voz alterada, poder seguir mejorando la herramienta en varios aspectos. Desde un punto de vista técnico y de investigación, una tarea por cumplir es dar más robustez al sistema frente a diferentes entornos de trabajo y condiciones de ruido, y poder implementar técnicas de tracking en tiempo real tanto en la estimación de pitch como de formantes teniendo en cuenta que en la voz infantil estos parámetros varían considerablemente.

- Con respecto a la parte de desarrollo, ampliar la sección de evaluación y la generación de reportes para los tiempos máximos de fonación y espiración, y la ampliación de actividades en ataque vocal pero considerando la intensidad del inicio de actividad glotal. Otra característica que ampliará la utilización de la herramienta en la

comunidad terapéutica será, el poder ejecutarla en otros sistemas operativos como linux o mac y que la herramienta esté disponible como una aplicación web, sin la necesidad de instalar el programa localmente.

- Con respecto a ARTICULA, al ser ésta la primera versión de una aplicación que trabaje articulación vocálica en español en tiempo real, la convierte en una herramienta con todo el potencial por delante como extender su aplicación a otros idiomas, dotarla de robustez, y empezar a trabajar en sonidos co-articulados con consonantes sin olvidar la extrema complejidad de ésta tarea. Ésta línea de investigación es la que más retos conlleva al tener inherente una altísima complejidad técnica, pero al mis tiempo, un alto potencial de ayuda en terapia de voz y habla especialmente en el tratamiento de dislalias como la /s/ y la /r/, y también su aplicación en herramientas de aprendizaje de segundo idioma (L2) en donde no existen alteraciones en la voz.

- La herramienta VocalClick abre también un gran abanico de posibilidades de aplicación en personas con impedimentos físicos, una promisoria línea de investigación consistirá en que la herramienta utilice las emisiones sonoras que un determinado usuario pueda producir, sin obligarlo a generar los sonidos vocálicos que solo identifica el sistema, esto permitirá que la herramienta la utilicen personar con serios problemas de articulación adicionales a sus impedimentos físicos, y también, que la utilicen indistintamente del idioma nativo de la persona, ya que la herramienta se adaptaría a las emisiones vocales que cada usuario pueda generar.

12.4 Indicios de Calidad

Durante el tiempo empleado en la realización de la presente tesis, varios han sido los méritos alcanzados. Algunos de ellos están directamente implicados en el trabajo presentado, mientras que otros se encuentran simplemente relacionados. En esta Sección se hace una breve reseña de ellos.

12.4.1 Ponencias en Congresos.

- W.-R. Rodríguez, O. Saz, A. Miguel y E. Lleida. *lugar: Vigo University, Spain, libro de actas: Proceedings of VI Jornadas en Tecnología del Habla, FALA 2010, mes: Noviembre, título: On Line Vocal Tract Length Estimation for Speaker Normalization in Speech Recognition, año : 2010.*

- W.-R. Rodríguez, O. Saz y E. Lleida. *lugar: Waseda University, Tokio - Japan, libro de actas: Proceedings of the 2010 Workshop on Second Language Studies: Adquisition, Learning, Education an Technology, Interspeech 2010 satellite workshop, mes: September, título: ARTICULA - A Tool for Spanish Vowel Training in Real Time, año : 2010.*

- W.-R. Rodríguez, y E. Lleida. *lugar: Wroxall Abbey Estates, United Kingdom, libro de actas: Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE), mes: September, título: Formant Estimation in Children's Speech and its Application for a Spanish Speech Therapy Tool, año : 2009.*

- W.-R. Rodríguez, O. Saz, C. Vaquero y E. Lleida. *lugar: San José, Costa Rica, libro de actas: Proceedings of the VIII Congreso Iberoamericano de Informática y Educación Especial (CIIEE), mes: August, título: Habilitación del Prelenguaje y del Lenguaje con Comunica, año: 2009.*

