Universidad Central "Marta Abreu" de Las Villas

Facultad de Ingeniería Eléctrica

Centro de Estudios de la Electrónica y la Tecnología de la Información

TRABAJO DE DIPLOMA

Software para el análisis del acento en voces patológicas.

Autor: *Publio Osleiky Garcés Pérez*

Tutor: *Ing. Héctor Arturo Kairuz Hernández - Díaz*

Santa Clara

2013

"Año del 55 Aniversario del Triunfo de la Revolución"

Universidad Central "Marta Abreu" de Las Villas

Facultad de Ingeniería Eléctrica

Centro de Estudios de la Electrónica y la Tecnología de la Información

TRABAJO DE DIPLOMA

Software para el análisis del acento en voces patológicas.

Autor: *Publio Osleiky Garcés Pérez*

E-mail: osleiky@uclv.edu.cu

Tutor: *Ing. Héctor Arturo Kairuz Hernández - Díaz*

Prof. Instructor Centro de Estudios de la Electrónica y la Tecnología de la Información. Facultad de Ing. Eléctrica. UCLV

E-mail: akairuz@uclv.edu.cu

Santa Clara

2013

"Año del 55 Aniversario del Triunfo de la Revolución"

Hago constar que el presente trabajo de diploma fue realizado en la Universidad Central "Marta Abreu" de Las Villas como parte de la culminación de estudios de la especialidad de Ingeniería en Biomédica, autorizando a que el mismo sea utilizado por la Institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos, ni publicados sin autorización de la Universidad.

Firma del Autor

Los abajo firmantes certificamos que el presente trabajo ha sido realizado según acuerdo de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

Firma del Autor

Firma del Jefe de Departamento
donde se defiende el trabajo

Firma del Responsable de
Información Científico-Técnica

PENSAMIENTO

Si desea el éxito no lo busque, limítese a hacer lo que ama y lo que cree. El éxito vendrá por añadidura.

David Frost

DEDICATORIA

Dedico este proyecto a todas las personas que contribuyeron de una forma u otra durante toda la carrera en especial a mis padres que siempre estuvieron pendiente de mí, a mi novia que siempre me apoyo, a todos mis tíos, a mi hermana, a mis primos, a mis abuelos en fin toda mi familia que siempre estuvo presente en mis actos y de una manera u otra se preocuparon por mí.

AGRADECIMIENTOS

Agradezco a todos los que se esforzaron en brindar los conocimientos adquiridos.

A mi tutor Arturo por su ayuda y paciencia en los momentos difíciles.

A mis padres y a mi novia los cuales me apoyaron durante toda la carrera y por ayudarme a convertirme en la persona que soy hoy, a toda mi familia por siempre apoyarme en los momentos más difíciles.

A mis amigos de universidad que los llevo en el alma: Yendys, Sandy, Sergio, José Daniel, Lisbel, Eddy, Roger, por estar siempre presente y recuerda que el que tiene fe en sí mismo no necesita que los demás crean en él.

TAREA TÉCNICA

- Revisión Bibliográfica y definición del Marco Teórico.
- Programación de un software para el análisis del acento en voces patológicas.
- Análisis del desempeño del software en una muestra representativa de pacientes disártricos. (Tabulación de resultados con los especialistas)
- Confección del informe final.

Firma del Autor

Firma del Tutor

RESUMEN

El país requiere de personal médico con recursos tecnológicos que ayuden a la exactitud y confiabilidad de los diagnósticos. Los estudios realizados sobre la fisiología de la voz y su procesamiento han sido de gran utilidad para reducir la subjetividad humana a la hora de diagnosticar algunas de las enfermedades de la voz. Esta investigación parte de la necesidad de que las consultas de Logopedia y Foniatría cuenten con herramientas tecnológicas que permitan una mayor objetividad y una valoración certera de las enfermedades.

En el presente trabajo se desarrolla un software para el análisis de la acentuación en voces patológicas orientado a la utilización en consultas de Logopedia y Foniatría. Permite la extracción de parámetros de la voz como la frecuencia fundamental, la intensidad y la duración de las sílabas. Para la extracción de la frecuencia fundamental se utiliza el algoritmo RAPT, mientras que el algoritmo de Wang se utiliza en la obtención de la intensidad y duración de las sílabas. El software brinda además la posibilidad de cargar una señal de voz, posee una interfaz gráfica amena y de fácil interpretación haciendo más sencilla la utilización de técnicas matemáticas del procesamiento de voz a usuarios del software que no estén familiarizados con las mismas.

Para la validación se toma una muestra representativa de una base de datos de las clínicas Mayo y se escogen 17 pacientes. En la misma aparece un párrafo leído por cada paciente (Grandfather Passage), del mismo se seleccionan las dos primeras oraciones y se procesan de forma independiente por los algoritmos propuestos.

Palabras claves: acentuación, frecuencia fundamental, intensidad, duración, software, interfaz gráfica.

Contenido

PENSAMIENTO 4
DEDICATORIA 5
AGRADECIMIENTOS 6
TAREA TÉCNICA 7
RESUMEN 8
INTRODUCCIÓN 10
Organización del informe 13
I. CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA 14
1.1 Fisiología de la producción del habla 14
1.1.1 Modelo general de producción de la voz 17
1.2 El acento y sus particularidades 19
1.2.1 La acentuación en los contornos de entonación 22
1.2.2 Desordenes de la acentuación y el ritmo 23
1.3 Desórdenes del habla 23
1.3.1 Asociación de medidas acústicas con aspectos prosódicos y paralingüísticos 24
1.4 Necesidad de mediciones objetivas 26
1.5 Consideraciones finales 27
II CAPÍTULO 2 MATERIALES y MÉTODOS 28
2.1 Detección de la frecuencia fundamental 28
2.1.1 Algoritmo seleccionado para detectar la frecuencia fundamental 28
2.1.2 Funcionamiento del RAPT 29
2.1.3 Modificaciones para la reducción de los errores de detección 31
2.2 Algoritmo para detectar intensidad 33
2.2.1 Algoritmo de Wang 33
2.2.2 Modificaciones introducidas al algoritmo de Wang 35
2.3 Materiales y métodos 36
2.3.1 El experimento 37
III. CAPÍTULO 3: RESULTADOS Y DISCUSIÓN. 38

3.1 Resultados de los algoritmos propuestos sobre la base de datos Aronson 38
3.2 Resultados de los algoritmos propuestos sobre un paciente de las consultas de Logopedia y Foniatría del país 42
3.3 La interfaz gráfica 44
CONCLUSIONES Y RECOMENDACIONES 50
Conclusiones 50
Recomendaciones 50
REFERENCIAS BIBLIOGRÁFICAS 51

INTRODUCCIÓN

El habla es un acto individual y voluntario que el hombre utiliza para materializar la lengua como modelo general y constante de todos los miembros de una colectividad lingüística, con el fin único de lograr comunicarse. Son estas manifestaciones de habla las que permiten la evolución del lenguaje; el cual ha sido objeto de estudio de importantes investigaciones en el campo de la medicina y el desarrollo de la tecnología. El conocimiento acerca de la fisiología de la voz y su procesamiento tiene gran utilidad en sistemas de diagnóstico y constituyen la base de numerosas terapias de rehabilitación.

La voz es una señal compleja, pues de ella pueden extraerse diferentes parámetros los cuales permiten su caracterización. Existen varias aplicaciones en el área de los sistemas digitales de comunicación de la voz, incluyendo el reconocimiento de locutores o palabras y la síntesis de habla por parte de los microprocesadores. [1] Ya en la década de los sesenta los estudios subjetivos del habla comienzan a ser formalizados, ejemplo de ello son los conocidos trabajos de las Clínicas Mayo, donde se localiza la lesión neurológica a partir

de las manifestaciones en el habla y se caracterizan los distintos tipos de disartria. En las décadas del 70 y 80 del siglo pasado empiezan a distinguirse otras esferas de aplicación para el tratamiento digital de la voz y en particular la detección de variables que la caracterizan, como el período fundamental, la frecuencia fundamental y los formantes. Se hacen más notables los esfuerzos realizados por la comunidad científica, específicamente de nuestro país, entre finales del siglo pasado y principios del presente [2].Para la segunda mitad del siglo XX se inicia junto con el desarrollo del procesamiento digital de señales, la era de mediciones acústicas. En el libro editado por Kent & Ball en el 2000, aparece un inventario de esta etapa, caracterizada por incontables medidas relacionadas con la calidad vocal y pocos intentos asociados a la articulación, la nasalidad y la prosodia, que también se ven afectadas y cuentan para el diagnóstico. Hoy día las medidas se concentran en buscar una integración hacia una medida objetiva de la inteligibilidad, lo cual requiere de medidas de articulación, nasalidad y prosodia, que en su mayoría, se calculan en segmentos cortos debido a la complejidad del análisis del habla fluida [3].

El personal médico necesita cada vez más de nuevos algoritmos y equipos especializados que permitan una mayor efectividad de la determinación de las diferentes enfermedades por sus síntomas. El desarrollo de la tecnología para el diagnóstico de las voces patológicas está siendo cada vez más utilizado por el personal de salud y el producto de su utilización cuenta con el aval de la comunidad científica moderna. El uso de técnicas de procesamiento digital de señales en la caracterización de señales de voz permite obtener medidas objetivas de los diferentes rasgos que caracterizan a estas señales biológicas y reduce la subjetividad introducida por la apreciación humana.

Para la clasificación de las distintas patologías, en las consultas de Logopedia y Foniatría las evaluaciones subjetivas han sido la herramienta principal [4], esto provoca que los resultados de los diagnósticos no presenten el 100 % de confiabilidad y de validez. A esto se suma que el personal no cuenta con los recursos especializados y tecnológicos para el proceso terapéutico.

El país debido a diferentes factores externos que lo afectan como las políticas internacionales y la globalización del conocimiento, se ve imposibilitado para lograr la obtención de herramientas que permitan la efectividad en los

diagnósticos de las enfermedades de la voz. Por tanto se hace necesario que las consultas de Logopedia y Foniatría cuenten con recursos tecnológicos que permitan luego de una valoración certera de las enfermedades la mejora de los pacientes en el proceso terapéutico; la realización de este trabajo va encaminado hacia esa necesidad y por tanto los resultados del mismo están dirigidos al desarrollo de un interfaz gráfico, que contribuya al análisis de la acentuación en voces patológicas y ofrecer una respuesta a la constante demanda de los especialistas para el diagnóstico.

Lo anterior conduce a la siguiente formulación del **problema**: ¿Cómo contribuir al diagnóstico y rehabilitación en las consultas de logopedas y foniatras, teniendo en cuenta los principios de la acentuación en voces patológicas y la aplicación de un algoritmo para el análisis de la misma? De este problema se originan las siguientes **interrogantes científicas**:

- ¿Cuál es la situación actual que presenta el desarrollo de soluciones para crear un método mediante el procesamiento digital de voz (PDV) en el diagnóstico y rehabilitación en las consultas de Logopedias y Foniatría dentro del campo del análisis de la acentuación en las voces patológicas?
- ¿Cómo elaborar un software basado en los principios de la acentuación en las voces patológicas?
- ¿Cómo evaluar la efectividad de la acentuación en las voces patológicas?

Basado en lo anteriormente expuesto y en busca de dar solución al problema planteado, el **objetivo general** de este trabajo es:

- Desarrollar un software basado en la programación de MATLAB para la visualización de parámetros de la voz utilizables en el análisis del acento de los pacientes de las consultas de Logopedia y Foniatría.

Y como **objetivos específicos**:

- Dilucidar los parámetros de la voz que estén relacionados con el acento.
- Analizar las características de dichos parámetros en voces patológicas.
- Crear una interfaz gráfica para el software.

- Validar los resultados.

Organización del informe

El informe de la investigación se estructura en introducción, capitulario, conclusiones, recomendaciones, referencias bibliográficas y anexos. En la introducción se deja definida la importancia, actualidad y necesidad del tema que se aborda. El Capítulo 1 se dedica a la caracterización del problema a partir de un análisis de la literatura (Fundamentación Teórica). Se comentan varios tipos de medidas para el análisis del habla y se realiza una selección de las que serán utilizadas posteriormente. El Capítulo 2 explica el diseño metodológico de la investigación y los algoritmos seleccionados con las modificaciones efectuadas para utilizarlos en el análisis de voces patológicas. En el Capítulo 3 se realiza la validación de la efectividad del algoritmo mediante un análisis objetivo de los resultados con los datos adquiridos y las evaluaciones subjetivas, realizadas por expertos, de la muestra de pacientes seleccionada.

I. CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

Resumen. El presente capítulo introduce el estudio de las disartrias, la necesidad e importancia de implementar un algoritmo para el análisis de la acentuación en voces patológicas de acuerdo a las evaluaciones clínicas subjetivas con medidas acústicas. Se discuten los compontes relacionados con el mecanismo de producción del habla, estableciendo así la necesidad de abordar de forma global y no fragmentada el problema de las medidas acústicas en voces patológicas.

1.1 Fisiología de la producción del habla

Para la producción del habla se necesitan tres grupos o sistemas de órganos: sistema respiratorio, sistema fonatorio y sistema articulatorio, (intervienen estructuras de naturaleza nerviosa, como los centros nerviosos específicos de control del habla situados en la corteza cerebral: Área de Brocca y Wernicke). Cada uno realiza una determinada función durante el proceso, el cual comienza con la generación de la energía suficiente (flujo de aire) en los pulmones, la modificación de ese flujo de aire en las cuerdas vocales, y su posterior perturbación por algunas constricciones y configuraciones de los órganos superiores. El habla, como señal acústica, se produce a partir de las ondas de presión que salen de la boca y las fosas nasales de una persona. Existen dos funciones mecánicas que permiten la producción del habla: la fonación y la articulación.

