del 26 al 30 de noviembre de 2012
16 CONVENCIÓN CIENTÍFICA DE INGENIERIA Y ARQUITECTURA
PALACIO DE CONVENCIONES DE LA HABANA
cujae

SISTEMA DE AYUDA VISUAL PARA APOYAR APRENDIZAJE DE FONEMAS ESPAÑOLES

VISUAL AID SYSTEM TO SUPPORT LEARNING SPANISH PHONEME

Dr. Enrique San Juan, Dr. Francisco Watkins, Dr. Héctor Kaschel
Departamento de Ingeniería Eléctrica. Universidad de Santiago de Chile.
Av. Ecuador 3519, Estación Central. Santiago de Chile. Chile.
enrique.sanjuan@usach.cl, francisco.watkins@usach.cl, hector.kaschel@usach.cl

RESUMEN

El presente trabajo tiene como propósito mostrar un sistema computacional con la capacidad de apoyar el aprendizaje de la pronunciación de un conjunto determinado de fonemas españoles, dirigido a personas sordas o con dificultades auditivas, que tienen como consecuencia trastornos del habla. La posibilidad de "ver lo que se dice", puede resultar muy útil como método para la implantación y rehabilitación del Habla. Visualizar de forma inmediata, mediante una gráfica, los perfiles acústicos de los principales parámetros de la señal de voz y asociarlos con imágenes que representan lo dicho, ha resultado una alternativa adicional muy estimulante en el campo de la Foniatría. Dicho sistema basa su funcionamiento en la extracción y comparación de parámetros fundamentales de las señales de voz, entre los cuales se puede mencionar los LPC (*Linear Predictive Coding*), Formantes y Coeficientes Ceptrales en la escala de frecuencias de Mel (MFCC,). Se espera que el sistema constituya una herramienta de apoyo a la rehabilitación de trastornos del habla, reemplazando el canal de realimentación auditivo por un canal de realimentación visual. Es decir, mediante gráficas de los perfiles acústicos y principales parámetros característicos de la señal de voz, con imágenes e indicadores de avance, se estructura en conjunto una importante herramienta alternativa adicional para la rehabilitación en trastornos del habla.

PALABRAS CLAVES: Procesamiento digital de señales, LPC, Autocorrelación, Análisis de voz, Formantes, *Mel-Frequency Cepstral Coefficients*.

ABSTRACT

This paper aims to show a computer system with the ability to support the learning of the pronunciation of a particular set of phonemes Spanish, aimed at people who are deaf or hard of hearing, which result in speech disorders. The ability to "see what it says," can be very useful as a method for the implementation and speech rehabilitation. Display immediately, using a graph, acoustic profiles of the main parameters of the speech signal and associate them with images that represent what this has proved very stimulating an additional alternative in the field of Pathology. The system operation is based on the extraction and comparison of key parameters of the speech signals, among which one can mention the LPC (Linear Predictive Coding), Formant and Cepstral Coefficients in Mel frequency scale (MFCC, Mel-Frequency Cepstral Coefficients). It is expected that the system constitutes a tool to support the rehabilitation of speech disorders, replacing the auditory feedback channel for visual feedback channel. Namely, by graphical and acoustical profiles main characteristic parameters of the speech signal, with images and progress indicators, allow to set an important additional alternative tool for rehabilitation of speech disorders.

1.- INTRODUCCIÓN

Las patologías auditivas y de la voz, son consideradas uno de los principales problemas en la comunicación humana. Es por esto la necesidad de desarrollar tecnologías orientadas a la rehabilitación de estos problemas y/o fortalecer su apoyo. Desde mediados de la década de los 80 se inició el desarrollo y comercialización de sistemas de análisis de voz mediante la gráfica de perfiles paramétricos de la señal que la representa, entre los parámetros más comunes tenemos la intensidad de la señal y su cruce por ceros. Estos perfiles paramétricos no sólo se realizaban para la señal pura de voz, también se realizaban para determinadas bandas de frecuencias en las cuales está el mayor contenido de la información hablada: formantes o frecuencias de resonancias del tracto vocal, así como la parte del espectro que caracteriza a los sonidos fricativos y la frecuencia fundamental. En la década de los 90 aparecieron sistemas, que sin mostrar los perfiles de parámetros acústicos, presentaban imágenes capaces de ser movidas o alteradas por la presencia de determinado nivel o duración de un parámetro en específico. A inicios del siglo XXI se continúa el desarrollo de aplicaciones para la educación y se inicia el diseño y programación de sistemas para el análisis de voz en el área médica de consultas de foniatría. A pesar que el campo de investigación en el área del análisis y síntesis de voz lleva varias décadas de desarrollo, los aspectos sobre los sistemas automáticos para el reconocimiento de la misma aún no han sido resueltos totalmente, siendo muy común en la actualidad sistemas de reconocimiento dependiente del hablante y sobre la base de exigencias de entrenamiento previo. Además se debe tomar en cuenta la importancia de la lengua a la cual nos referimos, principalmente porque los fonemas son muy distintos en sonidos, dependiendo de la lengua o idioma de que se trate. En relación con esto, la lengua que tratamos en este trabajo es la española y en específico fonemas producidos por hablantes chilenos.
La tarea de análisis de voz constituye la base para el entendimiento y desarrollo de la producción y síntesis de voz, así como de los algoritmos para la identificación, clasificación y posterior ayuda a la rehabilitación de patologías en la producción del lenguaje hablado. Este sistema, busca entregar al usuario en rehabilitación índices que le ayuden a aprender y a mejorar su pronunciación, basándose en la correlación de los parámetros propios de cada hablante con respecto a parámetros patrones almacenados en el sistema.