- W.-R. Rodríguez y E. Lleida. *lugar: Bilbao, Spain, libro de actas: Proceedings of the V Jornadas en Tecnologías del Habla, mes: November, título: PRELINGUA Una Herramienta para el Desarrollo del Pre-Lenguaje., año: 2008.*

- W.-R. Rodríguez, O. Saz, E. Lleida, C. Vaquero y A. Escartín. *lugar: Chania, Greece, libro de actas: Proceedings of the 2008 Workshop on Children, Computer and Interaction, mes: October, título: COMUNICA - Tools for Speech and Language Therapy, año: 2008.*

- W.-R. Rodríguez, C. Vaquero, O. Saz y E. Lleida. *lugar: Kuala Lumpur, Malaysia, libro de actas: Proceedings of the 4th Kuala Lumpur International Conference on Biomedical Engineering, mes: June, pages: 247–250, título: Speech Technology Applied to Children with Speech Disorders, año: 2008.*

- W.-R. Rodríguez, C. Vaquero, O. Saz y E. Lleida. *lugar: Isla Margarita, Venezuela, libro de actas: Proceedings of the 2007 Congreso Latinoamericano de Ingeniería Biomédica (CLAIB), mes: June, páginas: 1064–1067, título: Aplicación de las Tecnologías del Habla al Desarrollo del Prelenguaje y el Lenguaje, año: 2007.*

- O. Saz, E. Lleida y W.-R. Rodríguez. *lugar: Cambridge (MA), USA, libro de actas: Proceedings of the 2009 Workshop on Children, Computer and Interaction, mes: November, título: Avoiding Speaker Variability in Pronunciation Verification of Children Disordered Speech, año: 2009.*

- O. Saz, V. Rodríguez, E. Lleida, W.-R. Rodríguez y C. Vaquero. *lugar: Wroxall Abbey Estates, United Kingdom, libro de actas: Proceedings of the 2009 Workshop on Speech and Language Technologies in Education (SLaTE), mes: September, título: An Experience with a Spanish Second Language Learning Tool in a Multilingual Environment, año: 2009.*

- O. Saz, E. Lleida y W.-R. Rodríguez. *lugar: Madrid, Spain, libro de actas: Proceedings of the 3rd Advanced Voice Function Assessment International Workshop (AVFA09), mes: May, páginas: 129–132, título: Acoustic Phonetic Decoding for Assessment of Mispronunciations in Speakers with Cognitive Disorders, año: 2009.*

- O. Saz, W.-R. Rodríguez, E. Lleida, C. Vaquero y A. Escartín. *lugar: Bilbao, Spain, libro de actas: Proceedings of the V Jornadas en Tecnologías del Habla, mes: November, páginas: 37–40, título: COMUNICA - PLATAFORMA PARA EL DESARROLLO, DISTRIBUCIÓN Y EVALUACIÓN DE HERRAMIENTAS LOGOPÉDICAS ASISTIDAS POR ORDENADOR, año: 2008.*

- O. Saz, W.-R. Rodríguez, E. Lleida y C. Vaquero. *lugar: Chania, Greece, libro de actas: Proceedings of the 2008 Workshop on Children, Computer and Interaction, mes: October, título: A Novel Corpus of Children's Impaired Speech, año: 2008.*

- C. Vaquero, O. Saz, W.-R. Rodríguez y E. Lleida. *lugar: Rome, Italy, libro de actas: Proceedings of the LangTech2008, mes: February, páginas: 129–132, título: Human Language Technologies for Speech Therapy in Spanish Language, año: 2008.*

- C. Vaquero, O. Saz, E. Lleida y W.-R. Rodríguez. *lugar: Las Vegas (NV), USA, libro de actas: Proceedings of the 2008 International Conference on Acoustics, Speech and Signal Processing (ICASSP), mes: April, páginas: 4509-4512, título: E-Inclusion Technologies for the Speech Handicapped, año: 2008.*