El conjunto de órganos que intervienen en la fonación Figura 1.1 pueden dividirse en tres grupos bien delimitados: [5]

- Cavidades infraglóticas (sistema sub-glotal) u órgano respiratorio
- Cavidad laríngea u órgano fonador
- Cavidades supraglóticas

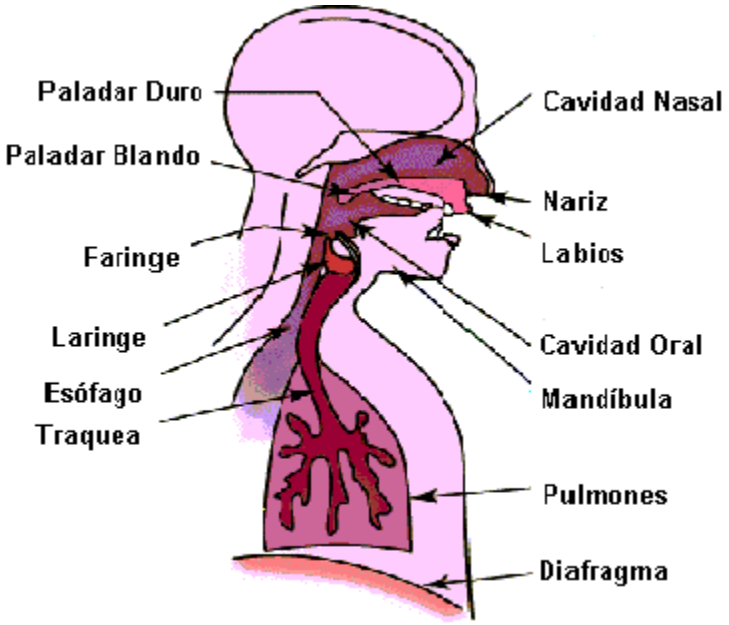

***Fig. 1.1**- Conjunto de órganos que intervienen en la Fonación.*

Cavidades infraglóticas

Las cavidades infraglóticas constan de los órganos propios de la respiración (pulmones, bronquios, y tráquea), que son la fuente de energía para la producción de voz. En el proceso de inspiración, los pulmones toman aire, bajando el diafragma y agrandando la cavidad torácica. En el momento de la fonación, la espiración, provocada por la contracción de los músculos intercostales y del diafragma, aporta la energía necesaria para generar la onda de presión acústica que atravesarán los órganos fonadores superiores.

Cavidad laríngea

La cavidad laríngea es la responsable de modificar el flujo de aire generado por los pulmones y convertirlo (o no), en una señal susceptible de excitar adecuadamente las posibles configuraciones de las cavidades supraglóticas. El último cartílago de la tráquea, el cricoides, forma la base de la laringe, cuyo principal órgano son las cuerdas vocales; estas son dos pares de repliegues compuestos de ligamentos y músculos. El par inferior son las llamadas cuerdas vocales verdaderas, que pueden juntarse o separarse mediante la acción de los músculos crico-aritenoides lateral y posterior, y que están protegidas en su parte anterior por el cartílago tiroides, el más importante de la laringe, abierto

por su parte posterior. Finalmente, la parte superior de la laringe está unida al hueso hioides.

Cavidades supraglóticas

Las cavidades supraglóticas están constituidas por la faringe, la cavidad bucal y la cavidad nasal. Su misión fundamental de cara a la fonación es perturbar adecuadamente el flujo de aire procedente de la laringe, para dar lugar finalmente a la señal acústica generada a la salida de la nariz y la boca. El volumen de la faringe laríngea puede ser modificado por los movimientos de la laringe, la lengua y la epiglotis mientras que el volumen de la faringe bucal se modifica por el movimiento de la lengua.

La faringe nasal y las restantes cavidades nasales forman, desde el punto de vista de su acción sobre el flujo de aire procedente de la faringe, un resonador que puede o no conectarse al resonador bucal mediante la acción del velo del paladar. Según el resonador nasal esté o no conectado, el sonido será nasal u oral, respectivamente. Si se hace una descripción de la cavidad bucal (esquematizada en la Figura 1.2), se pueden señalar las siguientes partes:

- Los labios en el extremo
- Los dientes
- La zona alveolar, entre los dientes y el paladar duro
- El paladar, en el que a su vez, y de forma simplificada, se puede distinguir el paladar duro y el paladar blando o velo.

La raíz de la lengua forma la pared frontal de la faringe laríngea, y sus movimientos le permiten modificar la sección de la cavidad bucal (movimiento vertical), adelantar o retrasar su posición frente a la de reposo (movimiento horizontal), así como poner en contacto su ápice o la parte trasera con alguna zona del paladar.

El movimiento de los labios también interviene en la articulación, pudiendo ser de apertura o cierre y de protuberancia, alargando en este último caso la cavidad bucal.

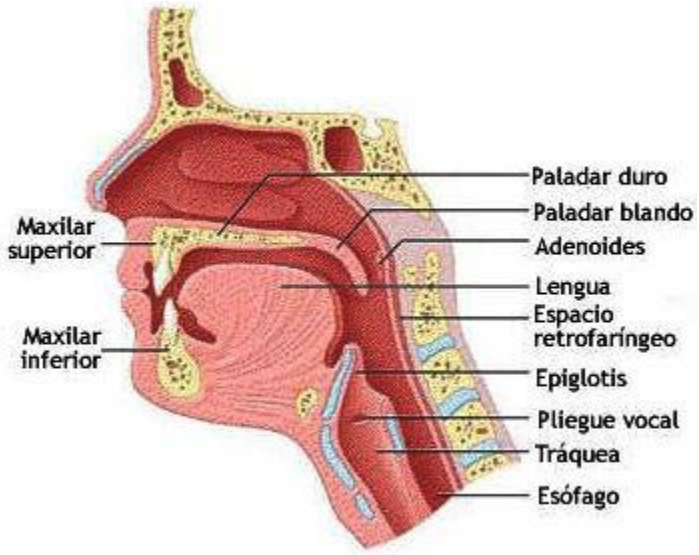

Fig. 1.2 *- Sección vertical de la cavidad oral.*

1.1.1 Modelo general de producción de la voz

El tracto vocal modelado se manifiesta como un filtro variable, cuyos parámetros varían en el tiempo en función de la acción consiente que se realiza al pronunciar una palabra. El filtro variable en el tiempo tiene dos posibles señales de entrada que dependerán del tipo de señal, sorda o sonora. Para señales sonoras la excitación será un tren de impulsos de frecuencia controlada, mientras que para las señales sordas la excitación será ruido aleatorio. La combinación de estas señales modela el funcionamiento de la glotis. El espectro de frecuencias de la señal de la voz puede obtenerse a partir del producto del espectro de la excitación por la respuesta en frecuencia del filtro (tracto vocal). El tracto vocal manifiesta un número muy grande de resonancias, sin embargo se consideran solo las primeras tres o cuatro que toman el nombre de formantes y cubren un rango de frecuencias entre 100 y 7500 Hz. Esto es debido a que las resonancias de alta frecuencia son atenuadas por la característica del tracto, el cual tiende a actuar como un filtro paso-bajo con una caída de aproximadamente - 12 dB por octava [6].

Desde un punto de vista mecano acústico, y además considerando que la voz es fisiológicamente normal, las vocales son los sonidos emitidos por la sola vibración de las cuerdas vocales sin ningún obstáculo entre la laringe y la abertura oral. Dicha vibración se genera por la vibración de las cuerdas vocales, donde interviene una fuente de energía constante en la forma de un

flujo de aire proveniente de los pulmones. Son siempre sonidos de carácter sonoro y por consiguiente de espectro discreto. Las consonantes, por el contrario, se emiten interponiendo alguna constricción formado por los elementos articulatorios. Los sonidos correspondientes a las consonantes pueden ser sonoros o no, dependiendo si las cuerdas vocales están vibrando. Funcionalmente, en el español las vocales pueden constituir palabras completas, no así las consonantes [6]. El reconocimiento de una consonante a través de su percepción depende esencialmente de la presencia de un cambio de frecuencias en sus elementos acústicos consecutivos, mientras que el de una vocal depende de la estabilidad en la frecuencia fundamental. Todos los cambios apreciables en la frecuencia de los formantes, excepto aquellos que aparecen en la unión de dos vocales contiguas, ayudan a la percepción de las consonantes; un cambio no apreciable en la frecuencia de las formantes contribuye a la percepción de las vocales. Todas las consonantes necesitan de otros parámetros (transiciones) para ser percibidas claramente. Cuando un espectrograma presenta cambios en las formantes colaboran a la identificación de las consonantes, mientras que los que presentan una relativa estabilidad en la frecuencia de las formantes, incluso durante un tiempo breve, se identifican como vocales. Los sonidos que funcionan como una consonante en un caso y como vocal en otro no son fonéticamente los mismos, tienen algo en común, pero son diferentes, se distinguen entre sí por un rasgo fonético marcado, es decir, uno se percibe por medio de un cambio en la frecuencia de los formantes, el otro no.

Este modelo (Figura 1.3) es una simplificación del proceso general de producción del habla. El mismo supone que las dos señales (señales de ruido y señales sonoras) pueden interactuar entre ellas sin sufrir ninguna dependencia. Si se consideran las señales sonoras que son filtradas en el tracto y las señales sordas que provienen del generador de ruido aleatorio, éste modelo supone igualdad en la extensión del filtro para ambas señales, por lo que lo hace más preciso y más confiable a la hora de ser considerado como tal. Por último, para este modelo digital de producción de la voz, es necesario considerar la radiación de salida por la importancia que significa la impedancia de radiación al momento de la emisión de la voz [7].

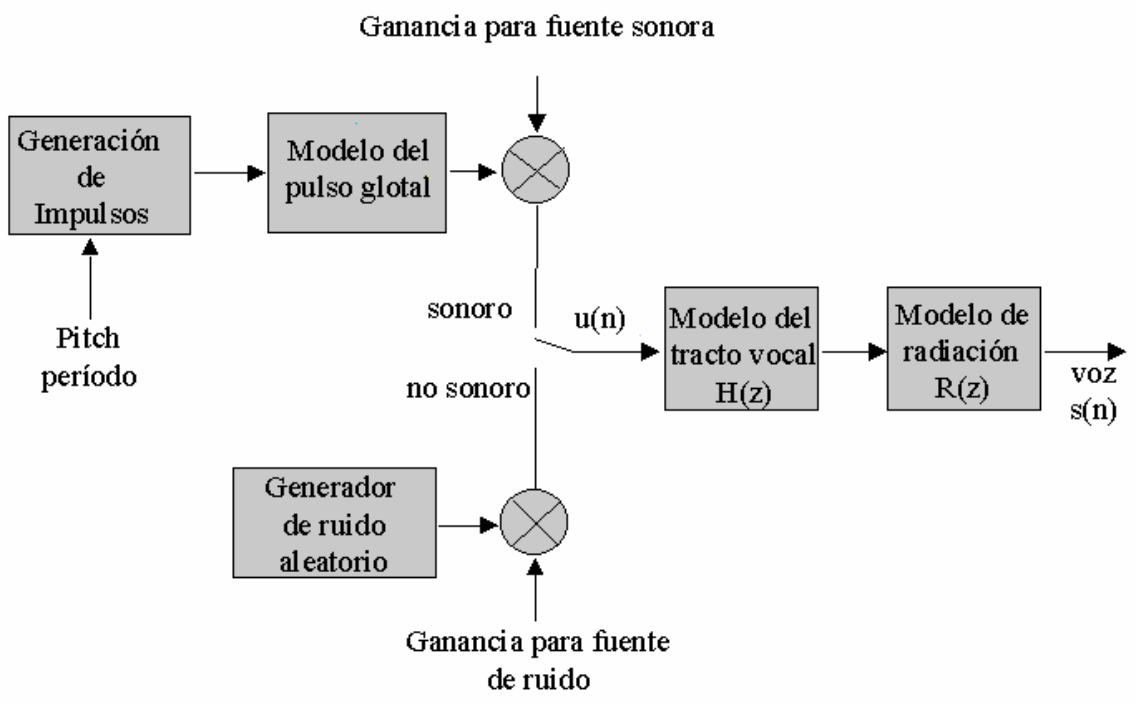

Fig. 1.3 *- Modelo general de producción de la voz.*

1.2 El acento y sus particularidades

La acentuación es una prominencia relativa colocada en la sílaba de una palabra o en una unidad mayor del habla como frases u oraciones. Por ejemplo, el sustantivo *content* (contenido en inglés) se pronuncia regularmente con mayor acentuación en la primera sílaba, mientras que el adjetivo *content* contiene mayor acentuación en la segunda [8].

Existe consenso respecto al patrón de la acentuación en palabras: pero hay desacuerdo respecto a las bases para su definición. Por ejemplo se puede esperar que las sílabas acentuadas tengan valores más elevados de frecuencia fundamental, mayor duración y mayor intensidad que las sílabas no acentuadas. Sin embargo hay excepciones a tales generalizaciones [9],[10].

En una frase u oración, la localización del mayor acento depende del significado implícito en la misma. El adjetivo *content* normalmente recibe mayor acentuación en la oración: "*I am content*", pero no en la oración: "*I am not content*", en la cual el mayor acento recae en *not.* En las Figuras 1.4 y 1.5 se

aprecian los espectrogramas de la palabra *impact* (impacto en inglés) con acento en la primera y la segunda sílaba respectivamente. El espectrograma indica la duración, intensidad y frecuencia fundamental de las vocales en cada sílaba. Los valores de estos parámetros en las sílabas acentuadas deben exceder a aquellos de las sílabas no acentuadas. Como se puede observar en las Figuras 1.4 y 1.5 no siempre ocurre así. La evidencia contradice la teoría: primero *im* no es más prolongado que *pact* en el niño ni en el adulto; 2 *im* no es mayor en intensidad que *pact* en el adulto; 3 *pact* no posee una F0 mayor que *im* en la pronunciación del niño.