2.- TÉCNICAS DE ANÁLISIS

A continuación se describen brevemente las técnicas de análisis utilizadas para la implementación de este sistema.

Formantes de la voz

Los formantes son frecuencias *peaks* del espectro de voz, en torno al cual se concentra la mayor parte de la energía. En el espectro de voz humana, para sonidos sonoros (con uso de las cuerdas vocales) en las señales de voz están presentes muchos formantes, no obstante para obtener una representación adecuada del tracto vocal los primeros tres formantes son esenciales, mientras que los de orden superior son progresivamente menos importantes. Los dos primeros Formantes llevan la mayor parte de la potencia del sonido lo que se hace evidente en el nivel o volumen. El tercer Formante posee un efecto relevante en la inteligibilidad; aspecto indispensable para la buena comprensión de los mensajes hablados. Los Formantes se ubican en todas las vocales y algunas de las consonantes. [1]

Predicción Lineal en el dominio del tiempo

En la teoría de predicción lineal [2] es ampliamente utilizado el Modelo para todo polo, conocido como modelo Autorregresivo. En este modelo (1), la señal s_n se da como una combinación lineal de los valores pasados y algunas entradas Un presentes.

$$s_n = -\sum_{K=1}^{P} a_K s_{n-K} + G U_n \qquad (1)$$

En donde G es un factor de ganancia y a_k los parámetros de predicción lineal, más conocidos como parámetros LPC. La obtención de estos parámetros es fundamental, ya que la voz se puede parametrizar a partir de éstos. Siguiendo con la fundamentación matemática, es posible obtener la función de transferencia H(z) del sistema, la que queda definida como:

$$H(z) = \frac{G}{1 + \sum_{K=1}^{P} a_K z^{-K}} \qquad (2)$$

Uno de los métodos para la determinación de los parámetros a_k es a través de la minimización del error, dado por la ecuación (4). En donde $\bar{s}_n$ (3) se obtiene considerando U_n (entrada) desconocido totalmente, lo cual es el caso en muchas aplicaciones. Por lo tanto, la señal s_n sólo puede predecirse en forma aproximada a partir de una sumatoria valorada linealmente de las muestras pasadas. Siendo $\bar{s}_n$ una aproximación de s_n, en donde

$$\bar{s}_n = -\sum_{k=1}^{P} a_k s_{n-k} \qquad (3)$$

Luego minimizando el error entre el valor real s_n y el valor predicho $\bar{s}_n$, se obtiene el siguiente sistema:

$$\frac{\partial E}{\partial a_i} = \sum_{k=1}^{p} a_k \sum_{n=0}^{N-1} s_{n-i} s_{n-k} + \sum_{n=0}^{N-1} s_{n-i} s_n = 0 \qquad i = 1\ldots\ldots p \qquad (4)$$

Las sumatorias que contienen s_n son conocidas como los coeficientes de autocorrelación y el sistema de ecuaciones es comúnmente resuelto por el algoritmo de Levison Durbin [2][3].