12.4.2 Publicaciones en Revistas.

- O. Saz, S.-C. Yin, E. Lleida, R. Rose, W.-R. Rodríguez y C. Vaquero. *journal: Speech Communication, número: 10, páginas: 948 -967, título: Tools and Technologies for Computer-Aided Speech and Language Therapy, volumen: 51, año: 2009.*

- O. Saz, J. Simón, W.-R. Rodríguez, E. Lleida y C. Vaquero. *journal: EURASIP Journal on Advances in Signal Processing, páginas: Article ID 159234, 11 pages, título: Analysis of Acoustic Features in Speakers with Cognitive Disorders and Speech Impairments, volumen: Special Issue on Analysis and Signal Processing of Oesophageal and Pathological Voices, año: 2009.*

- O. Saz, W.-R. Rodríguez, C. Vaquero, A. Escartín, J.-M. Marcos y C. Canalís. *journal: Maremagum - Publicación Galega sobre os Trastornos do Espectro Autista, páginas: 131–138, título: Consideraciones en el Desarrollo de Herramientas Informáticas para Logopedia en Educación Especial, volumen: 13, año: 2009.*

12.4.3 Capítulos de Libro.

- O. Saz, E. Lleida, V. Rodríguez, W.-R. Rodríguez y C. Vaquero. *serie: Computer Synthesize Speech Technologies: Tools for Aiding Impairment, editor: J.-W. Mullenix and D.-E. Stern, nota: In press, editorial: IGI Global Publishing, título: The Use of Synthetic Speech in Language Learning Tools: Review and a Case Study, año: 2010.*

- O. Saz and V. Rodríguez and E. Lleida and W.-R. Rodríguez and C. Vaquero. *serie: Language Teaching: Techniques, Developments and Effectiveness, editor: F. Columbus, nota: In Press, editorial: Nova Science Publishers, título: The Use of Multimodal Tools for Pronunciation Training in Second Language Learning of Preadolescents, año: 2010.*

12.4.4 Otros Méritos.

- Segundo lugar en los Premios Accesibilidad Universal 2010, Fundación *DFA* Disminuidos Físicos de Aragón.

Parte VI

Apéndices

Apéndice A

Motor Gráfico Allegro

Un requerimiento importante a la hora de desarrollar herramientas informáticas libres para logopedia y discapacidad es, que su desarrollo sea en lo posible también con herramientas y librerías gratuitas. En este caso, el motor gráfico necesario para estas aplicaciones debía cumplir además ciertas características como: un relativo fácil uso, fácil integración a los algoritmos de tratamiento de voz, bajo coste computacional y, en lo posible, bien documentado. Se analizaron motores gráficos como: Irrlicht, OGRE, ALLEGRO y Fenix games, seleccionado finalmente ALLEGRO por cumplir las características mencionadas anteriormente en mayor proporción que los demás.

ALLEGRO es un acrónimo recursivo de Allegro Low LEvel Game ROutines [rutinas de bajo nivel para videojuegos], es una biblioteca para programación de videojuegos desarrollada en Lenguaje C, originalmente escrita por Shawn Hargreaves para la computadora Atari ST que más tarde adaptó y amplió para el compilador DJGPP. Actualmente funciona en plataformas como: DOS, Unix (Linux, FreeBSD, Irix, Solaris), Windows, QNX, BeOS y MacOS X. Allegro tiene varias funciones especialmente diseñadas para: gráficos, sonidos, entrada del usuario (teclado, ratón, joystick) y temporizadores, también tiene funciones matemáticas en punto fijo y coma flotante, algunas funciones 3D, y funciones para manejar ficheros[1].

Las gráficas en ALLEGRO se crean con primitivas geométricas de dibujo en donde se especifica el tipo de primitiva, longitud o radio, posición y color. Por ejemplo, en la figura A.1 pueden verse gráficas de líneas, rectángulos, triángulos y círculos de diferentes tamaños y colores (escala de grises). También se pueden cargar imágenes estáticas y utilizando subrutinas de animación propias de ALLEGRO, se obtienen animaciones y movimientos como la suma de elementos estáticos. Es así como con el conjunto de imágenes del dragón de la figura A.2, pueden generar el efecto de vuelo ya que el programa dibuja una imagen tras otra continuamente.

