RECONOCIMIENTO DE CARACTERÍSTICAS VOCALES ENFOCADO A LA IDENTIFICACIÓN DE HABLANTES

Katherine García Cruz
Universidad de San Buenaventura Bogotá
katherine.gc@hotmail.com

Marcelo Herrera Martínez
Universidad de San Buenaventura Bogotá
mherrera@usbbog.edu.co

Andrea Lorena Aldana Blanco
Universidad de San Buenaventura Bogotá
aaldana@usbbog.edu.co

(Tipo de Artículo: Investigación. Recibido el 18/11/2014. Aprobado el 05/12/2014)

RESUMEN
El siguiente documento describe el desarrollo de un software de análisis de parámetros acústicos de la voz (APAVOIX), que puede ser utilizado con énfasis en acústica forense, basado en el reconocimiento e identificación de hablantes. Este software permite observar de manera clara los parámetros de la voz suficientes y necesarios en el momento de realizar una comparación entre dos señales de voz, una dubitada y una indubitada. Estos parámetros se utilizan de acuerdo al método combinado clásico, utilizado generalmente por las entidades del estado para realizar cotejos de voz.

Palabras clave. Software, Identificación de hablantes, Parámetros Acústicos, Voz, Procesamiento Digital de Señales.

VOICE FEATURE RECOGNITION FOR SPEAKER IDENTIFICATION

ABSTRACT
This document describes the development of software for voice feature analysis (APAVOIX), intended for voice recognition and speaker identification tasks in forensic acoustics field. This software allows observing clearly important voice features when comparing two voice recordings; one that is known and other unknown. These features are based on the classic combined method, which is generally used by national security organizations for voice comparison.

***Keywords**. Software, Speaker Identification, Acoustic Features, Voice, Digital Signal Processing.*

Reconnaissance des caractéristiques vocales pour l'identification des locuteurs

Résumé
Cet article présente le développement d'un logiciel d'analyse des paramètres acoustiques de la voix (APAVOIX), qui peut être utilisé avec emphase dans juri-acoustique, en se basant sur la reconnaissance et l'identification des locuteurs. Ce logiciel permet d'observer clairement les paramètres de la voix suffisants et nécessaires au moment de réaliser une comparaison entre deux signaux de voix, une connue et une inconnue. Ces paramètres sont utilisés d'après la méthode combinée classique qui est utilisée couramment par les institutions de l'état pour réaliser confrontations des voix.

***Mots-clés.** Logiciel, Identification des locuteurs, Paramètres acoustiques, Voix, Traitement numérique de signaux.*

1. INTRODUCCIÓN

La acústica forense es una rama, en su mayoría desconocida en el campo de la ingeniería de sonido, debido a que el contacto que se tiene con ella no es muy fuerte a lo largo del desarrollo del programa; sin embargo es una rama completamente necesaria y ampliamente utilizada para realizar cotejos o comparaciones entre dos o más señales de voz, con el fin de comprobar la culpabilidad o inocencia de un implicado en algún proceso judicial. Para realizar este trabajo, los peritos responsables de cada caso, requieren una herramienta que les permita hacer una comparación correcta de ciertos parámetros que logran dar una caracterización específica a la voz de cualquier hablante, ya que esta actúa de la misma forma que lo haría una huella dactilar, es decir que dos voces no pueden ser iguales, debido a las características fisiológicas de cada hablante. Cada una de estas comparaciones se realiza mediante el método combinado clásico, que es el resultado de años de investigación en el campo de la acústica forense, para tener un referente exacto de los elementos a analizar y los resultados que pueden llegar a obtenerse en la comparación de las voces.

Este método es de cierta manera analizado cualitativamente, sin valores cuantitativos que permitan darle más peso a las grabaciones de voz en un estrado, sin embargo, si es posible cuantificar algunos de estos valores, para dar más precisión a un veredicto, pero los peritos encargados de esta labor no tienen un conocimiento tan amplio y específico en cuanto al procesamiento digital de señales, y a otros parámetros cuantificables, por lo que es necesario contar con un profesional con este tipo de conocimientos, para realizar un trabajo conjunto y poder utilizar con mayor poder muchas de estas pruebas, y que seguramente permitirían que los dictámenes de un juez sean lo más justos posible, y que las personas implicadas en procesos judiciales puedan comprobar su inocencia o culpabilidad, y así tomar las medidas respectivas, con un porcentaje más alto de confiabilidad.