Estimación del Período Fundamental

En los sonidos sonoros las cuerdas vocales vibran y en los sonidos sordos las cuerdas vocales no vibran. Se define el Período Fundamental T_o, también llamado *Pitch* [1], como el tiempo transcurrido entre dos aperturas sucesivas de las cuerdas vocales. Las cuerdas vocales al vibrar producen un sonido tonal o periódico, de esto se desprende que los sonidos sonoros tienen *Pitch* y los sonidos sordos carecen de este. Para una secuencia real *s*[*n*], se define la autocorrelación de s[*n*] como:

$$r_{ss}(l) = \sum_{n=-\infty}^{\infty} s(n)s(n-l) \qquad (5)$$

Si *s*[*n*] es una secuencia periódica, de periodo T, la función de Autocorrelación $r_{ss}(5)$, es una secuencia periódica con período T. Esta característica es utilizada para obtener el período fundamental (*Pitch*) de señales de voz.

COEFICIENTES CEPSTRALES EN LA ESCALA DE FRECUENCIAS DE MEL

Los Coeficientes Cepstrales en la escala de frecuencias de Mel son más robustos que los coeficientes LPC y Cepstrales. Esto se fundamenta principalmente, porque estos coeficientes adaptan las frecuencias de fonemas a la manera que el oído humano percibe los sonidos [5] [6].
El cálculo de los MFCC se obtiene aplicando la siguiente ecuación.

$$\sum_{k=1}^{N} \left| X_i(k) \right| \cdot H(k.m) \qquad m = 1, 2, \ldots, M \qquad (6)$$

Donde $|X_i(k)|$ es la Transformada de Fourier de usada ventana de análisis, M es el número de bancos de filtros que se utilizan. Se debe tener en cuenta que *M*<<*N*. Los Bancos de Filtros en la Escala de Mel son una serie de filtros pasa-bandas triangulares, centrados en una frecuencia $f_c(m)$.
Una vez aplicado el Banco de Filtros en la Escala de Frecuencias de Mel a cada una de los coeficientes de *Fourier*, se deben calcular una serie de parámetros de transición denotados por $X'(m)$.

$$X'(m) = \ln\left(\sum_{k=1}^{N} |X_i(k)| \cdot H(k.m)\right) \qquad m = 1,2,\ldots,M \tag{7}$$

Finalmente se aplica a los parámetros de transición la Transformada Discreta del Coseno (DCT).

$$c(l) = \sum_{m=1}^{M} X'(m) \cdot \cos\left(\frac{l \cdot \pi}{M} \cdot \left(m - \frac{1}{2}\right)\right) \quad l = 1,2,\ldots,M \tag{8}$$

3.- DISEÑO CONCEPTUAL DEL SISTEMA

El sistema diseñado corresponde a una primera etapa de desarrollo de un sistema mayor, el cual permitirá el entrenamiento para el aprendizaje de palabras en forma independiente del hablante. Este sistema contempla el que el usuario (paciente con dificultades de audición) se entrene en la pronunciación de una sílaba, la cual es seleccionada apropiadamente desde un menú en una interfaz gráfica lo más amigable posible. Se considera que el aprendizaje de la pronunciación de fonemas es fundamental para el aprendizaje posterior de la pronunciación de palabras, partiendo de la base que las mismas están conformadas por fonemas, por lo que para un sistema mayor que identifique y sirva para el entrenamiento de palabras, se tendrá que considerar un sistema que previamente segmente las palabras en silabas, de una forma similar a la que se realiza en este trabajo, segmentando las sílabas en fonemas [7].
El diseño conceptual del sistema computacional para rehabilitación de trastornos del habla es esquematizado mediante el diagrama de flujo señalado en la figura 3.1. En el podemos visualizar cada una de las etapas principales de este sistema y a continuación se describe cada una de ellas.

Captura de señal de voz

En esta etapa el usuario, utilizando un micrófono, graba la pronunciación de una sílaba española. Esta grabación es almacenada y guardada en formato de audio WAV (PCM) para posterior análisis de la siguiente etapa. Los parámetros extraídos a la señal ingresada son posteriormente comparados con los parámetros correspondientes a fonemas patrones, previamente procesados y almacenados. La elección de los fonemas patrones ha sido consensuada por el equipo de trabajo.

Segmentación en 'N' fonemas

En este sistema las sílabas (señal de voz entrante) son segmentadas en N fonemas para lograr una adecuada comparación e identificación. Por ejemplo la sílaba /FA/ está compuesta por N=2 fonemas (/F/ y /A/).
Una vez individualizados los fonemas de una sílaba, el sistema extrae los parámetros característicos de cada fonema y a partir de estos es posible aplicar funciones estadísticas, en relación a los parámetros de fonemas patrones, para determinar el grado de exactitud en la pronunciación de las sílabas españolas.