[1] http://alleg.sourceforge.net/index.es.html

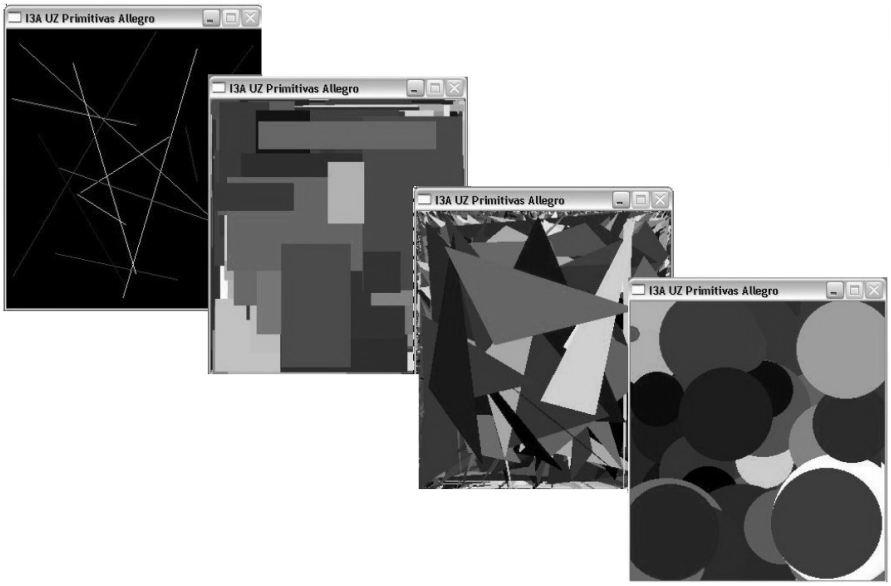


Figura A.1: *Primitivas de Dibujo en ALLEGRO.*

Figura A.2: *Imagenes Estáticas para Animación.*

Apéndice B

Evaluación Logopédica

Este apéndice muestra la evaluación logopédica aplicada antes y después de realizado el estudio a cada usuario participante. La evaluación logopédica de voz fue creada por profesionales del área de audición y lenguaje de la "Junta de Andalucía"[1] de la comunidad autónoma de Andalucía en España, su uso fue recomendado por los profesionales del colegio público de educación especial "Alborada" en Zaragoza donde se realizó el estudio, quienes participaron también en el diseño final.

La aplicación de esta evaluación permitirá comparar de manera cualitativa si el alumno presento alguna modificación o mejora en su voz y demás habilidades pre-lingüísticas después de haber trabajado con *PreLigua*, la evaluación abarca tópicos de la comunicación pre-lingüística y de la voz misma en escalas subjetivas según el aspecto evaluado, y en donde la experiencia del profesional que la aplica es fundamental. La evaluación abarca:

- Aspectos previos al lenguaje como: la capacidad de atención, la percepción visual, la percepción auditiva, y la imitación de sonidos.
- Cualidades de la voz como: el tipo de voz, la entonación y el ritmo.
- Una evaluación anatómica de: paladar, lengua, velo de paladar, frenillo, úvula, labios, dientes y amígdalas.
- La capacidad de relajación, y todo lo relativo a la respiración.
- La imitación de expresiones faciales y práxias de: lengua, labios, mejillas, y maxilares.
- La movilidad del velo del paladar.