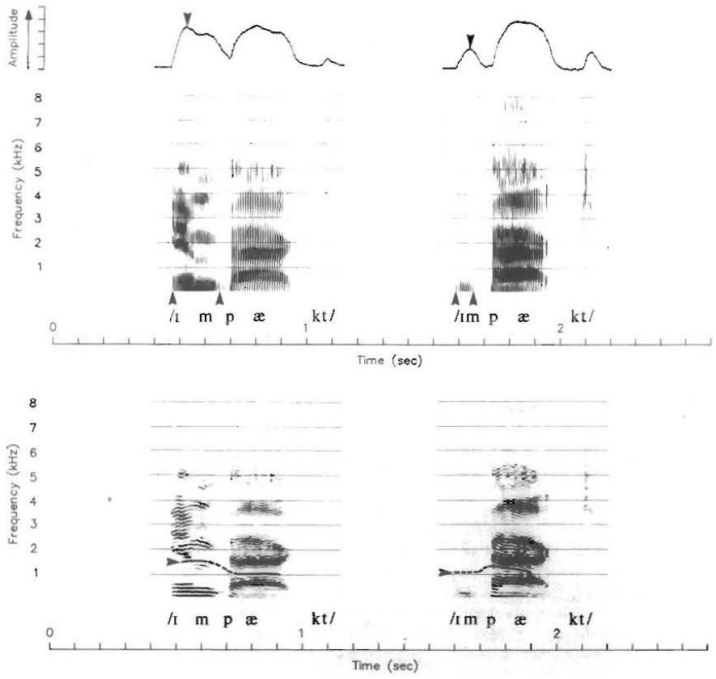

Fig. 1.4 - *Espectrograma y frecuencia fundamental de la palabra ´impact´ producida por un adulto.*

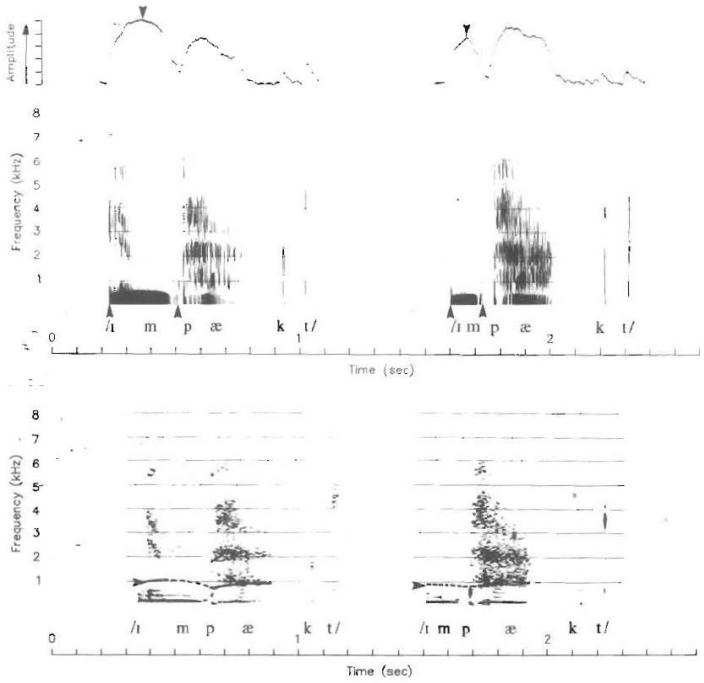

Fig. 1.5 - *Espectrograma y frecuencia fundamental de la palabra ´impact´ producida por un niño.*

A pesar de estas contradicciones la evidencia acústica muestra que no hay dificultad en diferenciar el patrón de acentuación. Esto demuestra que los oyentes de alguna manera permiten para este tipo de distinciones utilizar umbrales relativos en vez de absolutos para diferenciar los parámetros de la percepción de la acentuación [11].

Como se indica en las Figuras 1.4 y 1.5, las siguientes características pueden diferenciar el sustantivo de la forma verbal de *impact*:

- La primera sílaba es mayor en el sustantivo que en el verbo.
- La intensidad de la primera sílaba es mayor en el sustantivo que en el verbo.
- La frecuencia fundamental decrece entre la primera y la segunda sílaba del sustantivo y se incrementa entre la primera y la segunda sílaba del verbo para el adulto.

Además de la falta de evidencia acústica para diferenciar la acentuación, hay desacuerdo entre los especialistas con respecto del número de niveles de acentuación que se deben utilizar. Algunos utilizan solo dos niveles: presencia o ausencia de acentuación, [10] mientras que otros describen tres niveles [12],[13] o 4 niveles [14]. Algunos prefieren poner marcas de acentuación antes de las sílabas; otros lo hacen en la vocal de la sílaba acentuada.

La acentuación de las palabras tiene dos funciones:

1- Diferenciar pares de sílabas

2- Proveer un patrón apropiado de pronunciación de las palabras.

Colocar el acento en la sílaba incorrecta o no acentuarlas resulta un patrón anormal de pronunciación o extranjero. Se propone que el factor fonológico más influyente en la colocación de la acentuación sea el peso o la fuerza de la sílaba[15], [16], [17]. Sílabas ligeras o débiles son aquellas que incluyen vocales cortas no terminadas en consonantes. Sílabas fuertes son aquellas que:

1- Incluyen vocales largas o diptongos

2- Incluyen vocales cortas terminadas en consonantes.

1.2.1 La acentuación en los contornos de entonación

De acuerdo a Crystal, D. [18], la acentuación de una oración pronunciada por un adulto está en la última unidad léxica, en el 90% de los casos. Esto puede cambiar modificando el énfasis de las siguientes maneras:

- La acentuación se puede colocar en una palabra como objeto de clasificación o énfasis de la misma en el contexto de la oración.
- Para aclarar o refutar la acción del sujeto.
- Para contrastar intensiones con una tercera persona.

Dentro del contorno de entonación las unidades tonales de ciertas palabras se encuentran acentuadas. Estas pueden ser sustantivos, verbos, adjetivos y adverbios. Usualmente las no acentuadas son pronombres personales, verbos auxiliares, preposiciones, conjunciones y artículos.

Ritmo: es el patrón de acentuación en el habla y los intervalos de duración entre acentuaciones. En una oración como "*I am walking to the store*" algunas sílabas tienen acentuación y otras no, lo cual es usual en los patrones de acentuación. Como se mencionó anteriormente, las vocales en sílabas no acentuadas usualmente son reducidas y esto ayuda a describir el patrón de acentuación para el inglés como acentuado en tiempo. Otros lenguajes como el francés se describe como cronometrados en sílabas porque cada una de ellas es aproximadamente igual en duración y existen pocas vocales reducidas. La

implicación del término acentuado en tiempo es que el ritmo involucra el control de la localización de la acentuación y el intervalo de tiempo entre las mismas.

1.2.2 Desordenes de la acentuación y el ritmo

Hay dos problemas relacionados al uso correcto de la acentuación y en el ritmo.

- Localización incorrecta del acento tónico en el contorno
- Falta de diferencia entre sílabas acentuadas y no acentuadas.

La incorrecta localización se considera un problema de la mala acentuación de la oración y la falta de diferenciación puede considerarse un problema de ritmo. La incorrecta localización de la acentuación puede variar en su naturaleza.

El acento tónico refiere a la mayor intensidad con que se pronuncia una sílaba y la elevación en el tono. Este puede desplazar hacia adelante y puede recaer en una sílaba no acentuada.

La pérdida de diferenciación entre sílabas acentuadas y no acentuadas es característico de individuos con ataxia [19][Kent and Rosenbek 1982]; no solo se reduce las diferencias en duración y tiempo entre sílabas continuas acentuadas y no acentuadas, sino que también, se incrementan los intervalos intersílabicos. Estas alteraciones del patrón cambian el cronometraje típico de la acentuación.

1.3 Desórdenes del habla

Para producir una señal de voz estable y adecuada es necesaria la coordinación en el acoplamiento de los diferentes subsistemas involucrados en su producción. En ocasiones existen afecciones que provocan alteraciones en el control sobre movimientos musculares y por consiguiente en la producción del habla, las que se agrupan bajo el término de Trastornos Motores del Lenguaje (TML) y son consecuencia de una lesión en el Sistema Nervioso (SN). Se conocen dos clases de TML: la apraxia y la disartria del lenguaje.

La disartria es una alteración del habla que tiene como base un trastorno neurológico, y que generalmente se acompaña de alteraciones de los movimientos biológicos de los órganos buco faríngeos y en ocasiones de incoordinación fono respiratorio. Es fundamentalmente un problema del habla,

se acompaña en la mayoría de los casos de problemas concomitantes de voz, dada la íntima relación anatómica y funcional que tienen estos dos niveles de comunicación y de la participación frecuente de la misma inervación para ambas funciones. (Según Darley, Aronsm y Brown) [20]

Las disartrias son aquellas perturbaciones del habla causadas por parálisis, debilidad o incoordinación de la musculatura del habla de origen neurológico que ocasiona trastorno motor sobre la respiración, fonación, resonancia, articulación de la palabra y prosodia. Por su parte Prater [21] la define como alteraciones de la inervación motora de los músculos del mecanismo vocal que se caracterizan por trastornos de la articulación, de la fonación, la resonancia y la respiración que traen como consecuencia anormalidades neuromusculares como trastornos de la fuerza muscular, o del tono o de excesivos movimientos involuntarios.

La apraxia consiste en la disminución de la capacidad para ejecutar voluntariamente los movimientos adecuados en la articulación del habla, siempre que no exista parálisis, debilidad o descoordinación de la musculatura en el proceso de producción de la voz. Puede ser pura o asociarse a una afasia de Broca. En contraste con la disartria no hay distorsión del sonido del lenguaje sino más bien sustituciones fonéticas [22]. Casi todos los pacientes disártricos están asociados con algún problema de prosodia en acentuación, contornos de entonación, y/o razón del habla y ritmo. Solo recientemente se ha considerado la prosodia como un problema de primera línea en el tratamiento, dado que históricamente los clínicos consideraban la prosodia como un asunto a tratar luego de haber remediado otros trastornos en la producción del habla [23].

1.3.1 Asociación de medidas acústicas con aspectos prosódicos y paralingüísticos

Kent et al, 1999 [24] sugiere explorar el área sobre la envolvente de la energía, duración de los sonidos y las pausas, fragmentación y variaciones espectrales, así como varias medidas sobre la frecuencia fundamental, para correlacionar con la prosodia. Las principales correlaciones acústicas son: la ruptura del patrón temporal, que afecta la envolvente de la energía, el contorno de F0 y las

propiedades espectrales de las vocales y las consonantes. Los síntomas de disartria son: repetición de sílabas o palabras, prolongación de sonidos, bloques de silencios o duda, múltiples oclusiones y liberaciones. La prosodia evalúa las variaciones del tono de los fonemas sonoros en base al seguimiento de su frecuencia fundamental. Se distinguen dentro de ella dos patrones lingüísticos de interés: la acentuación y la entonación. La acentuación es un rasgo que permite poner en relieve un fonema para diferenciarlo de otras unidades del mismo nivel dentro de un nivel morfológico (palabra). El concepto de entonación es similar al de acentuación sólo que representa expresión a nivel de oraciones, siendo además mucho más clara la variación en la frecuencia fundamental. Las mismas no pueden ser descritas dentro de un sonido, ellas abarcan de una sílaba a un segmento en el habla. El habla es más útil para detectar problemas en prosodia en pacientes disártricos; pero el problema con el habla fluida es la falta de control sobre propiedades de la pronunciación como: longitud, estructura sintáctica y composición fonética. Las medidas que aparecen referenciadas en dicho artículo son análisis basado en consideraciones de unidades de tonos y regulaciones de F0, donde se ha reportado que individuos con disartria severa tienen unidades cortas de tono y alta media de F0 con respecto a pacientes con disartria leve o individuos sanos. Por otra parte, pacientes con disartrias leves tienen menos variaciones de F0 que pacientes con disartria severa o grupos de control. Leuschel & Docherty (1996) [25] realizaron una aproximación estocástica multidimensional para estudiar prosodia en disartria; específicamente en lectura y habla fluida, con mejor resultado en habla fluida. Para ello utilizaron las siguientes variables: razón de articulación, duración media de la pausa, número de pausas, razón de tiempo articulación/pausa, longitud media de la alocución, duración media de la vocal no acentuada, porcentaje de vocales no acentuadas, intervalo de intensidad, envolvente de intensidad, F0 medio, intervalo de F0, envolvente de F0 y la variación de F0 entre vocales. Consideran el procesamiento matemático insuficiente para las medidas de prosodia en voces patológicas en gran parte por lo antes mencionado de que no existe una medida cuantitativa para la prosodia.

1.4 Necesidad de mediciones objetivas

El proceso de la rehabilitación en pacientes con problemas del habla y la articulación de las palabras, constituye un fenómeno de gran interés en las consultas de Logopedia y Foniatría de los centros de salud en el país. Las evaluaciones subjetivas han sido la herramienta principal para la clasificación, descripción y seguimiento de problemas del habla, por este motivo surgen las dudas sobre la confiabilidad y la validez de las evaluaciones subjetivas, especialmente por el hecho de que el proceso terapéutico y su seguimiento se lleven a cabo por técnicos y personal médico con escasos recursos especializados. Esto hace que los mismos sean ineficientes y den poca información de la evolución de dichas patologías [26].