Cálculo de Parámetros A

Los Parámetros A son el Periodo Fundamental (*Pitch*) y los Formantes, y su propósito es generar valores representativos que permitan la comparación de las vocales. Principalmente porque las vocales tienen la característica de de poseer claramente *Pitch* y Formantes.

Cálculo de Parámetros B

Los Parámetros B son los coeficientes LPC y los MFCC, y su propósito es generar valores representativos que permitan la comparación de las consonantes.

Cálculo Estadístico

Aplicación del coeficiente de correlación de *Pearson*

Una vez estimados los Parámetros del usuario, se está en condiciones de aplicar técnicas estadísticas para obtener índices que revelen el grado de cercanía con respecto a la pronunciación correcta de los fonemas (parámetros patrones). La función estadística utilizada es el Coeficiente de Correlación de *Pearson* y mide que tan cerca se encuentran los pares de variables a comparar de su recta de regresión lineal.
La función que rige el cálculo de este parámetro es la siguiente:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})}} \qquad (9)$$

Donde x_i e y_i corresponden a los parámetros de la señal patrón y señal entrante, respectivamente. $\bar{x}$ e $\bar{y}$ son el promedio de los parámetros patrones y parámetros de la señal entrante, respectivamente.

Promedio de identificación

En esta etapa los coeficientes de correlación son ponderados para determinar el Promedio de identificación (PI).

$$PI = 0{,}2 \cdot r_{LPC} + 0{,}2 \cdot r_{MFCC} + 0{,}3 \cdot r_{FO} + 0{,}3 \cdot rr \quad (10)$$

Donde r_{LPC}: Coeficiente correlación de los parámetros LPC,
r_{MFCC}: Coeficiente correlación de los MFCCs,
r_{FO}: Coeficiente correlación de los Formantes,
rr: Autocorrelación(utilizada en el cálculo de Pitch).

La ponderación total entre los coeficientes r_{LPC} y r_{MFCC}, que corresponden a los coeficientes de la consonante, es de un 40% y para los coeficientes de la vocal, r_{FO} y rr, un 60%. Esta diferencia de porcentaje radica que la vocal tiene un porcentaje mayor de participación en la extensión temporal de la sílaba en comparación a la consonante.
El Promedio de identificación (PI) varía entre 0 y 1. Mientras más cercano este a la unidad, mejor será la pronunciación del paciente en comparación a las sílabas patrones.

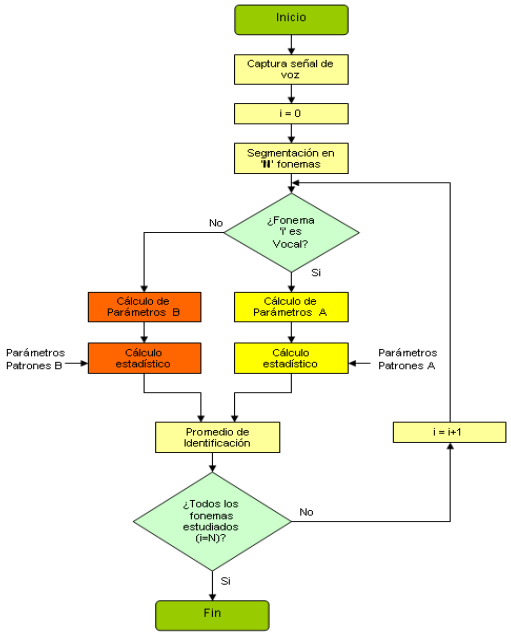


Figura 3.1.- Diagrama de flujo del sistema.

4.- SISTEMA COMPUTACIONAL

4.1. Ventana principal del sistema

La interfaz gráfica del sistema computacional es una única ventana, en la cual se encuentran distintas etapas, que representan los distintos procesos que se llevan a cabo. En la figura (4.1) muestra la ventana dicha ventana.

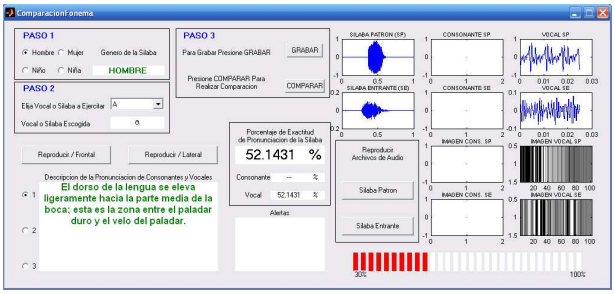


Figura 4.1. - Ventana principal del software de comparación de sílabas, a nivel de fonemas.

Para la utilización de este sistema de debe seguir la secuencia de pasos indicadas a continuación.