[1] http://usuarios.multimania.es/maestrosayl/evaluacion-lenguaje.htm

EVALUACIÓN LOGOPÉDICA

Nombre y apellidos: _____ Sexo: _____
Edad: _____ Fecha de exploración: _____
Institución / Curso: _____Tutor/a: _____
Fecha de nacimiento: _____Observación/Dx: _____

ASPECTOS PREVIOS AL LENGUAJE

CAPACIDAD DE ATENCIÓN: (Comprobar si el niño/a es capaz de mantener la mirada o escuchar intencionalmente al menos unos instantes, ante la demanda o ante un estímulo).
PERCEPCIÓN VISUAL: (Verificar si puede seguir con la mirada un objeto animado o inanimado que se desplace, si se reconoce ante el espejo, si reconoce personas y objetos).
PERCEPCIÓN AUDITIVA: (Cerciorarse de que oye reaccionando ante ruidos y discrimina diferentes sonidos, voces,...).
IMITACIÓN: (Asegurarse de que es capaz de imitar sonidos, gestos y movimiento ante el modelo que se le proporcione).
RITMO: (Observar si el niño/a consigue seguir diferentes ritmos con diversos instrumentos o partes del cuerpo): Sigue las siguientes secuencias rítmicas: 0 0 0 Ritmo lento 000 Ritmo normal 000000 Ritmo rápido

VOZ: ENTONACIÓN Y RITMO

VOZ:	ENTONACIÓN:	RITMO:
Normal	Normal ….....	Normal ….....
Baja ….....	Monótona ….....	Rápido ….....
Fuerte ….....	Robótica ….....	Repeticiones ….....
Susurrada ….....		Entrecortado ….....
Disfónica ….....		
Nasal ….....		

ASPECTO ANATÓMICO

ANATOMÍA: Paladar …..... Velo…..... Úvula…..... Dientes…..... Lengua …..... Frenillo..... Labios..... Amígdalas.....
CAPACIDAD DE RALAJACIÓN: Relajación global …..... Relajación segmentaria: Cara..... Cuello.....

1 FUENTE: http://usuarios.multimania.es/maestrosayl/evaluacion_lenguaje.htm

Figura B.1: *Evaluación Logopédica hoja 1.*

RESPIRACIÓN:
Inspiración nasal …..... Espiración bucal ….....…..... Alternancia …..... Ritmo.....
(normal 11/15 insp/min aprox.) Retención del aire ..... Respiración ….....
Resp. Costal …..... Resp. Diafragmática ….....
Soplo: Intensidad …..... Duración …..... Direccionalidad ….....
Higiene nasal: Expulsa las mucosidades ….....

IMITACIÓN EXPRESIONES FACIALES:
Reír …..... Llorar …..... Comer …..... Dormir …..... Beber ….....
Sorpresa..... Miedo …..... Tristeza ..... . Alegría …..... Enfado .....

PRAXIAS DE LENGUA:
Sacar la lengua …..... Sacar y esconder deprisa ….....
Elevar hacia la nariz..... Bajarla hacia la barbilla ..….....
Llevarla hacia las comisuras: Izqda ..... Presionar las mejillas por dentro.....
Dcha …..... Hacer vibrar la lengua en el alveólo dental superior ….....
Llevar a los alveólos dentales: Sup …..
Inf …..

PRAXIAS DE LABIOS:
Llevarlos a la Izqda
Fruncir ...…..... Sonrisa con labios juntos .....
Sonrisa con labios separados ..... Esquema vocálico a-e-i-o-u .....
Llevarlos a la Izqda ..... Llevarlos a la Dcha .....
Hacer vibrar los labios ….....

PRAXIAS DE MEJILLAS:
Hinchar mejillas ..... Hincharlas alternativamente .....

VELO DEL PALADAR: Movilidad.
Ante bostezo ….....
Ante vocalización .....
Ante gárgaras .....

PRAXIAS MAXILARES:
Abrir ..... Cerrar ….....
Desplazamiento hacia: Izqda ..... Adelante – Atrás .....
Dcha ….....