La respuesta al problema de la confiabilidad de las medidas subjetivas descansa en la creación de medidas objetivas que surgen a partir del análisis acústico. Este se puede clasificar como: análisis en el dominio del tiempo, en el dominio de la frecuencia, y en el dominio tiempo-frecuencia de la señal de voz. Por este motivo existe una necesidad importante de incorporar al proceso de evaluación de pacientes con problemas del habla algunas medidas objetivas que cuantifique de manera real el desempeño del individuo en un proceso de rehabilitación o de diagnóstico. Las medidas objetivas para la calidad vocal, articulación y prosodia son componentes esenciales para calcular un índice de inteligibilidad. Hay gran cantidad de trabajos reportados en medidas de calidad vocal, a pesar de las dificultades inherentes al procesamiento digital y la variabilidad de la señal, que se amplifica por los desórdenes del habla; por otro lado hay una gran carencia en cuanto a medidas objetivas relacionadas con la articulación y la prosodia debido a la dificultad que acarrea analizar unidades más complejas del lenguaje junto con las dificultades antes mencionadas [27].

Sin embargo, dado que la voz es el resultado de la interacción de múltiples procesos fisiológicos que no se pueden capturar con una sola técnica de medición, no se puede obviar la necesidad de las evaluaciones subjetivas. Así la mejor opción es lograr una combinación entre las medidas subjetivas y objetivas, con el propósito de conseguir un mayor aporte en los procesos de diagnóstico y rehabilitación en las consultas de Logopedia y Foniatría.

1.5 Consideraciones finales

La prosodia es en los últimos tiempos objeto de estudio en numerosas investigaciones, ya que históricamente se dio prioridad a otras medidas acústicas de la producción del habla. La entonación como componente de la prosodia aporta información relacionada al estado de la patología, por lo cual, es menester del presente trabajo continuar la investigación de las medidas relacionadas con aspectos suprasegmentales del habla disartria, en este caso la acentuación ya que aporta información vital relacionada con la inteligibilidad del habla.

Los algoritmos desarrollados para el análisis del habla no han aportado una solución definitiva en el diagnóstico de los pacientes disártricos. Esto se debe a las complicaciones mencionadas en el epígrafe 1.3. Por lo cual se propone realizar modificaciones en algoritmos bien referenciados en la literatura científica con el propósito de mejorar su rendimiento en la estimación de la acentuación en voces patológicas. Con el objetivo final de crear una herramienta que ayude a los especialistas en sus evaluaciones diagnósticas y seguimientos al paciente durante el tratamiento.

II CAPÍTULO 2 MATERIALES y MÉTODOS

Resumen: En el desarrollo de este capítulo se definen los materiales y métodos utilizados para la confección del presente trabajo. Se realiza una descripción de los algoritmos seleccionados, se llevan a cabo una serie de modificaciones de los mismos y se describen las diferentes funciones implementadas para llegar a los resultados esperados. Se explican las características y composición de la muestra de pacientes, así como las distintas herramientas utilizadas.

2.1 Detección de la frecuencia fundamental

Para obtener una representación acústica de la evolución temporal de la frecuencia fundamental (F0) a lo largo de un enunciado, se emplean normalmente algoritmos de detección de la F0 que actúan directamente sobre la señal en el tiempo para detectar la periodicidad de la misma y la longitud del período. A continuación se describe el algoritmo empleado y los procedimientos que permiten eliminar algunos de los errores inherentes a este proceso.

2.1.1 Algoritmo seleccionado para detectar la frecuencia fundamental

RAPT (¨*A Robust Algorithm for Pitch Tracking*¨ *en inglés*) es un algoritmo robusto para el estimado de F0. La principal meta de este algoritmo es obtener, con el menor costo computacional posible, un estimado de la F0 más preciso. Este detector muestra que varias mejorías en la eficiencia son incorporadas reduciendo la complejidad computacional y logrando la precisión buscada.

RAPT opera de forma continua, está diseñado para trabajar a cualquier frecuencia de muestreo (Fs) y también acepta cualquier variación en el tamaño de la ventana. Esto se puede aplicar en condiciones donde, además del segmento del habla que se esté utilizando, pueden haber otras personas hablando y diversas condiciones de ruido. Este algoritmo permite ajustar parámetros para establecer una relación de compromiso entre la velocidad de la detección de la F0 y la precisión para diferentes tipos de voces y condiciones de grabación.

Después del análisis de una voz patológica se pueden observar algunas de las características siguientes: La F0 puede presentar cambios abruptos como

sustituciones de F0 por múltiplos y submúltiplos de F0. El espectro de corta duración presenta marcadas diferencias para las ventanas donde hay sonoridad y para las ventanas donde hay silencios. La amplitud de la señal se incrementa dentro de las ventanas donde hay sonoridad y disminuye donde hay silencios.

RAPT usa la NCCF (Función de Correlación-Cruzada Normalizada), esta función, independientemente de los cambios rápidos que puedan ocurrir en la amplitud de las muestras, ve las formas y los períodos sucesivos como similares. Las propiedades que presenta la NCCF son independientes de las muestras que se analizan. Posee para la parte sonora máxima del segmento en los intervalos de retraso una amplitud comparable que se corresponde a los múltiplos enteros del T0. Y para los segmentos silenciosos, los máximos aparecen donde no hay retrasos. La estimación precisa de la F0 con esta función se realiza efectivamente aumentando la Fs, y luego realizando un proceso de relocalización de los picos en la tasa de muestreo más alta.

2.1.2 Funcionamiento del RAPT

Este algoritmo posee dos variantes para muestrear los datos. Puede muestrearlos usando la tasa de muestreo original o puede usar también una tasa de muestreo reducida. Luego utiliza la NCCF para registrar periódicamente una señal con baja tasa de muestreo en el rango de interés de F0 para todos los retrasos de la señal. Después guarda las posiciones de los máximos locales detectados con baja tasa de muestreo. También busca los vecinos de los picos ya encontrados anteriormente pero con una tasa de muestreo alta. Busca los máximos locales de nuevo para obtener una posición más precisa. Los picos obtenidos con la alta tasa de muestreo son los candidatos para detectar la curva de la F0. Se auxilia de la Programación Dinámica para seleccionar el conjunto de picos de la NCCF o las hipótesis de silencios que mejor equivalencia tengan con las características mencionadas anteriormente.

La idea del RAPT está inspirada en el Integrated Pitch Tracker y posee iguales características. Solo se diferencia en algunos puntos importantes. Primero: el NCCF se calcula a partir de la señal de habla en lugar del residuo de los

coeficientes de predicción lineal (LPC). Segundo: las dos variantes de la NCCF para muestrear los datos, son utilizadas para reducir la carga computacional del proceso. Usa la interpolación de picos a una tasa de muestreo original para incrementar la precisión.

El RAPT no requiere ningún preprocesamiento de la señal de entrada y ofrece un buen rendimiento para las muestras de entrada de la señal de voz a cualquier tasa de muestreo típica de audio [6 KHz $\leq$ Fs $\leq$ 44 KHz]. El costo computacional crece linealmente y a grandes rasgos a medida que aumenta la frecuencia de muestreo. En algunos casos puede ser económico disminuir las muestras en la preparación del segmento seleccionado. Para segmentos de habla donde el ruido de fondo tiene un componente periódico significativo, deben realizarse acciones para remover la periodicidad. Mientras la estimación de F0 es solamente afectada débilmente por bajos niveles de ruido periódico, la determinación del estado de sonoridad puede ser fuertemente afectada. También es necesario usar filtros para la cancelación del ruido introducido por la línea de 60 Hz. En casos extremos de ruidos se corta al centro ese segmento de la oración para aumentar la credibilidad de la determinación del habla y se adiciona un ruido blanco para enmascarar la periodicidad del fondo que esté varios dB por debajo de la amplitud normal del habla (para más información sobre este método ir a [28]). NCCF es la fuente para detectar los períodos candidatos para formar la curva de F0. El costo mayor en el algoritmo es calcular estos períodos. Limitar el rango de los valores de F0 ayuda a reducir el cálculo computacional del proceso.

En este algoritmo se usa la programación dinámica para seleccionar la mejor F0 y el mejor estado sordo-sonoro, basado en la combinación de la información local y la información contextual. Aunque esté disponible todo el discurso de la señal para esta optimización, la F0 solo converge para unas pocas decenas de milisegundos alrededor de la muestra en estudio [28].

Conclusión

Este algoritmo de selección de la F0 hace una clasificación binaria del habla en sordo/sonoro dando evidencias de la presencia o la ausencia de la misma en la señal analizada. Dentro del modelo limitado de sonoridad en sus dos estados de trabajo y la excitación en una sola frecuencia el RAPT provee una curva de F0 confiable y brinda estimados de la señal mediante la consideración de todas las posibilidades al mismo tiempo dentro de un contexto temporal largo.

2.1.3 Modificaciones para la reducción de los errores de detección

Al procesar las oraciones de los pacientes con afectaciones en el habla por el algoritmo detector de F0 del RAPT se puede concluir lo siguiente: presenta una baja tasa de errores refiriéndonos a inserción de múltiplos y submúltiplos de F0. Este algoritmo es muy bueno siguiendo la curva de F0 y tiene baja calidad en cuanto a la detección de silencios, por lo que se hace necesario incorporar un detector de sonoridad que utiliza cruces por cero y energía. Se ha elegido el detector creado por Rabinel & Sambur en 1974 [29]. Para llevar a cabo el proceso de corrección de errores se realizan los pasos siguientes:

- Se eliminan todos los puntos aislados en la detección de sonoridad, es decir si hay un silencio que dure 10 ms es un error y entonces se invierte el signo de la señal en ese punto. El umbral que se ha utilizado es de 10 ms porque el tracto vocal se mantiene constante en un tiempo estimado de 20 ms según plantea van Beinum [30].
- Corrección de F0:

  Se buscan los límites máximos y mínimos de los segmentos sonoros. En aquellos menores o iguales a 40 ms se verifica que el valor medio y el rango de F0 se encuentren en el intervalo definido por la media de F0 $\pm$1.2 por la desviación estándar de F0 (resultado experimental) y el rango del segmento debe ser menor que la desviación estándar de F0. Esto se hace experimentalmente porque en los segmentos cortos pueden ocurrir errores en el detector de sonoridad, como por ejemplo atrapar una ráfaga de aire. Si F0 no es estable se descarta el segmento como sonoro y se continúa el análisis.

- Se buscan errores de sustitución de F0 por sus múltiplos o submúltiplos y posibles errores del detector de sonoridad en el segmento sonoro en estudio.

  Se toma el primer valor del segmento, sí se encuentra entre el valor medio de la F0 ±1.2 por la desviación estándar de F0 se toma como el umbral del segmento. De lo contario se multiplica por 2 y se repite el chequeo, de pertenecer al rango, se duplica el valor inicial y se toma como umbral. La misma operación se repite para los valores que estén a la mitad (es decir como submúltiplos de F0). De no ocurrir lo anterior se abre una ventana de 5 muestras y se toma la mediana como valor inicial y umbral.

  Luego para cada valor de una región de sonoridad se verifica que se encuentre por encima o por debajo de 0.6 por el umbral y 1.8 por el umbral, de ser así se divide por el entero más cercano entre el cociente de umbral y el valor de análisis. De no ser así se mantiene y se pasa a la comparación. La comparación es entre el umbral y el valor de la F0 en estudio. Si este valor no está en un rango de ±0.1 por el umbral se abre una ventana de 5 muestras y se le da el valor de la mediana a la F0 en estudio, luego de la detección actualiza el umbral.

- Se realiza el chequeo de los puntos adyacentes al final del segmento sonoro.

  Primero se verifica que no se haya llegado al final de la señal. En tal caso no es necesario realizar este paso de corrección. Se actualiza el intervalo de búsqueda para chequear los puntos entre dos segmentos sonoros. Si el punto adyacente al final de un segmento sonoro está en el intervalo de ±0.1 por el umbral, o presenta una sustitución de F0 por un múltiplo o submúltiplo del mismo, corregido como en el paso anterior (pero solo para un valor de 2), y que se encuentra en el intervalo de búsqueda se considera como sonoro; se actualiza el detector de sonoridad y los umbrales. Los valores menores de 63 Hz se eliminan de la detección para eliminar el ruido que produce la fuente de potencia. La búsqueda termina con el primer valor que quede fuera del intervalo.

2.2 Algoritmo para detectar intensidad

La intensidad de la voz en la mayoría de los casos se toma como una curva suavizada producida por el proceso de elevar al cuadrado las muestras de la señal de voz, en algunas ocasiones se introducen procesos de filtrado para eliminar ruidos y componentes de la señal de voz que no aporten información a la curva antes mencionada. En el presente trabajo se toma de la literatura del tema un algoritmo que realiza una selección cuidadosa del proceso de filtrado de las bandas de frecuencias y el suavizado de la curva de intensidad; el cual resalta notablemente las regiones de interés para el análisis de la acentuación y disminuye las contribuciones de ruidos y regiones de menos interés.

2.2.1 Algoritmo de Wang

Este algoritmo propone un método directo para estimar la razón del habla a partir de las características acústicas sin ayuda de transcripciones automáticas de la misma Figura 2.1. La señal de habla pasa a través de un banco de filtros de 19 canales compuestos por filtros de segundo orden de Butterworth espaciado como en el trabajo de Holmes [31]. El suavizado de energía se realiza a 50 Hz para dar una razón de tramas de 100Hz y también de los 19 canales se conservan 12 que son los de mayor energía. Después se segmenta la señal en tramas de 11 muestras y estas tramas son ponderadas por una ventana gaussiana, luego se le aplica la correlación temporal realizando el solapamiento entre tramas con la ecuación 1.

$$y(n) = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} x_i(n) x_j(n) \qquad (1)$$

Donde *n* es el número de bandas y *N(N-1)/2* es el valor del coeficiente *M*. Finalmente el conteo de máximos se realiza sobre la envolvente suavizada con validación del F0 y varios mecanismos de umbralación [32].