4.2 Paso 1: Elección de la base de datos

En el paso 1, se realiza la acción de escoger entre cuatro bases de datos, que corresponden a las sílabas patrones. Entre dichas bases de datos se definen cuatro grupos: Hombres, Mujeres, Niños y Niñas. Para clasificar dichos grupos, se consideraron los rangos de edades mostrados en la tabla 4.1.

Estos rangos de edades fueron seleccionados a partir de las edades de los individuos de pruebas utilizados durante todo el proceso de programación del software.

Tabla 4.1. - Rango de edades.

	Rangos de Edad
Niños y Niñas	6 a 15 años
Hombres y mujeres	16 a 50 años

La etapa 1 se encuentra representada en la interfaz según lo indicado en la figura 4.2.

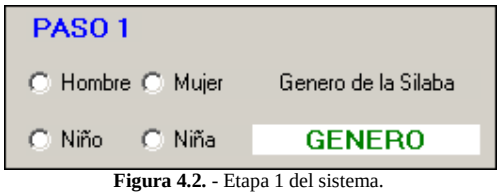


Figura 4.2. - Etapa 1 del sistema.

Como se puede apreciar, la elección de la base de datos es a través de un menú de punto. Se puede seleccionar sólo un tipo de base de datos a la vez, en el cual se complementa su elección, con un mensaje con letras verdes en mayúscula bajo un título “Genero de la Sílaba”.

4.3. Paso 2: Elección de la sílaba a entrenar

El paso 2 consiste en la elección de la sílaba a ensayar, esta puede ser elegida a través de una lista desplegable en donde se muestran cada una de las opciones disponibles para el entrenamiento del paciente. En la figura 4.3 se muestra el menú de dicha etapa.

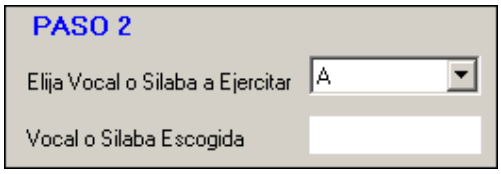


Figura 4.3. - Paso 2: Elección de la sílaba a entrenar.

Ya seleccionada la sílaba con la cual se quiere practicar, ésta es mostrada, por un motivo interno de las rutinas de programación, en un recuadro blanco, como se observa en la figura 4.3.

4.4. Paso 3: Grabación de la sílaba entrante y comparación entre sílaba patrón y sílaba entrante

En el paso 3, se hace uso de los tres conjuntos de parámetros enunciados anteriormente. En éste se graba la sílaba pronunciada por el usuario, y se compara (a nivel de fonemas) con la sílaba patrón escogida. En la figura 4.4 se muestra la interfaz gráfica de esta etapa.

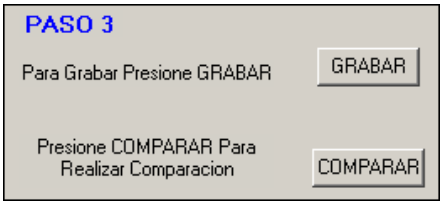


Figura 4.4. - Etapa de grabación y comparación de sílabas, a nivel de fonemas.

Inicialmente esta etapa permite grabar la pronunciación de una sílaba, dicha por el usuario. Una vez presionado el botón “GRABAR”, el software le entrega dos segundos al usuario para pronunciar la sílaba para ensayar.
La señal de voz grabada atraviesa por procesos que realizan las tareas de: verificar que el fonema haya sido pronunciado en un volumen adecuado, y por otro lado, la de acortar la señal grabada para mostrarla en un grafico, de manera de que el usuario pueda ver la señal en el dominio del tiempo y comprobar de manera visual que fue lo que se pronuncio. En caso de que el volumen de la sílaba se encuentre fuera de los rangos preestablecidos, el software muestra una advertencia para que la sílaba sea pronunciada nuevamente. Posteriormente, presionando en botón “COMPARAR” se acciona el procedimiento mostrado en la figura 3.1.

4.5. Visualización de resultados de la comparación de sílabas, a nivel de fonemas

Una vez hecha la comparación, a nivel de fonemas, entre la sílaba patrón y la sílaba entrante, la metodología mostrada anteriormente entrega un porcentaje final que representa la exactitud que existe entre las sílabas comparadas. Dichos porcentajes son mostrados, en la interfaz gráfica, en un recuadro mostrado en la figura 4.5.