2. DESARROLLO DEL ARTÍCULO

Con el fin de desarrollar un software de análisis de parámetros acústicos de la voz, es necesario conocer acerca de los fenómenos ocurridos en la generación del habla, así como los parámetros que pueden ser utilizados para la realización de los cotejos de voz en el campo de la acústica forense.

2.1. Conceptos generales de la voz

La voz es un conjunto de sonoridades producidas por el funcionamiento de los órganos de la fonación.
El instrumento de la voz comprende:

- El aparato respiratorio es el motor que proporciona al sonido la intensidad, la fuerza, la potencia y el sostenimiento. Está comprendido por los pulmones, el fuelle y el depósito de aire. El aparato respiratorio se divide en dos partes:

  - Vías respiratorias superiores: constituidas por las fosas nasales, la faringe nasal, y los senos o cavidades accesorias es la primera parte del trayecto que debe efectuar el aire, el cual debe penetrar por las aletas nasales.

  - Las vías respiratorias inferiores están constituidas por la laringe, la tráquea, los bronquios y los pulmones. El pulmón es el órgano esencial de la respiración. El aire se almacena en los alvéolos pulmonares.

- El órgano vocal vibrante es el generador del sonido, y le proporciona la altura por las vibraciones de las cuerdas vocales. Está compuesto por la laringe con la glotis, las cuerdas vocales y los ventrículos. La laringe es el órgano donde nace el sonido y se encuentra tapizada por una membrana mucosa, provista de glándulas secretoras. En medio de la laringe hay una región llamada glotis, que está constituida por las cuerdas vocales que son dos bandas móviles que, unidas en su parte anterior, dejan entre sí un espacio triangular que es la glotis. Para determinar la apertura o cierre de la glotis existen los músculos tensores y constrictores respectivamente. el músculo de las cuerdas vocales tensa las cuerdas vocales, llamado también tiroaritenoides.

- El sistema de resonancia es el que proporciona al sonido el timbre, el color y la riqueza armónica. Es el reforzamiento del sonido. También permite la colocación de la voz y el alcance. Está compuesto por los resonadores y las cavidades de resonancia. Este puede dividirse en:

  - Partes duras fijas: son las partes óseas: el maxilar superior, los huesos de las fosas nasales, de los senos y de la bóveda palatina ósea, y los dientes. Estas partes son duras, rígidas y fijas. Para que favorezcan la resonancia es necesario que sean lisas y parejas. Si hay vegetaciones en la nasofaringe, pólipos en las fosas nasales, líquido o pus en los senos, un a mucosa espesa, la voz será sorda y la resonancia será mal producida.

  - Partes blandas móviles: Son las paredes musculo-membranosas de la faringe: el velo del paladar blando, la lengua, las mejillas y los labios. Pero existe un hueso móvil, el maxilar inferior. Estas partes deben estar sanas, libres, y ser bien móviles. Si hay una amígdala lingual voluminosa o amígdalas aumentadas de

volumen, los movimientos de la lengua o del velo del paladar serán trabados, dificultados, y, sobre todo, esas masas constituirán por su volumen un obstáculo a la salida de los sonidos. La colocación de la voz será defectuosa, disminuirá la resonancia y el alcance será menor [1].