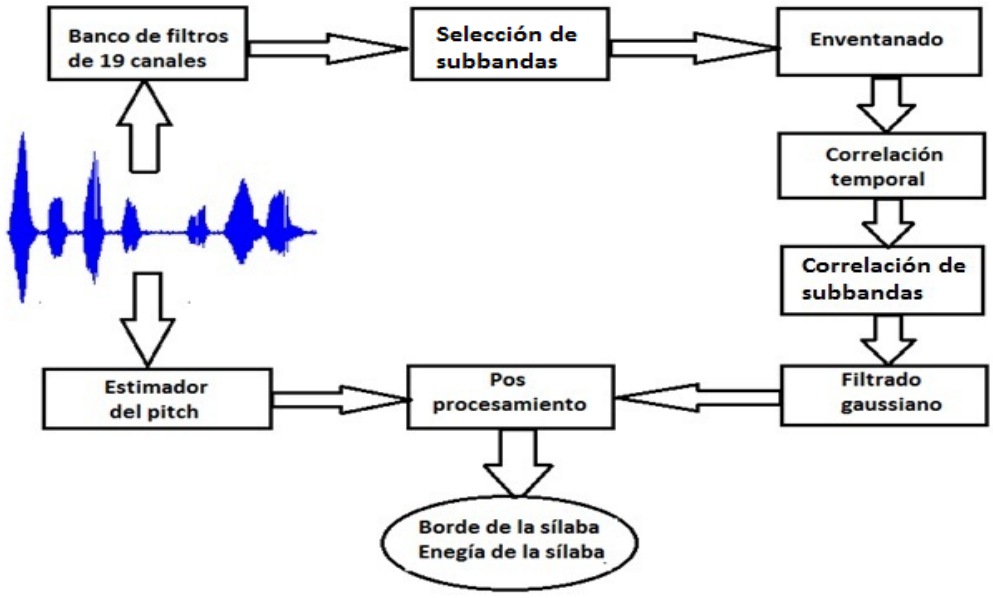

Fig. 2.1 - *El diagrama de flujo del sistema para la estimación de los bordes de sílabas del discurso.*

Según Wang, [33] si se concentran en las bandas prominentes donde están los formantes, el segmento vocálico puede ser amplificado, mientras que la energía de las consonantes disminuye significativamente. Luego se realiza una correlación temporal. En lugar de escoger solo cuatro subbandas se aplican 19 subbandas provistas en el *Speech Filing System Tool* [34] y de aquí se toman los 12 canales de mayor energía.

La correlación temporal está inspirada en la cros-correlación espectral y en el hecho de que cada sílaba dura varias decenas de milisegundos. Ahora si x_t, $x_{t+1},\ldots..,x_{t+k-1}$ representan los incrementos de tiempo de una subbanda de energía de longitud K, la correlación se puede calcular con la ecuación 2.

$$y_t = \sqrt{\frac{1}{2K(K-1)} \sum_{j=0}^{K-2} \sum_{p=j+1}^{K-1} x_t + j \bullet x_t + p} \qquad (2)$$

A través de esta correlación cada sílaba va a tener un máximo en el centro de las sílabas. Comparado con la ponderación lineal de tramas vecinas, la fórmula anterior utiliza el producto que magnifica a la trama que tiene la sílaba, proceso que muestra una marcada mejoría en los resultados. También ofrece una mejora al efecto del enventanado sobre el envolvente de energía.

Para enfatizar las discontinuidades intersilábicas se aplica una ventana de ponderación gaussiana centrada en el medio de la trama (antes de la cros-correlación temporal) por lo tanto la parte central, en el caso de que haya una pequeña discontinuidad, es amplificada y esta trama tiene más peso en el proceso. Después de la correlación espectral y la correlación temporal se obtiene una envolvente de una dimensión. Esta puede contener picos espurios y por lo tanto un suavizamiento de la misma es necesario. En el suavizamiento se aplica un filtrado gaussiano estándar con varianza de 1.3 y longitud de 15 muestras.

2.2.2 Modificaciones introducidas al algoritmo de Wang

Se realiza un submuestreo a 16 kHz para disminuir el número de cálculos (se corrobora que disminuye 6 veces el tiempo requerido para los cómputos para el software MatLab R2009a sobre sistema operativo Windows 7, microprocesador AMD Turionx2 64). Se elimina el proceso de submuestreo que propone Wang que limitaba la señal a 50Hz, es decir una salida con frecuencia de muestreo de 100Hz, esto se debe a la dificultad que acarrea el posterior escalado de la curva de energía sobre el espectro y la señal de voz, que resultaría en gráficas con menos resolución y menor utilidad para el diagnóstico. El enventanado cambia su tamaño y el solapamiento con ventanas de 10 ms cada 5 ms; en lugar de 110 ms con avance de 10ms. Con la envolvente de la energía propuesta se procede al cálculo del centro de la sílaba y se estiman los bordes del núcleo silábico. Para ello se considera que cada máximo de la envolvente que se encuentre en una región sonora (que posea un valor de F0 distinto de 0 Hz), es candidato a ser el centro de una sílaba. Cada candidato debe pasar la prueba del umbral izquierdo donde la magnitud del máximo es comparada con la magnitud del mínimo que se encuentra a su izquierda. Si la magnitud del máximo no excede a la del mínimo en un 10%, este es eliminado y no se considera como una sílaba.

Para estimar los bordes del núcleo silábico requeridos para determinar la duración de la sílaba, se crea un arreglo que contiene la posición en el tiempo de los mínimos de la envolvente de energía y los instantes de inicio y fin de las regiones de sonoridad. Se consideran bordes del núcleo silábico a los valores más cercanos de este arreglo, a ambos lados del máximo de la sílaba, dado

que el núcleo silábico es una región sonora y la envolvente de energía estimada está diseñada para tener mínimos en las regiones donde se pronuncia una consonante. En resumen, los límites del núcleo silábico se determinan por la consonante más cercana o por el inicio/fin de vibración de las cuerdas vocales.

2.3 Materiales y métodos

Se escogieron 17 pacientes de la base de datos de las clínicas Mayo [22]. Estos pacientes se seleccionaron de modo que presenten desviaciones en las dimensiones subjetivas evaluadas por De Bodt et al, 2002 [35], calidad vocal, articulación, nasalidad y prosodia, y como consecuencia en la inteligibilidad. Al mismo tiempo representan varios grupos disártricos como se aprecia en la Tabla II-1. Sus edades están entre 20 y 70 años aproximadamente.

Tabla II-1 DISTRIBUCIÓN DE PACIENTES POR SEXO Y GRUPOS DISÁRTRICOS.

Grupos Disártricos	Pacientes		
	Masculino	Femenino	Total
Atáxicos	2	1	3
Flácidos	1	2	3
Corea	0	3	3
Parkinson	3	0	3
Espástica	0	2	2
Temblor Orgánico	0	3	3

El programa ECAH se utiliza para la selección de la muestra de pacientes, atendiendo a que sean los casos con mayor grado de dificultad para someter el algoritmo propuesto a las situaciones más extremas.

El MatLab R2009a [36] sobre el sistema operativo Windows 7 es la plataforma sobre la que se desarrollan los algoritmos propuestos en los epígrafes 2.1 y

2.2, así como las modificaciones realizadas y se calculan las corridas de datos correspondientes al experimento.

2.3.1 El experimento

En la base de datos aparece un párrafo leído por cada paciente (Grandfather Passage), del mismo se seleccionan las dos primeras oraciones: la primera es una interrogación, lo cual facilitará emitir algunos criterios sobre la prosodia, al igual que en la segunda, en la cual aparece una pausa. Las grabaciones de las voces se convierten a un formato PCM, con frecuencia de muestreo de 44.1 Khz y con resolución de 16 bits con un canal (mono). El almacenamiento se efectúa en un archivo con extensión .WAV utilizando para todo el proceso la herramienta "Sound Recorder" que brinda el sistema operativo Windows 7. Estas oraciones se procesan de forma independiente por los algoritmos propuestos.

El resultado son: el contorno de F0, la posición y el valor del máximo de energía de cada sílaba detectada, así como la frontera del núcleo silábico. Las marcas se colocan en un espectrograma de banda estrecha con mejor resolución frecuencial y esto último sirve tanto como un resultado o un método eficaz de corroborar el desempeño de los algoritmos. Una vez sintonizado los algoritmos se procede a su incorporación a una interfaz gráfica que permita su manejo por parte de los especialistas de la consultas de Logopedia y Foniatría.

Este proyecto está orientado a la visualización del patrón de acentuación. La interpretación del mismo queda por parte de los especialistas debido a que no existe un consenso sobre qué parámetros medir. En esta aproximación se propone medir la intensidad máxima de la palabra y/o la oración apoyada por la duración del núcleo silábico y el valor máximo de F0 comprendido en dicho intervalo.

Conclusiones del capítulo: Se seleccionan y se modifican los algoritmos necesarios para dar cumplimiento a los objetivos del presente trabajo. También se ha diseñado un experimento que ponga a prueba las debilidades del procedimiento propuesto.

III. CAPÍTULO 3: RESULTADOS Y DISCUSIÓN.

Resumen

En el presente capítulo se muestran los resultados obtenidos, la evaluación de los algoritmos de Frecuencia Fundamental (F0), la envolvente de intensidad o energía y la duración de las sílabas, por último se muestra el funcionamiento de la interfaz gráfica confeccionada.

3.1 Resultados de los algoritmos propuestos sobre la base de datos Aronson

Para dar respuesta a la interrogante científica del presente trabajo, se seleccionan algoritmos bien referenciados en la literatura a los que se les introduce un cierto número de modificaciones para su ajuste y uso en voces patológicas. El presente epígrafe valora los resultados obtenidos en la evaluación de los algoritmos sobre una muestra de voces con disímiles enfermedades tomadas de la base de datos Aronson. Los resultados se muestran en la Tabla III.1. En la misma se analizan los errores que más afectan el proceso de detección; que serían la incorrecta detección de sílabas (omisión o inserción) o diversos errores en el detector de F0, como puede ser la sustitución del valor de F0 por un múltiplo o submúltiplo del mismo. Ambos errores poseen el poder de modificar notablemente la correcta detección de las sílabas acentuadas, ya que influyen en el valor de F0, en la detección de la sílaba y en la posición de los bordes del núcleo vocálico.

Tabla III.1. Resultados de la detección.

Paciente	Oración	No. Inserciones	No. Omisiones	No. Sílabas	Errores del detector de F0
AT3F	1		1	12	
	2	1		16	
AT5M	1			9	Múltiplo de F0 (3%)
	2			17	
AT7M	1			8	
	2		2	14	
FD3M44	1		1	9	Submúltiplo de F0 (5%)
	2		1	16	
FD4F	1			8	

	2			17	
FD5F	1		1	10	Submúltiplo de F0 (20%)
	2		1	19	
KR2F	1		1	10	Submúltiplo de F0 (5%)
	2		1	19	
KR6F	1	1	1	10	Submúltiplo de F0 (3%)
	2	1		19	Submúltiplo de F0 (2%)
KR7F	1	1	3	11	Submúltiplo de F0 (20%)
	2	2		19	Submúltiplo de F0 (5%)
PK2M	1		3	8	
	2		3	18	
PK7M	1		2	11	
	2		2	16	
PK8M	1		3	10	
	2		2	14	
SD3F	1		1	10	
	2			18	Submúltiplo de F0 (20%)
SD5F	1	1		12	
	2	2	1	22	
VT2F	1			11	Submúltiplo de F0 (3%)
	2	2	1	18	
VT5F	1		1	11	Submúltiplo de F0 (10%)
	2	3	1	20	
VT6F	1	1	1	11	
	2		1	20	Submúltiplo de F0 (10%)

En síntesis, la evaluación de los algoritmos fue satisfactoria, en 34 oraciones pronunciadas por 17 pacientes, solo 12 presentaron errores en la detección de F0, de ellas solo 5 presentaron errores mayores que un 5% del contorno detectado. De 473 sílabas pronunciadas solo se omitieron 35 y se insertaron 15 para un error total de un 10.57%. De estos errores, 11 se deben al mal

suavizamiento de la curva de intensidad (con respecto a este punto Wang sugiere un suavizado adaptativo), 5 al umbral izquierdo anormalmente elevado y el resto a disímiles razones como ruidos, ininteligibilidad de la voz, etc. Sin embargo es una cifra sorprendentemente favorable, por lo que se propone verificar el desempeño del detector de sílabas en otras bases de datos de pacientes disártricos. El grupo con mejores resultados es el de atáxicos con solo 1 inserción, 3 omisiones en 6 oraciones y un 3% de error en el detector de F0 en una sola oración. En la Figura 3.1 se muestra un ejemplo de los resultados de un paciente con dicha enfermedad. El peor grupo es el de Korea que muestra 5 inserciones, 6 omisiones y errores en el detector de F0 en 5 de las 6 oraciones; el cual se observa en la Figura 3.2.