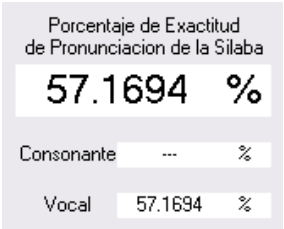


Figura 4.5. - Resultados de la comparación de sílabas, a nivel de fonemas.

Como se puede apreciar en la figura 4.5, el recuadro muestra los porcentajes por separado, de la comparación entre consonantes y vocales. En un recuadro más amplio se muestra el porcentaje final o Promedio de Identificación PI (ec. 10), el cual se obtiene ponderando ambos resultados nombrados.

4.6. Recuadro de audición

En el proceso de formulación del software se consideró que el usuario, la persona con deficiencias auditivas, puede utilizar el programa con cierta asistencia de alguna persona. Para aquella persona se dispuso un cuadro en el cual se pueden reproducir las sílabas, patrón y entrante, para que pueda apreciar la pronunciación de la sílaba ensayada, y dar algunas indicaciones que sirvan para mejorar dicha pronunciación. En la figura 4.6 se muestra el recuadro comentado.

Figura 4.6. - Recuadro de audio de sílabas.

Como se puede observar en la figura, existen dos botones claramente identificables los cuales reproducen la sílaba correspondiente al título mostrado en ellos.

4.7. Ayudas visuales

Una de las características principales del Sistema desarrollado, se basa en las ayudas visuales que se puedan entregar al usuario, de tal forma que a partir de la retroalimentación visual el pueda aprender, corregir o mejorar su aprendizaje del lenguaje hablado, es decir en la medida que mejoremos el canal visual se espera que permita un mejor *performance* del usuario.

4.7.1. Ayuda mediante texto

Este tipo de ayuda posee alguna utilidad si se considera que el usuario perdió su capacidad auditiva después de aprender el lenguaje. En caso contrario esta ayuda no tiene ninguna relevancia.
El tipo de ayuda definida en este punto consiste en describir como se pronuncia la consonante, o la vocal, en entrenamiento. En la tabla 4.2 se muestran algunas de las descripciones comentadas, siendo un total de 22 ayudas textuales.

Tabla 4.2. - Descripción de la pronunciación de letras.

Letra	Descripción
/a/	El dorso de la lengua se eleva ligeramente hacia la parte media de la boca; esta es la zona entre el paladar duro y el velo del paladar. La lengua extendida en el hueco de la mandíbula inferior toca con sus bordes los molares inferiores. La punta de la lengua roza la cara interior de los incisivos inferiores. Los labios forman una abertura mayor que la de todas las demás vocales. El velo del paladar permanece elevado. Las cuerdas vocales vibran.
/b/	El labio inferior hace contacto con el labio superior creando una oclusión completa que interrumpe la salida del aire. El velo del paladar permanece elevado. El aire aprisionado tras la oclusión escapa por la boca al abrirse los labios. Las cuerdas vocales vibran.
/c/, /k/, /q/	El postdorso de la lengua se eleva hasta tocar el velo del paladar creando una oclusión completa que interrumpe la salida del aire. La punta de la lengua desciende hasta las encías de los incisivos inferiores. El velo del paladar permanece elevado. El aire aprisionado tras la oclusión escapa por la boca. Las cuerdas vocales no vibran.

En la interfaz gráfica del programa existe un recuadro en donde se muestran tales descripciones. En la figura 4.7 se muestra el recuadro en cuestión. El recuadro posee un menú de puntos, los cuales se encuentran enumerados del 1 al 3, que permiten apreciar las descripciones, de la tabla 4.2, en tres etapas.

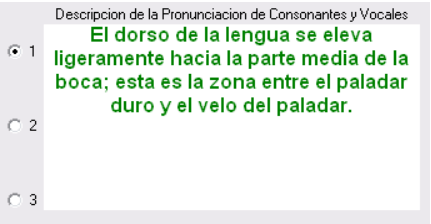


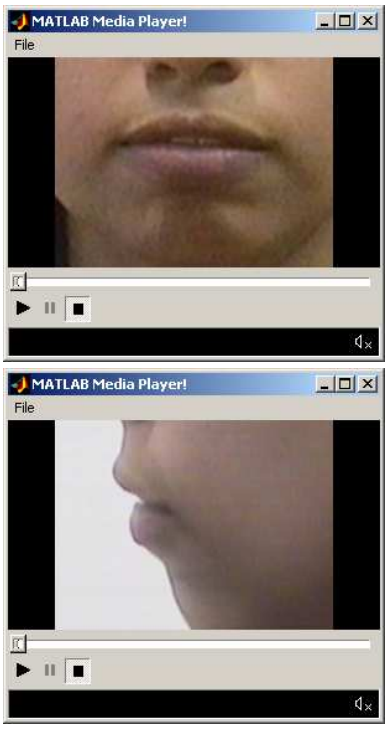
Figura 4.7. - Cuadro de descripción de letras.