2 FUENTE: http://usuarios.multimania.es/maestrosayl/evaluacion_lenguaje.htm

Figura B.2: *Evaluación Logopédica hoja 2.*

Bibliografía

[Arias and Estape, 2005] Arias, C. and Estape, M. (2005). *Disfonía Infantil*. Ed. Ars Medical, Barcelona, Spain.

[Bonet, 2009] Bonet, N. (2009). Rehabilitación de la voz infantil. *Audiología Práctica*, 1:10–13.

[Bosch, 2004] Bosch, L. (2004). *Evaluación Fonológica del Habla Infantil*. Ed. Masson, Barcelona, Spain.

[Cabero et al., 2008] Cabero, J., Córdoba, M., and Fernández, J. (2008). *Ordenador y Discapacidad*. Ed. CEPE, Sevilla, Spain.

[Deller et al., 1993] Deller, J., Proakis, J., and Hansen, J. (1993). *Discrete-Time Processing of Speech Signals*. MacMillan, New York, USA.

[E. Soria, 2003] E. Soria, M. Martínez, J. F. y. G. C. (2003). *Tratamiento Digital de Señales*. Prentice-Hall. Capítulo 1.

[Eide and Gish, 1996] Eide, E. and Gish, H. (1996). A parametric approach to vocal tract length normalization. In *Proceedings of ICASP-96*, pages 346–348, ,.

[El-Jaroudi and Makhoul, 1991] El-Jaroudi, A. and Makhoul, J. (1991). Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 39:411–423.

[Falcó et al., 2006] Falcó, J., Plaza, I., Marcos, J.-M., and Canalís, C. (2006). Dispositivo de orientación temporal: Ayuda técnica desarrollada a partir del acuerdo de colaboración entre el c.e.e. "alborada" y el centro politécnico superior de la universidad de zaragoza. In *Proceedings of the Jornadas Nacionales de Sistemas Aumentativos de Comunicación*, Zaragoza, Spain.

[Faúndez, 2000] Faúndez, M. (2000). *Tratamiento digital de voz e imágen*. Marcombo - Boixareu Editores. Capítulo 2.

[Gauvain and Lee, 1994] Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

[Goldstein, 1980] Goldstein, U. (1980). *An articulatory model for the vocal tracts of growing children*. PhD thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technolog.

[Gouvea and Stern, 1997] Gouvea, E.-B. and Stern, R.-M. (1997). Speaker normalization through formant-based warping of the frequency scale. In *Proceedings of Eurospeech*, pages 1139–1142, Rhodes, Greece.

[Green et al., 2003] Green, P., Carmichael, J., Hatzis, A., Enderby, P., Hawley, M., and M.Parker (2003). Automatic speech recognition with sparse training data for dysarthric speakers. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-Interspeech)*, Geneva, Switzerland.

[Hawley et al., 2003] Hawley, M., Enderby, P., Green, P., Brownsell, S., Hatzis, A., Parker, M., Carmichael, J., Cunningham, S., O'Neill, P., and Palmer, R. (2003). Stardust – speech training and recognition for dysarthric users of assistive technology. In *Proceedings of the 7th Conference of the Association for the Advancement of Assistive Technology in Europe, AAATE*, Dublin, Ireland.

[Hirano, 1981] Hirano, M. (1981). *Clinical Examination of Voice.* Springer, New York, USA.

[Hurtado and Soto, 2005] Hurtado, D. and Soto, F. (2005). *Tecnologías de Ayuda en Contextos Escolares.* Consejería de Educacioón y Cultura, Murcia, Spain.

[j. Benesty, 2008] j. Benesty, M. Mohan, Y. H. (2008). *Springer Handbook of speech processing.* Springer. Capítulo 2.

[J. Gurlekian and Eleta, 2000] J. Gurlekian, N. E. and Eleta, M. (2000). Caracterización articulatoria de los sonidos vocálicos del espanol de buenos aires mediate técnicas de resonancia magnética. Technical report, Laboratorio de Investigaciones Sensoriales.