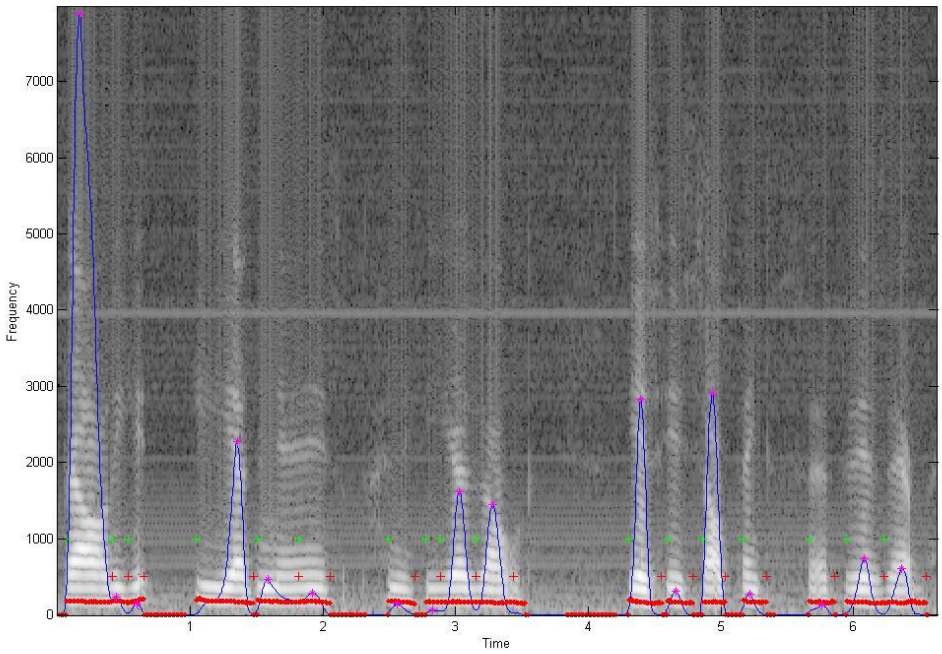

Fig. 3.1 - *Ejemplo de un paciente atáxico (AT5M). Línea azul envolvente de energía, línea roja F0, las marcas verdes y rojas comienzo y el fin de la sílaba y marcas lilas el centro de la sílaba.*

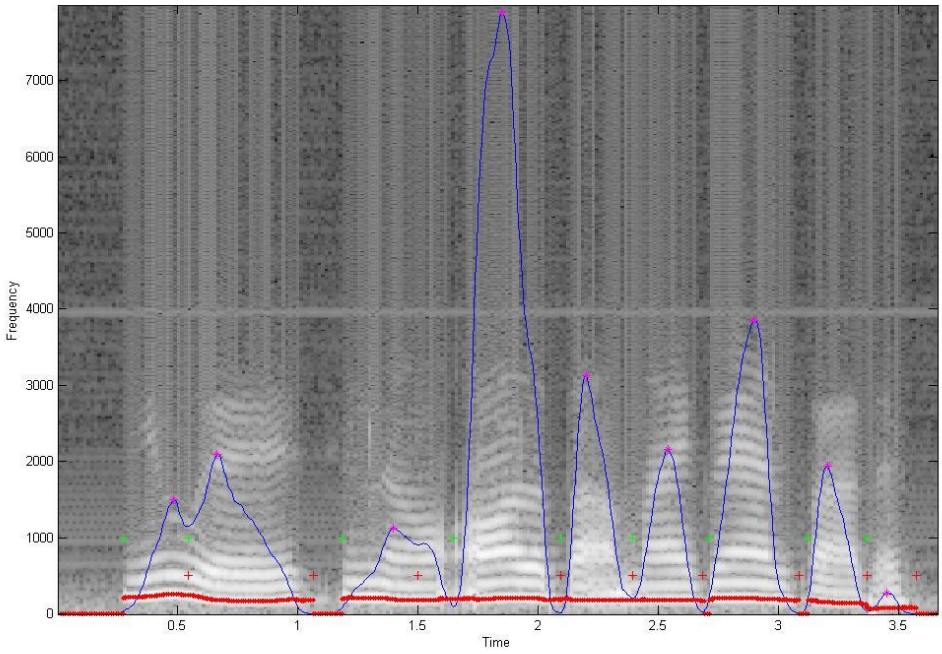

Fig. 3.2 - *Ejemplo de un paciente afectado de Korea (KR2F). Línea azul envolvente de energía, línea roja F0, las marcas verdes y rojas comienzo y el fin de la sílaba y marcas lilas el centro de la sílaba.*

En las Figuras 3.1 y 3.2 se muestran algunos de los resultados que se obtienen mediante los algoritmos antes propuestos, donde la curva azul representa la envolvente de energía, la roja representa la variación de F0, las marcas verdes y rojas indican el comienzo y el fin de la sílaba y por último las marcas lilas nos ofrecen el centro de la sílaba.

Con el objetivo de verificar en qué medida los algoritmos antes mencionados dan cumplimiento a los objetivos del presente trabajo. Se propone un segundo análisis de los errores contenidos en la Tabla III.1. En este caso solo se evalúa la correcta detección de sílabas acentuadas.

Las sílabas acentuadas son detectadas manualmente por dos especialistas. Para reducir la subjetividad de la detección se auxilian de la curva de intensidad generada por el algoritmo de Wang. Se estudian de forma independiente todas las oraciones con dificultades en la detección de sílabas y frecuencia fundamental.

Como resultado se encuentran 7 errores cometidos en sílabas acentuadas, de ellos 2 son omisiones de la misma y el resto solo errores en el detector de F0

que generan valores anómalos de frecuencia fundamental en la sílaba acentuada. Solo 5 oraciones se ven afectadas, de ellas 3 con solo un error. Este favorable resultado se debe a que la sílaba acentuada es más prominente en duración y energía, lo cual facilita su detección.

3.2 Resultados de los algoritmos propuestos sobre un paciente de las consultas de Logopedia y Foniatría del país

El presente proyecto ha de ser utilizado en las consultas de Logopedia y Foniatría, las cuales poseen características diferentes en cuanto a condiciones de grabación e idioma de los pacientes. Por lo tanto se graba a un paciente en la consulta de Logopedia y Foniatría con las siguientes características: Enfermedad: Disartria secundaria a Enfermedad de Parkinson; Operado, (Subtálamotomia). Edad: 43 años. Lenguaje: excelente/normal. Habla: Articulación general: Articulación superficial, desdibujada, pobre, torpe. Articulación aislada: /r/ distorsionada, y alternancias de /l/ simple y /l/ compleja. Fluidez: normal. Voz: Tono: grave. Timbre: áspero, espástico, aireado. Intensidad: hipofonía. Entonación: monotonía. Resonancia: nasalidad ligeramente aumentada.

Se analiza una muestra de habla fluida donde el paciente intercambia con la doctora una de sus vivencias diarias. La tabla III. 2 contiene 10 segmentos de habla de diferentes longitudes de tiempo, entre 4 y 8 segundos, donde cada uno contiene una oración; estas se analizan de forma independiente por el software antes mencionado. Entre los resultados más sobresalientes se observa que no hay errores por parte del detector de F0 y que no se omiten sílabas acentuadas como muestra la oración Dial2_seg 2.1 en la Figura 3.3. En cambio como se aprecia en la Tabla III. 2 se comenten 25 errores, en su mayoría omisiones de sílabas, esto se debe a las deficiencias en el habla del paciente en especial a su baja razón del habla que no supera las 3 sílabas por segundo, se muestra un ejemplo en la oración Dial2_seg1.2 la cual se observa en la Figura 3.4. En total el error en el detector de sílabas fue de un 19.23%, aunque se verifica que las sílabas acentuadas al ser las más prominentes son mucho menos sensibles a ser omitidas.

Tabla III.2 Resultados de la detección.

Archivo	No. inserción	No. omisión	No. sílabas
Dial1_Seg1	1	2	11
Dial2_Seg1.1	-	3	17
Dial2_Seg1.2	-	5	9
Dial2_Seg1.3	-	4	11
Dial2_Seg2.1	-	-	10
Dial2_Seg2.2	-	-	13
Dial3_Seg1.1	-	1	11
Dial3_Seg1.2	-	3	12
Dial3_Seg1.3	-	2	16
Dial3_Seg1.4	-	5	20

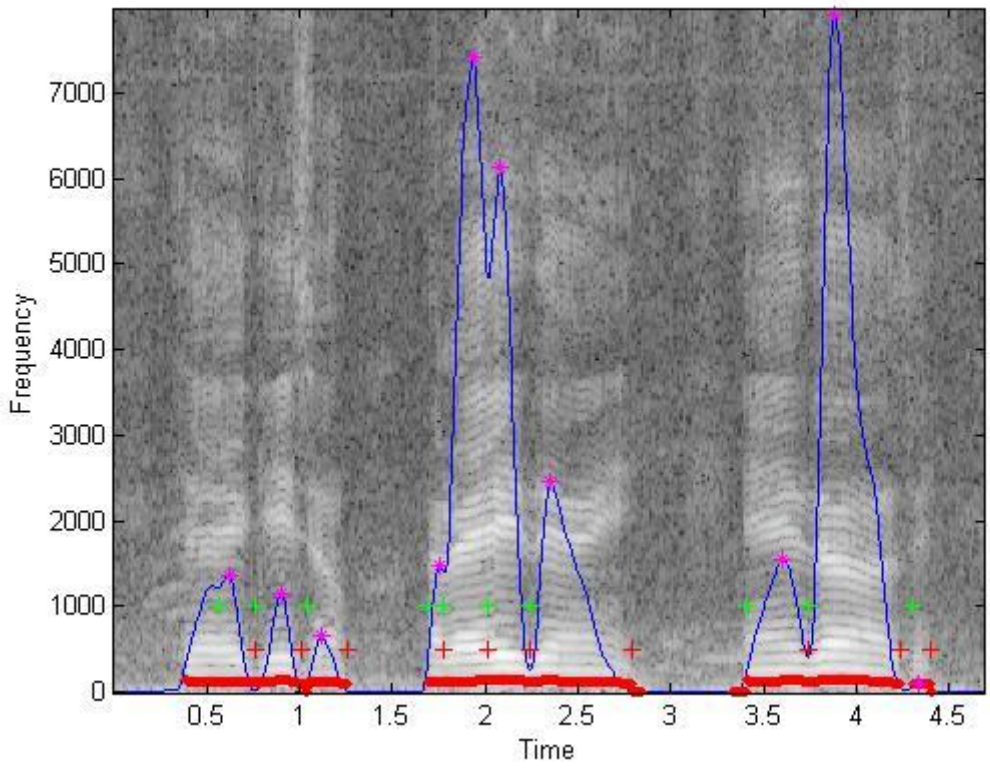

Fig. 3.3 - *Correcta detección de F0 y de sílabas acentuadas. Línea azul envolvente de energía, línea roja F0, las marcas verdes y rojas comienzo y el fin de la sílaba y marcas lilas el centro de la sílaba.*

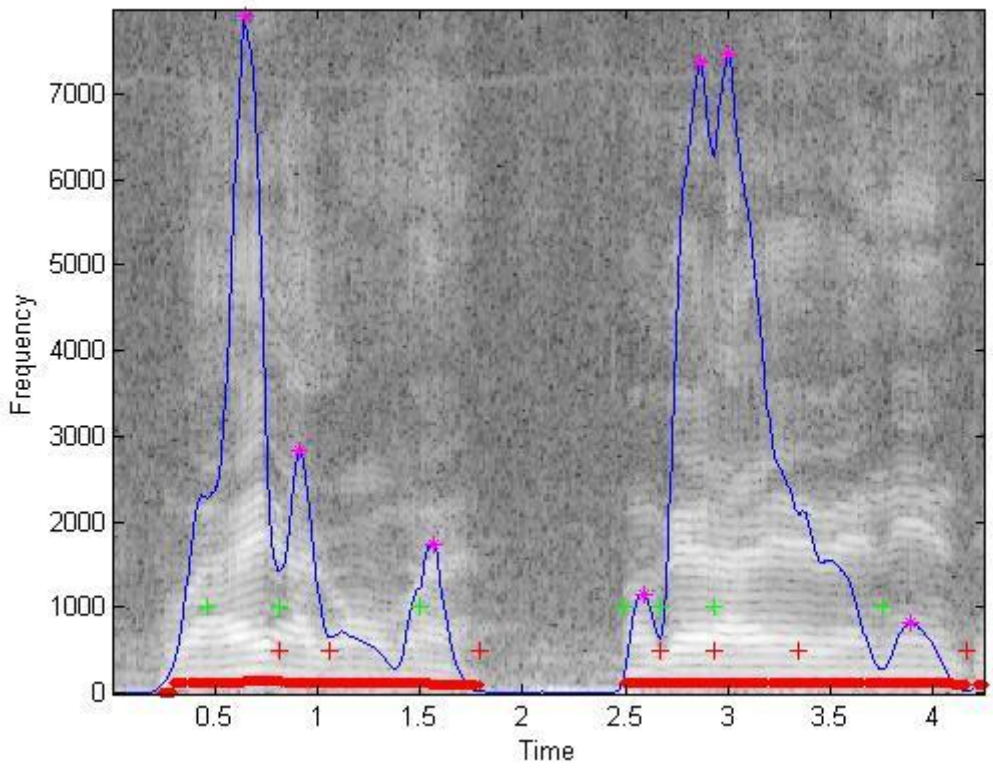

Fig. 3.4 - *Omisión de sílabas. Línea azul envolvente de energía, línea roja F0, las marcas verdes y rojas comienzo y el fin de la sílaba y marcas lilas el centro de la sílaba.*

3.3 La interfaz gráfica

A partir de los algoritmos explicados anteriormente en el Capítulo 2, utilizando el programa MatLab 7.8.0 R2009a, se implementa un software cuya interfaz gráfica se muestra en la Figura 3.5; el cual nos brinda facilidades como cargar una señal de voz y reproducirla. Además permite obtener la frecuencia fundamental de una señal, la intensidad y la duración de las sílabas. A continuación se explica su funcionamiento.