.7.2. Ayuda multimedia

La principal ayuda visual, independiente del tiempo que lleva el paciente con su padecimiento, se basa en la reproducción de videos multimedia que muestren la pronunciación de la sílaba en aprendizaje. Los videos fueron obtenidos a partir de la grabación de una persona que no presenta trastornos del habla la cual pronunció cada una de las sílabas, que se pueden entrenar, definidas para este trabajo. La reproducción se observa en una ventana independiente, separada de la interfaz gráfica principal, y se reproduce a partir del uso de dos botones que se aprecian en la figura 4.8.

Figura 4.8. - Botones que permiten la reproducción de los videos.

Como se acaba de comentar, el video es reproducido en una ventana anexa. A continuación, en la figura 4.9.a y 4.9.b, se muestran dos ejemplos de un video de tipo frontal, y uno del tipo lateral.



(a)

(b)

Figura 4.9. - (a) Video de la reproducción frontal, (b) Video de la reproducción lateral.

4.7.3. Ayuda a través barras

Una ayuda extra, complementaria a las anteriores, es mostrar una fila de barras para mostrar el porcentaje final de exactitud de manera visual. En la figura 4.10 se muestra dicha ayuda gráfica.

Figura 4.10. - Muestra gráfica del porcentaje final de exactitud.

Como se puede observar, la fila de barras muestra un mínimo de 30% de exactitud hasta un 100%.

4.7.4. Ayuda a través de imágenes estáticas

Finalmente se consideró la ayuda visual a través de imágenes estáticas. Esta etapa toma como base la transformación de las muestras en el dominio del tiempo, de la señal de voz, en el dominio de la escala de grises. En la figura 4.11 se muestra un ejemplo con la sílaba /do/.
Como se puede ver en la figura se tienen un par de ventanas, en el dominio del tiempo, para la sílaba patrón (denotados por la sigla SP) y para la sílaba entrante (denotados por la sigla SE). Como también para el caso en el dominio de la escala de grises.

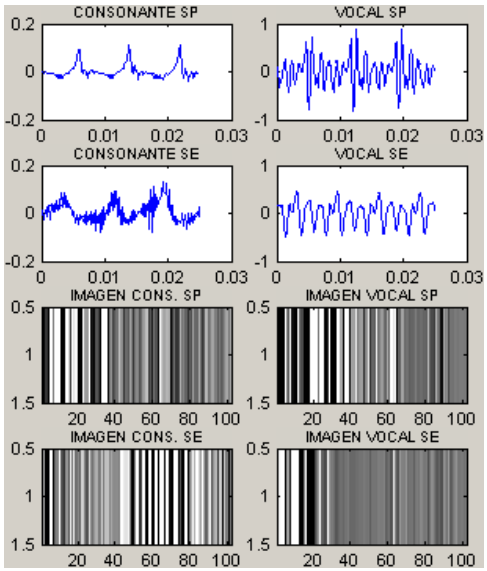


Figura 4.11. - Ayuda visual a través de imágenes.

4.8. Características finales

Como se ha dicho anteriormente, en la fase de grabación de la sílaba entrante, cuando ésta no se encuentra en un rango de volumen preestablecido, se muestra una advertencia que pide que se vuelva a pronunciar la sílaba. En la figura 4.12 se muestra el cuadro con el título “Alerta” que se muestra en la interfaz gráfica. Naturalmente en la etapa de inducción al software, un asistente deberá indicar al usuario el significado de esta alerta, dado que lo más común será encontrar usuarios que no saben leer.

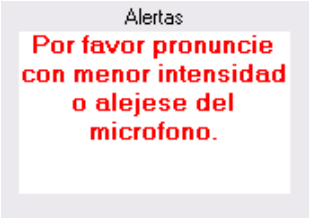


Figura 4.12. - Ventana donde se muestra cualquier advertencia programada.

En este recuadro, además, se realizan comentarios cuando el programa no puede realizar algunos de los procesos definidos en este trabajo. Entre los acontecimientos posibles de ocurrir podemos nombrar: "no se pudo graficar la sílaba entrante, no se pudo llevar a cabo la comparación, "falla".