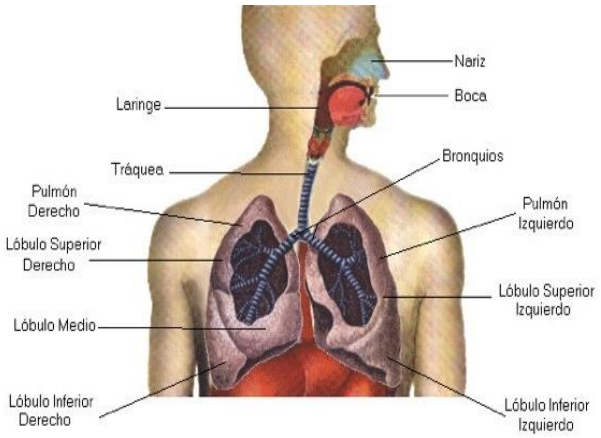


Fig. 1 Aparato Fonador [2]

2.2. Formantes y Espectrograma

Dado que el tracto vocal evoluciona en el tiempo para producir los distintos sonidos, la caracterización espectral de la señal de voz también es variante en el tiempo. Esta evolución temporal puede representarse mediante un espectrograma de la señal de voz o sonograma. Esta es una representación bidimensional que muestra la evolución temporal de la caracterización espectral.

Los formantes aparecen como franjas horizontales, mientras que los valores de amplitud en función de la frecuencia se representan en tonalidades de grises en sentido vertical.

Existen dos tipos de sonogramas: de banda ancha y de banda estrecha, en función del ancho de banda del filtro que se haya utilizado para realizar el análisis frecuencial. En el caso de los espectrogramas de banda ancha, se obtendrá una buena resolución temporal. En cambio, en el caso de espectrogramas de banda estrecha se obtendrá una buena resolución frecuencial, dado que el filtro utilizado es de banda estrecha, y permite obtener estimaciones espectrales más precisas. [1]

2.3. Análisis localizado en el dominio de la frecuencia

Un análisis convencional en el dominio frecuencial para una señal de voz aporta poca información dadas las especiales características que tiene dicha señal. En el espectro de una señal de voz aparecen dos componentes convolucionadas. Una es la proveniente de la frecuencia fundamental y sus armónicos y la otra de los formantes del tracto vocal. Por ello, la única información que nos puede aportar es desde el punto de vista visual y a través de espectrogramas. Dependiendo del tipo de espectrograma que usemos (de banda ancha o de banda estrecha) podremos observar una de las dos componentes en mayor detalle (los formantes o la estructura fina respectivamente).

a. Transformada rápida de Fourier

La transformada rápida de Fourier (FFT) basa su principio de funcionamiento en la reorganización de la señal que debe ser potencia de 2, como condición para la operación. Al igual que la TDF, en la FFT la señal transformada en el dominio de la frecuencia debe descomponerse en una serie de senos y cosenos representada por números complejos. Una señal de 16 puntos debe ser descompuesta primero en 4, luego en 2 y así sucesivamente hasta que la señal está punto a punto. Esta operación no es más que una reorganización de la señal para disminuir el número de operaciones y mejorar su velocidad. Luego de encontrar la frecuencia de los espectros, estos deben ser reorganizados para encontrar la operación en el dominio del tiempo. [3]

2.4. Análisis en el dominio Cepstral

El cepstrum (/kepstrum/), o coeficiente cepstral, c (τ), se define como la transformada inversa de Fourier del logaritmo del módulo espectral | X (ω) |

$$c(T) = IDFT[\log|X(w)|] \quad (1)$$

El término "cepstrum”• se deriva de la palabra inglesa “spectrum” (espectro), para dar una idea del cálculo de la transformada inversa del espectro. La variable independiente en el dominio cepstral se denomina (siguiendo la misma lógica) “quefrency”.
Dado que el cepstrum representa la transformada inversa del dominio frecuencial, la “quefrencia” es una variable en un dominio pseudotemporal.
La característica esencial del cepstrum es que permite separar las dos contribuciones del mecanismo de producción: estructura fina y envolvente espectral [4].

2.5. Análisis por Codificación Lineal Predictiva LPC

La predicción lineal es una buena herramienta para análisis de señales de habla. La predicción lineal modela el tracto vocal humano como un sistema de respuesta al impulso infinita (IIR) que produce la señal de voz. Para sonidos vocales y otras regiones con voz, que tienen una estructura resonante y un alto grado de similaridad sobre cambios de tiempo que son múltiplos de su periodo tonal, este modelo predice una representación eficiente del sonido.