[Kirlin, 1978] Kirlin, L. (1978). A posteriori estimation of vocal tract length. *IEEE Transactions on Acoustics, Speech and Signal Processing., VOL. ASSP-26, NO. 6.*

[Lee and Rose, 1998] Lee, L. and Rose, R. (1998). A frequency warping approach to speaker normalization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 6(1):49–60.

[Legetter and Woodland, 1995] Legetter, C.-J. and Woodland, P.-C. (1995). Maximum likelihood linear regression for speaker adaptation of the parameters of continous density hidden markov models. *Computer Speech and Language*, 9:171–185.

[Leonard, 1984] Leonard, R.-G. (1984). A database for speaker independent digit recognition. In *Proceedings of ICASSP'84*, pages 328–331, San Diego, CA (USA).

[Makhoul, 1975] Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings IEEE*, 63:561–580.

[Martínez et al., 2007] Martínez, B., Peguero, P., Ezpeleta, J., Falcó, J., Lleida, E., Mínguez, J., and Saz, O. (2007). Universidad y educaciÓn especial: Desarrollo y resultados de la colaboraciÓn entre el centro politÉcnico superior y el centro de educaciÓn especial "alborada". In *Proceedings of the III Congreso Nacional sobre Universidad y Discapacidad*, Zaragoza, Spain.

[Menéndez-Pidal et al., 1996] Menéndez-Pidal, X., Polikoff, J.-B., Peters, S.-M., Lorenzo, J., and Bunnell, H.-T. (1996). The nemours database of dysarthric speech. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP-Interspeech)*, Philadelphia (PA), USA.

[Molau et al., 2000] Molau, S., Kanthak, S., and Ney, H. (2000). Efficient vocal tract normalization in automatic speech recognition. In *Proceedings of ESSV*, pages –, Cottbus, Germany.

[Navarro-Mesa et al., 2005] Navarro-Mesa, J.-L., Quintana-Morales, P., Pérez-Castellano, I., and Espinosa-Yáñez, J. (2005). Oral corpus of the project hacro (help tool for the confidence of oral utterances). Technical report, Departmentof Signal and Communications, University of Las Palmasde Gran Canaria.

[Necioglu et al., 2000] Necioglu, B., Clements, M., and Barnwell, T. (2000). Unsupervised estimation of the human vocal tract length over sentende level uterance. *Acoustic Speech and Signal Processing*.

[Negre, 2005] Negre, F. (2005). Desarrollo de herramientas para la creación y utilización de tableros de comunicación en el ámbito de la educación especial [recurso electrónico]. Proyecto Fin de Carrera, Departamento de Ingeniería Informática y de Sistemas, , University of Zaragoza, Zaragoza, Spain. Dirigido por J. Ezpeleta.

[Negre et al., 2006] Negre, F., Ramos, D., Marcos, J.-M., and Canalís, C. (2006). Generador interactivo de tableros de comunicación: Ayuda técnica desarrollada a partir del acuerdo de colaboración entre el c.e.e. "alborada" y el centro politécnico superior de la universidad de zaragoza. In *Proceedings of the Jornadas Nacionales de Sistemas Aumentativos de Comunicación*, Zaragoza, Spain.

[Oppenheim and Schafer, 1968] Oppenheim, A. and Schafer, R. (1968). Homomorphic anañysis of speech. *IEEE Transaction on Audio and Electroacoustics*, 16:221–226.

[Ortega et al., 2004] Ortega, A., Sukno, F., Lleida, E., Miguel, A., and Buera, L. (2004). AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In *Proceedings of the 4th International Conferencece on Language Resources and Evaluation*, pages 763–767, Lisbon, Portugal.

[Paige and Zue, 1969] Paige, A. and Zue, V. (1969). Calculation of vocal tract length. *IEEE Transactions on Audio and Electroacoustics*.

[Proakis and Manolakis, 2007] Proakis, J. and Manolakis, D. (2007). *Tratamiento Digital de Señales*. Pearson Prentice Hall, Boston, USA.