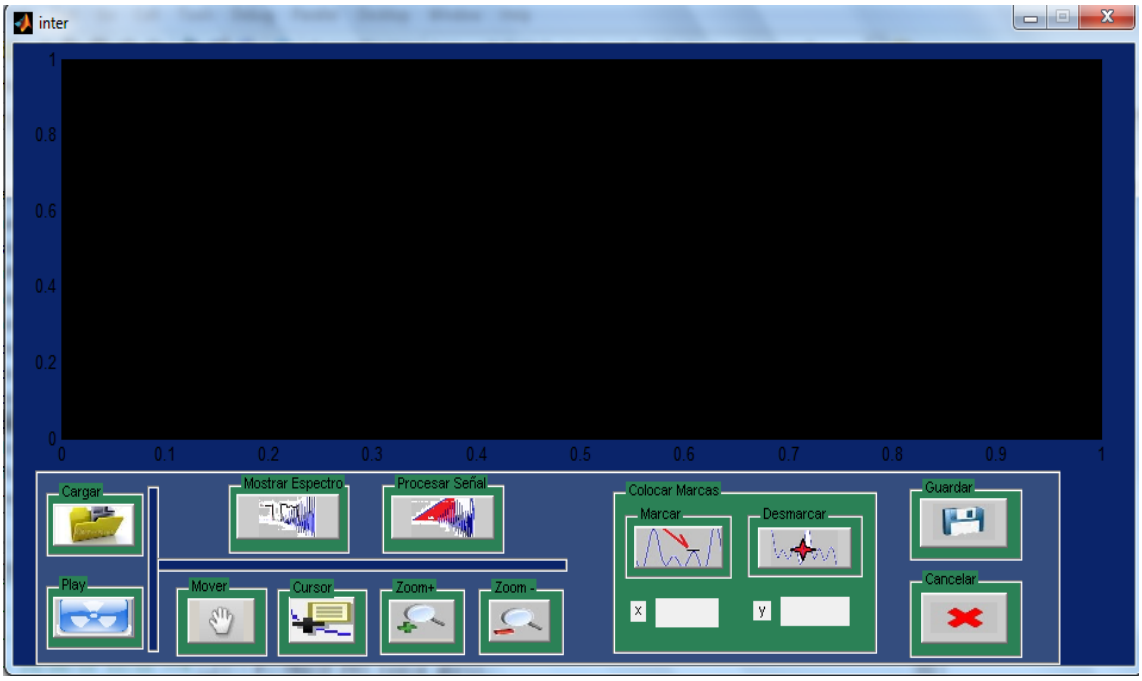

Fig. 3.5 - *Interfaz gráfica para análisis de acentuación.*

La interfaz gráfica posee una gráfica que muestra la señal cargada y los resultados de los algoritmos diseñados, también tiene una serie de botones que permite el manejo y procesamiento de la señal de voz. Estos botones se explican a continuación observándose los mismos en la Figura 3.6.

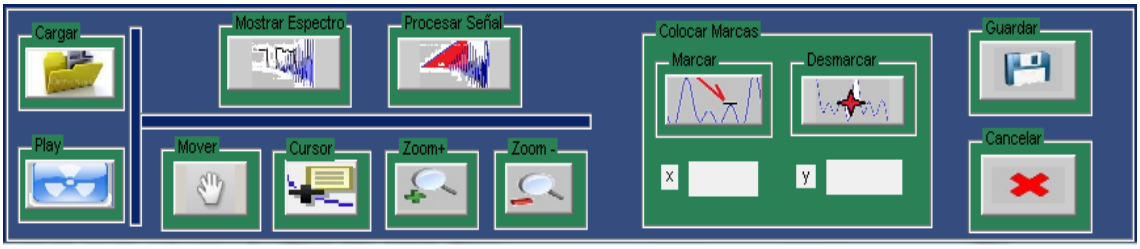

Fig. 3.6 - *Panel de control de la interfaz gráfica*.

Cuando se acciona en el botón **Cargar** se puede observar una ventana como la que se muestra en la Figura 3.7, esta permite cargar archivos con extensión *.WAV. Aquí se puede seleccionar cualquier archivo, el cual va a ser mostrado en la gráfica, como se observa en la Figura 3.8. El botón **Play** le permite al especialista oír la grabación ofreciéndole la posibilidad de obtener una evaluación subjetiva de la misma y comparar dicha evaluación con el análisis objetivo facilitado por esta herramienta.

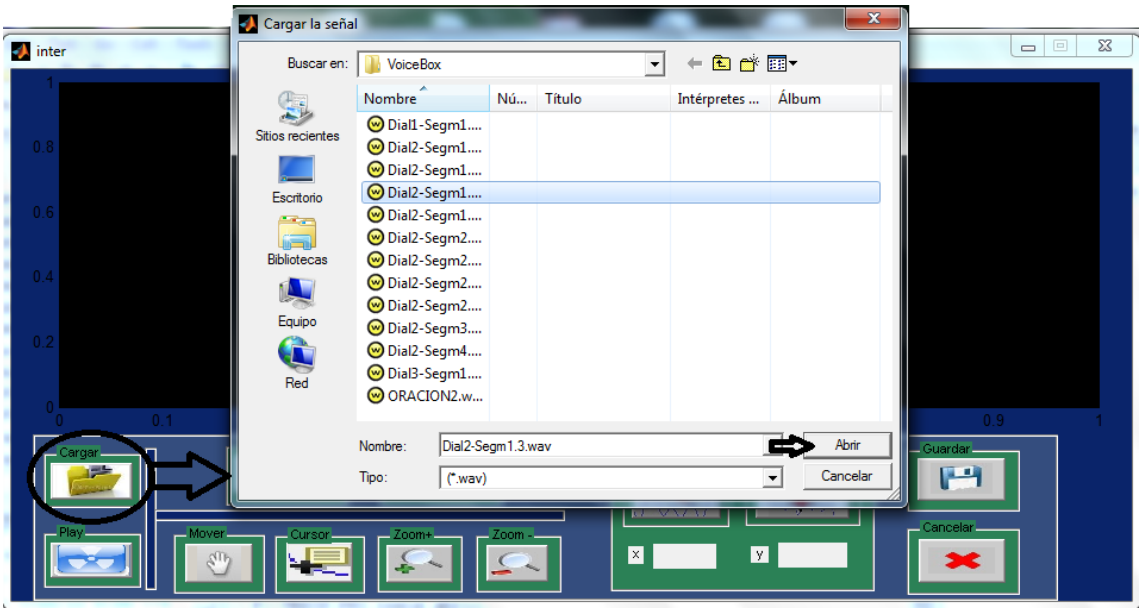

Fig. 3.7- *Cuadro de diálogo que se muestra al accionar el botón cargar*

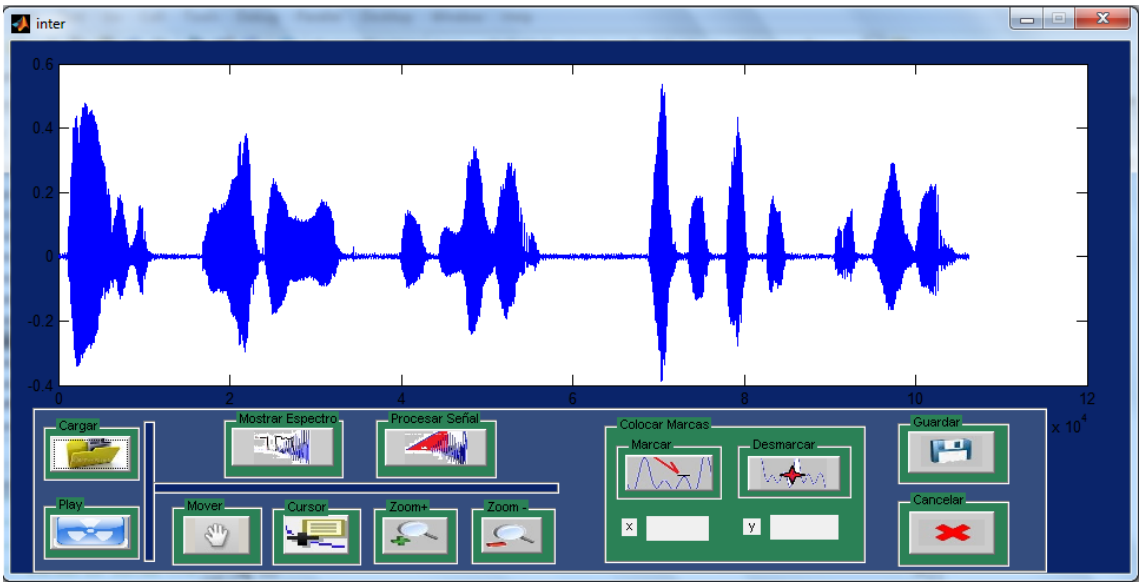

Fig. 3.8 - *Señal de voz cargada.*

Como se observa se han agregado varios botones para facilitar el trabajo con la señal de voz cargada como son: **Mostrar Espectro** y **Procesar Señal** los cuales procesan esta señal de voz y muestran la curva de F0 expresando la escala en Hz, la curva de intensidad, la duración de la sílaba y el espectrograma de la señal de voz cargada, como se observa en la Figura 3.9 y Figura 3.10. El botón **Mover** facilita al usuario desplazar las señales dentro del axes ofreciendo mejor dinamismo y soltura para la obtención de cualquier detalle necesario. El botón **Cursor** permite colocar una etiqueta de datos sobre

cualquier curva en la señal deseada con el objetivo de delimitar su posición respecto a los ejes de coordenadas. Los botones **Zoom +** y **Zoom –** permiten aumentar y disminuir el tamaño de las señales respectivamente para un mejor análisis. También posee dos botones **Marcar** y **Desmarcar** en los cuales se puede colocar y quitar marcas manualmente, mostrando las coordenadas (x, y) justo debajo del botón, de esta manera el operador puede colocar y remover sus marcas si así lo desea para cambiar la duración de las sílabas el botón, también se le agrega el botón **Cancelar,** el cual conduce a un cuadro de diálogo que pregunta si se desea salir o no. La Figura 3.11 permite observar el uso de este botón. Además se adiciona el botón **Guardar** con el objetivo de que el especialista pueda transportar en una memoria los datos del paciente como F0, intensidad y duración de las sílabas, los cuales se guardan en un Excel para su posterior estudio la Figura 3.12 muestra su uso.

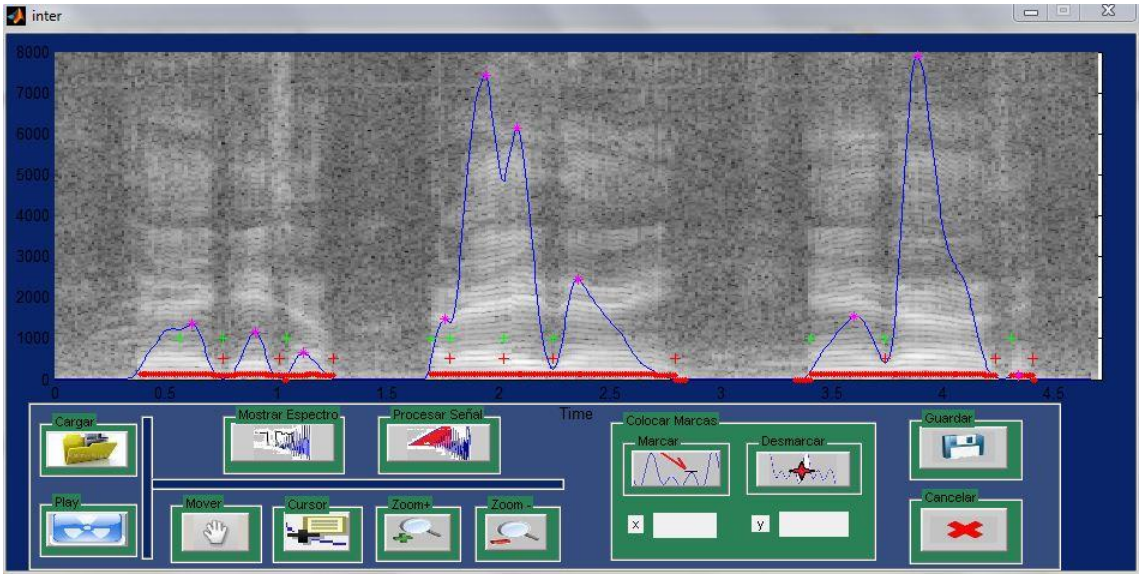

Fig. 3.9 - *Espectro de la señal cargada. Línea azul envolvente de energía, línea roja F0, las marcas verdes y rojas comienzo y el fin de la sílaba y marcas lilas el centro de la sílaba.*

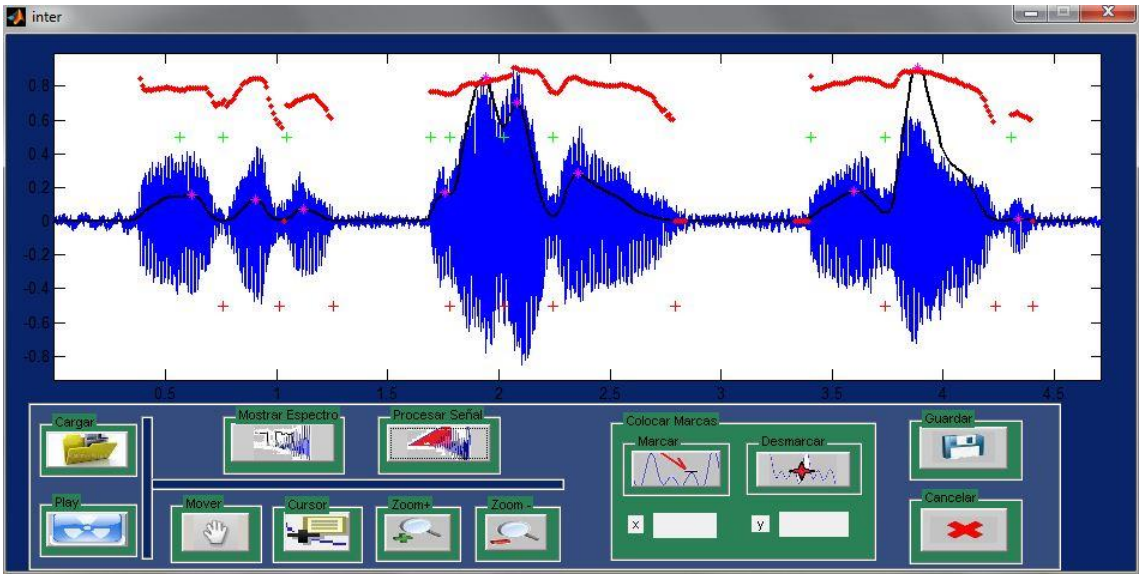

Fig. 3.10 - *Señal en el dominio de tiempo. Línea color negro envolvente de energía, línea roja F0, las marcas verdes y rojas comienzo y el fin de la sílaba y marcas lilas el centro de la sílaba.*

.