5.- CONCLUSIONES

El sistema presentado es el resultado de 2 años de trabajo en la línea de procesamiento de voz para aplicaciones de Foniatría. Este sistema fue probado por usuarios sordos mudos quienes lo calificaron de muy buena forma, contando con su aprobación y expresando una gran satisfacción por este esfuerzo. Considerando el análisis del sistema y el párrafo anterior es posible afirmar que lo presentado en este trabajo da muchas expectativas a futuro, para la obtención de un sistema eficiente para el apoyo del aprendizaje de hablado de palabras para personas sordas o con dificultades de audición. Los resultados indican que el sistema desarrollado permite un adecuado adiestramiento en la pronunciación de sílabas, las que constituyen la base para un sistema mayor capaz de realizar un entrenamiento en el aprendizaje de palabras. Los resultados obtenidos respecto de la medida cuantitativa del porcentaje de acertividad o cercanía con la sílaba patrón deben ser trabajados y afinados aún más, con la participación de usuarios con este tipo de afección y fonoaudiólogos, además se deben incorporar el mayor número de fonemas posibles, ya que existen sílabas como "ji" y "fi" que aún son difícil de distinguir, por lo que es recomendable utilizar otros métodos para su diferenciación. Por otra parte, el equipo de investigadores se encuentra trabajando en modelos que incorporan estructuras en base a redes Neuronales y Transformadas de Wavelet, de tal forma de obtener modelos más robustos de identificación de fonemas. Respecto de sistema desarrollado, la elección desde una base de datos del fonema para entrenamiento, proporciona una forma simple para ingresar la sílaba por parte del usuario, en donde dicho usuario puede apreciar lo que pronunció en forma de: señal en el dominio del tiempo, imágenes y mensajes visuales. Por tal motivo es que se implementaron ventanas de alertas y ayudas, las cuales muestran comentarios y explicaciones de los acontecimientos ocurridos durante la comparación. En resumen, se logra implementar un sistema complejo, en una forma simple y didáctica, que permite una interacción amigable para el usuario en rehabilitación. Finalmente es importante insistir que lo presentado es la primera aproximación a un sistema mayor, el cual deberá ser capaz de identificar palabras. Sin embargo, se espera que el sistema presentado para fonemas, constituya una herramienta de apoyo inicial a la rehabilitación de trastornos del habla, reemplazando el canal de realimentación auditivo por un canal de realimentación visual. Es decir, mediante gráficas de los perfiles acústicos y principales parámetros característicos de la señal de voz (Formantes, LPC y MFCC), con imágenes e indicadores de avance, permitan en conjunto una importante herramienta alternativa adicional para la rehabilitación de trastornos del habla.

6.- REFERENCIAS

[1] Faúndez, Marcos, Tratamiento Digital de Voz e Imagen, Alfaomega, México, 2001.
[2] Makhoul, J. "Linear Prediction: A Tutorial Review" Proc. IEEE. 1975
[3] Rabiner y Schafer "Digital Processing of Speech Signals" Prentice Hall. Englewood Cliffs, N.J. 1978
[4] B.S. Atal y S.L. Hanauer, Speech analysis and synthesis by linear prediction of the speech wave, J. Acoust. Soc. Amer., vol. 50, n°.2, pp. 637-655,1971.
[5] Shimamura, T.y Kobayashi, H., Weighted Autocorrelation for Pitch Extraction of Noisy Speech, IEEE Transactions on Speech and Audio Processing, Vol.9, No.7, pp. 727-730, Oct. 2001.
[6] Sigurdsson, Brandt y Lehn-Schiler, Mel Frequency Cepstral Coefficients An Evalution of Robustness of MP3 Encoded Music, Informatics and Mathematical Modelling Technical University of Denmark, 2006.
[7] San Juan Enrique, "Segmentación de sílabas en fonemas" Congreso Internacional de Telecomunicaciones, Senacitel 2008, Valdivia, Chile.
[8] Rabiner y Biing-Hwang Juang, Fundamentals Of Speech Recognition , Prentice Hall, Englewood Cliffs, N.J., 1993.
[9] Quilis, Antonio, Fonética Acústica de la Lengua Española, Gredos, Madrid,1988.
[10] Parsons, Thomas, Voice and speech processing, McGraw-Hill, New York, 1987.
[11] Rabiner y Schafer, Digital Processing of Speech Signals, Prentice Hall, Englewood Cliffs, N.J., 1978.
[12] Delores, M. Etter, Solución de problemas de Ingeniería con MatLab, 2a edición, Prentice Hall, México, 1998.
[13] Signal Processing Toolbox. For use with Matlab. User's, Guide version 5. The Math Works Inc. 2000.