[Puyuelo et al., 2004] Puyuelo, M., Rondal, J., and Wiig, E. (2004). *Evaluación del Lenguaje*. Ed. Masson, Barcelona, Spain.

[Quatieri, 1979] Quatieri, T. (1979). Minimum adn mixed phase speech analysis-sinthesis by adaotative homomorphic deconvolution. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 27:328–335.

[R. Schafer, 1978] R. Schafer, L. R. (1978). *Digital Processing of Speech Signals*. Prentice-Hall. Capítulo 4.

[Rabiner and Shafer, 2007] Rabiner, L. and Shafer, R. (2007). *Introduction to Digital Speech Processing*. The Essence of Knowledge, Santa Barbara CA, USA.

[Rodríguez et al., 2008] Rodríguez, W., Saz, O., Lleida, E., Vaquero, C., and Escartin, A. (2008). Comunica - tools for speech and language therapy. In *Workshop on Child Computer and Interaction, ICMI08*.

[Saz et al., 2008] Saz, O., Rodríguez, W.-R., Lleida, E., Vaquero, C., and Escartín, A. (2008). Comunica - plataforma para el desarrollo, distribución y evaluación de herramientas logopédicas asistidas por ordenador. In *Proceedings of V Jornadas en Tecnologías de Habla*, Bilbao, Spain.

[Schroeder, 1967] Schroeder, M. (1967). Determination of the geometry of the human vocal tract by acoustic measurements. *Journal of Acoustics Society America*, 41:1002–1010.

[Shahidur and Shimamura, 2005] Shahidur, M. and Shimamura, T. (2005). Formant frequency estimation of high-pitched speech by homomorphic prediction. *Acoustic Sci. and Tech.*, 26(6):502–510.

[Sánchez, 2002] Sánchez, R. (2002). *Ordenador y Discapacidad*. Ed. CEPE, Madrid, Spain.

[Stevens, 1998] Stevens, K. (1998). *Acoustic Phonetics*. The MIT Press, Cambridge, England.

[Tecumseh, 1997] Tecumseh, W. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journa of the Acoustical Society America*.

[Traunmuller and Eriksson, 1997] Traunmuller, H. and Eriksson, A. (1997). A method of measuring formant freqiencies at high fundamental frequencies. In *Proceedings of Eurospeech)*, –.

[Vallabha and Tuller, 2002] Vallabha, G. and Tuller, B. (2002). Systematic errors in the formant analysis of steady-state vowels. *Speech Communication*, 38:141–160.

[Vaquero, 2006] Vaquero, C. (2006). Reconocedor de comandos orales para eliminar barreras de comunicación y movilidad en personas con discapacidades motrices y de comunicación. Proyecto Fin de Carrera, Departamento de Ingeniería Electrónica y Comunicaciones, University of Zaragoza, Zaragoza, Spain. Dirigido por O. Saz (Ponente E. Lleida).

[Verhelst and Steenhaut, 1986] Verhelst, W. and Steenhaut, O. (1986). A new model for the short-time complex cepstrum of voiced speech. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 34:43–51.

[Vila, 2009] Vila, J. (2009). *Guia de Intervención Logopédica en la Disfonía Infantil*. Editorial Sintesis.

[Vorperian et al., 2005] Vorperian, H., Kent, R., Lindstrom, M., Kalina, C., Gentry, L., and Yandell, B. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *Journal of Acoustical Society America*, 117:338–350.

[Wakita, 1977] Wakita, H. (1977). Normalization of vowels by vocal tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech and Signal Processing., VOL. ASSP-25, NO. 2*.

[Watt and Fabricius, 2002] Watt, D. and Fabricius, A. (2002). Evaluation of a technique for improving the mapping of multiple speakers'vowel space in the f1 - f2 plane. *Leeds Working Papers in Linguistics and Phonetics*, 9:159–173.