Para utilizar la codificación lineal predictiva se tiene en cuenta que:

- Las vocales son más fáciles de reconocer utilizando LPC
- El error puede ser computado como a^T Ra donde R es la matriz de autocovarianza o autocorrelación de un segmento y a es el vector de coeficientes de predicción de un segmento estándar.
- Un filtro de pre-énfasis antes del LPC, enfatizando frecuencias de interés, puede mejorar el desempeño.
- El periodo del tono en hombres es diferente al periodo del tono en las mujeres.
- Para segmentos con voz (r_ss [T])/(r_ss [0])≃0.25 donde T es el periodo del tono [5].
- Para las consonantes el modelo LPC no es adecuado, debido a que la aleatoriedad de éstas, impide su predicción.

2.6. Tono (Pitch)

La información prosódica, es decir, la velocidad de entonación, se encuentra fuertemente influenciada por la frecuencia fundamental de vibración de las cuerdas vocalesf_0, cuyo inverso a su vez, se conoce como periodo fundamental T_0. En forma general, la definición de periodicidad está dada en intervalos de análisis infinitos, sin embargo, por practicidad, su estimación se realiza sobre intervalos finitos, de manera que permitan cubrir varios periodos del pitch, o instantáneamente a partir de la diferencia entre dos momentos consecutivos del cierre glótico [6].

2.7. Relación Señal-Ruido

La relación S/N proporciona una medida de la calidad de una señal en un sistema determinado y depende, tanto del nivel de señal recibida como del ruido total, es decir, la suma del ruido procedente de fuentes externas y el ruido inherente al sistema. En el diseño de sistemas, se desea que la relación señal a ruido tenga un valor tan elevado como sea posible. Sin embargo, en el contexto de cada aplicación particular, este valor puede variar, ya que por lo general, el obtener altos valores de S/N conlleva un aumento, a veces considerable, en el costo de implementación del sistema. Un valor adecuado de esta relación es aquél en el que la señal recibida puede considerarse sin defectos o con un mínimo de ellos [7].

3. DESARROLLO INGENIERIL

Para el diseño del software fue necesario realizar una exploración previa de los parámetros necesarios y comúnmente utilizados en la identificación de hablantes, con énfasis en acústica forense, por lo cual se tomó como base la referencia del software Computerized Speech LabCSL4500 de Kaypentax® y se tomaron las características que este proporciona al usuario, para realizar un software propio, cuyo nombre es APAVOIX (Análisis de Parámetros Acústicos de la Voz).
En el momento de desarrollar este software, la interfaz gráfica se realizó a la par con el software, utilizando la herramienta de diseño de interfaz gráfica guide en Matlab®.

Para el diseño del espectrograma, se contó con una gran ventaja que fue el toolbox para matlab de procesamiento de voz "VOICEBOX"[1] que contiene varias herramientas, para analizar los parámetros acústicos de la voz. De este toolbox se hizo uso de la función spgrambw, que permite graficar un espectrograma en matlab, utilizando un algoritmo de mayor precisión, que el que podría obtenerse utilizando la función specgram de matlab, propia del toolbox de procesamiento de señales. Para utilizar esta herramienta, el programa requiere utilizar el audio a analizar, que en este caso, es el fragmento seleccionado por los cursores, la tasa de muestreo, que en este caso siempre es 11025 Hz, debido a que para procesamiento de señales de voz, no es necesario utilizar una frecuencia de muestreo muy alta; el ancho de banda del espectrograma para determinar el equilibrio entre la resolución del tiempo y la frecuencia, el rango de frecuencias mínima y máxima, así como la resolución en la que se graficará. Además de esto, la función permite modificar la configuración de colores del espectrograma, para hacer más clara la visualización, modificar la escala en la que se quiere visualizar el espectrograma, ya sea en Hertz, Logaritmo, Erb, Mel y Bark y permite ver en la parte de arriba del espectrograma, una gráfica con la forma de onda que se está analizando.