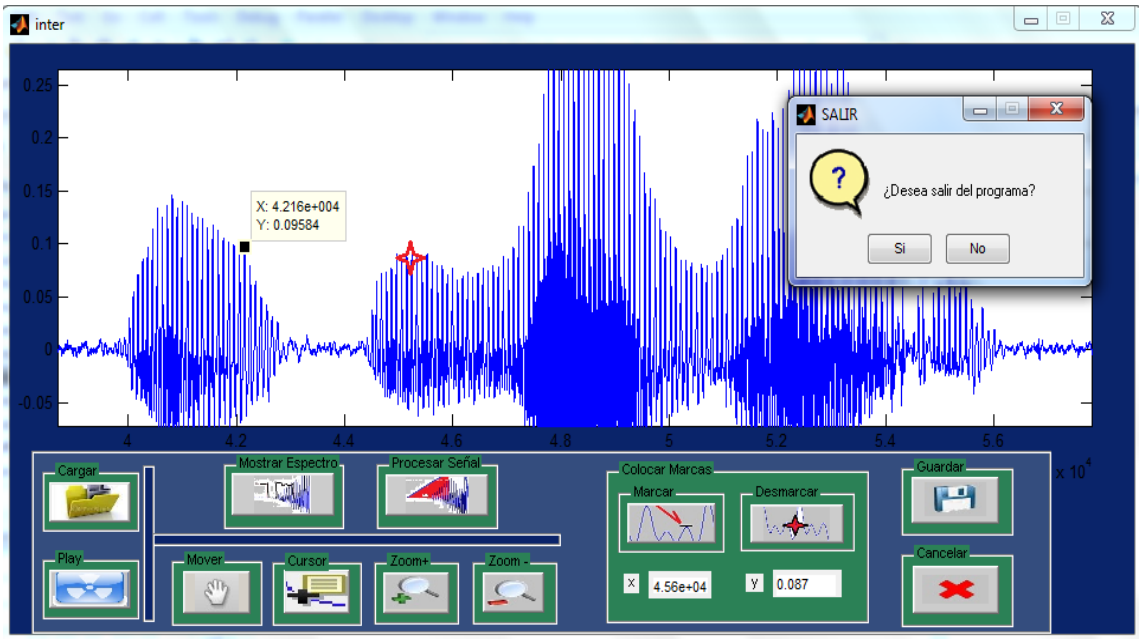

Fig. 3.11 - *Uso de los botones adicionados*

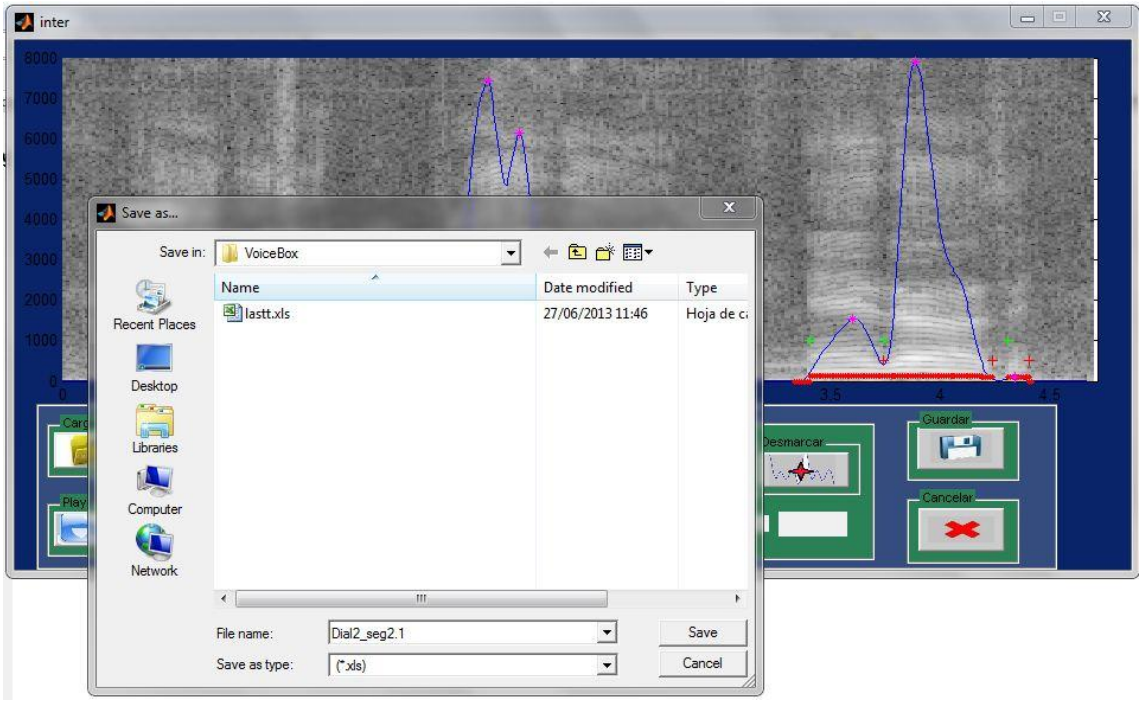

Fig. 3.12 - *Cuadro de diálogo que se muestra al accionar el botón guardar*

A manera de resumen el uso de la interfaz gráfica *inter* es relativamente sencillo. Este puede ser ejecutado en varios pasos como: Cargar la señal, analizar la misma mediante los botones **Mostrar Espectro** y **Procesar Señal**, realizar observaciones sobre la misma con los botones **Zoom +** y **Zoom –, Mover, Cursor** y **Colocar Marcas.** Así como realizar valoraciones subjetivas escuchando la grabación mediante el botón **Play.** De ser necesario se pueden guardar los parámetros calculados sobre la señal de voz mediante el botón **Guardar,** los datos se guardan en un archivo Excel con extensión ****.xls.*** Luego de concluir el análisis se puede cerrar la aplicación mediante el botón **Cancelar** o proceder a otro análisis siguiendo la misma metodología.

Como principal señalamiento a la herramienta *inter* se encuentra el tiempo que demora en ejecutar los cómputos relacionados con los parámetros necesarios para el análisis de acentuación. La mayor carga computacional recae en el cálculo de F0 y en el proceso de correlación temporal. Además la interfaz no posee facilidades de edición de la señal de voz, las cuales pudiesen facilitar el uso de la misma sin necesidad de la ayuda de otros programas.

CONCLUSIONES Y RECOMENDACIONES

Conclusiones

- Se acondicionaron exitosamente algoritmos refenciados en la literatura para el análisis de acentuación en voces patológicas.
- Se diseñó una interfaz gráfica de fácil utilización para el uso de los algoritmos por parte de los especialistas.
- Se validaron los cálculos de los parámetros cuantitativos asociados a la acentuación en voces patológicas.

Recomendaciones

- Mejorar el desempeño del detector de silabas a partir de una etapa de realimentación que sintonice los umbrales del mismo en aras de disminuir el número de errores de detección.
- Añadir posibilidades de edición y limpieza de ruido a las señales de voz.
- Extender el uso de la herramienta inter como apoyo al proceso de diagnóstico y rehabilitación en las consultas de logopedia y foniatría.

REFERENCIAS BIBLIOGRÁFICAS

[1] Arthur C & Hall Guyton, "Respiración.," in *Textbook of medical Physiology*., ch. VII, pp. 537-538.

[2] H. & Hernández-Díaz, M. , Kairuz, "Análisis del Triángulo Vocálico en pacientes con implante coclear," IV Encuentro Iberoamericano de Trastornos del Lenguaje, Habla y Voz y Congreso Cubano, Habana, Cuba., 2008.

[3] Héctor A. Kairuz Hernández – Díaz, "Detección y marca de los cambios espectrales notorios en voces patológicas," Centro de Estudios de Electrónica y Tecnologías de la Información (CEETI), Universidad Central de Las Villas (UCLV), Tesis de grado 2009.

[4] M. J. Llau Arcusa and J. González, "Medida de la inteligibilidad en el habla disartria," *Revista de Logopedia, Foniatría y Audiología*, vol. 24, no. 1, pp. 33-43, 2004.

[5] Pardo, J. M.; Sistema de producción del habla, Apuntes de Ingeniería Neusensorial; Universidad Politécnica de Madrid; España; 2002

[6] Kent, R., Read, Ch. Acoustic Analysis of Speech; Canada; Delmar. 2002.

[7] Communication technology Group, University of Zaragoza, España, 1997.

[8] Odhe, R. N., and Sharf, D. J. Phonetic Analysis of Normal and Abnormal Speech. Canada: Macmillan Publishing Company. (1992).

[9] Ladefoged. P, "A course in phonetics," *Harcourt Brace Javanovich,* 1975.

[10] Lehiste, I., "Suprasegmentals," *MA: MIT Press,* 1970.

[11] Cruttenden, A. , "Intonation," *Cambridge: Cambridge University Press,* 1986.

[12] Brontein, A. J. , "The pronunciation of American English," *Appleton-Century-Crofts,* 1960.

[13] Shriberg, L. D. and Kent, R. D. ,"Articulation judgments: Some perceptual considerations," *Journal of Specch and Hearing Research,* vol. 15, pp. 876-882, 1982.

[14] Tiffany, M. C. and Carrell, J. , "Phonetics: Theory and application," *McGraw-Hill,* 1977.

[15] Chomosky, N. and Halle, M. , "The sound pattern of English," *Harper and Row,* 1968.

[16] Hyman, L. M., "Phonology: Theory and analysis," *Holt, Rinehart and Winston,* 1975.

[17] Sloat, C. , Taylor, S. H and Hoard, J. E. , "Introducction to phonology," *Englewood Cliffs, NJ: Prentice-Hall,* 1978.

[18] Crystal, D. , "Clinical linguistic," *Viena: Springer-Verlag,* 1982.

[19] Kent, R. D. and Rosenbek, J. C, "Prosodic disturbance and neurologic lesion," *Brain and Language,* vol. 15, pp. 259-291, 1982

[20] Darley .F, Aronson. A, Brown. J. "Differential diagnostic patterns of dysarthria". Jourual of speech and hearing research, 12, 462-496, 1969. en: Neurología para especialistas del habla y del lenguaje. Russell J, Wanda G. edit panam. B. aires. 1992. pág.142.

[21] Prater, J. and Swift, R., "Manual de terapia de la voz. Salvat". Ed. 1989. pág.145.

[22] Aronson A.E., "Dysarthria: Differentia l diagnosis", (CD). Rochester, MN: Mentor Seminars, 1993.

[23] Dworkin, J. P. *Motor speech disorders: A treatment guide.* St. Louis, MO: Mosby-Yearbook. 1991.

[24] Kent, R.D., Weismer, G., Kent, J. F., Vorperian, H.K. and Duffy, J.R. "Acoustic Studies of Disarthric Speech: Method, Progress, and Potential". *Journal of Communication Disorders,* vol. 32, pp 141-186. 1999.

[25] Leuschel, A., and Docherty, G.J. "Prosodic assessment of dysarthria". In D. A. Robin, K.M. Yorkston and D.R. Beukelman (Eds.), *Disorder of motor speech: Assessment, treatment, and clinical characterization* (pp. 155-178). Balt imore: Paul H. Brookes Publishing Company. 1996.

[26] B.J., and Weisiger, B.E. Zski, "Identification of dysarthia types based on perceptual analysis," *Journal of Communication Disorders,* vol. 20, pp. 367-378, 1987.

[27] Schoentgen, J., “Vocal Cues of Disordered Vo iced". *Acta Acustica united with Acustica,* vol 92, pp-667, 2006.

[28] Talkin, D., "*Robust Algorithm for Pitch Tracking*", in *Voice epoch determination with dinamic programming*, A.S.o. America, Editor. 1989, Entropic Research Laboratory.

[29] Sambur, L.R.R.M.R., *An Algorithm for Determinig the Endpoints of Isolated Utterances.* 1974.

[30] Van Beinum, F.J.K., *What’s in a schwa?* Phonetica 51, 1994: p. 68–79

[31] Holmes, J.N., *The JSRU channel vocoder.* IEEE Proc. F. Commun Radar Signal Process, 1980. 127: p. 53-60.

[32] Fosler-Lussier, N.M.E., *Combining multiple estimators of speaking rate.* ICASSP, 1998. 2: p. 729-732.

[33] Narayanan, D.W.a.S.S., *Robust Speech Rate Estimation for Spontaneous Speech.* IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, 2007. 15(8).

[34] *Speech filing system*. Availablefrom: http://cwww.phon.ucl.ac.uk/resource/sfs/.

[35] De Bodt, M. S., Hernandez – Diaz Huici, M. E., Van De Heyning, P.; "Intellig ibility as a linear combination of dimensions in dysarthric speech"; *Journal of Communications Disorders* 5220, pp 1 – 10. 2000.

[36] *Matlab Signal Processing Toolbox*. The Mathworks, 2009