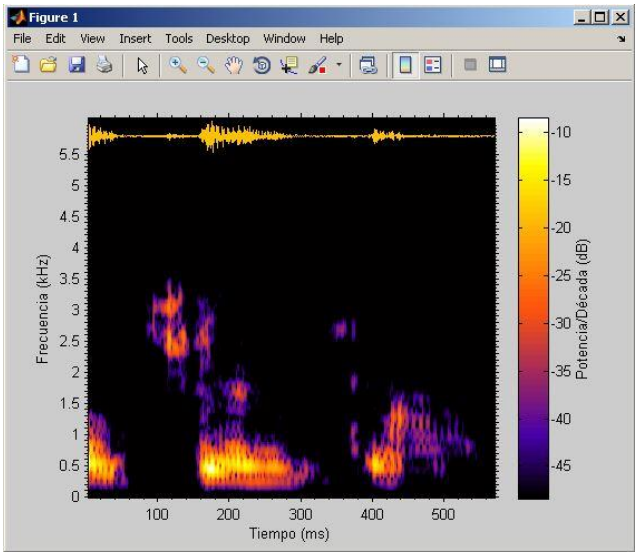


Fig. 2 Gráfica del Espectrograma

Para la transformada rápida de Fourier, inicialmente se toma el fragmento de audio seleccionado, que debe ser

[1] Brooks, Mike, <<VOICEBOX Speech Processing Toolbox for Matlab>> http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (consultada en Abril de 2012)

pequeño, como puede ser una vocal, debido a que este análisis requiere ser llevado a cabo en una porción muy pequeña del audio, pero que contenga información relevante, como son los formantes de la señal, presentes en las vocales. Posteriormente se aplica un filtro de pre-énfasis, que es opcional y puede ser modificado por los parámetros del software, en un rango de 0 a 1.5, luego se utiliza la función fft para obtener las componentes de la transformada rápida de Fourier, para ser graficadas, aplicando además el ventaneo de la señal, que puede ser Hamming, Hanning, Blackman, rectangular o triangular, además de utilizar el tamaño de la muestra que por defecto es de 8192 puntos, pero puede configurarse para la cantidad de puntos establecida en el software, y luego de esto se limita las frecuencias en las cuales se quiere graficar. Este parámetro también es modificable. Luego se toma el valor mínimo y máximo de frecuencia en el que se graficará y junto con los valores en dBu de la energía en cada valor frecuencias, se realizará la gráfica respectiva. Adicional a esto, en la parte superior, aparecerá una gráfica con el fragmento de la forma de onda que se está analizando.

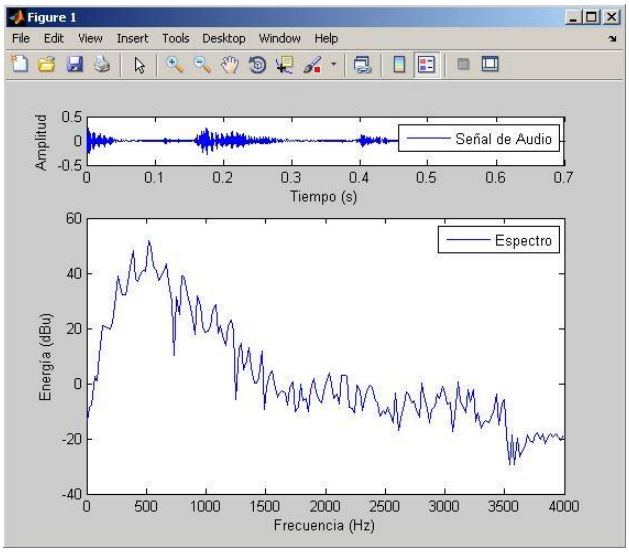


Fig. 3 Gráfica de la FFT

La codificación linear predictiva, se realizó, al igual que la FFT, tomando un fragmento de la señal con una vocal, y aplicando el filtro pre-énfasis opcional. A continuación, se utilizó la función spLpc, de un toolbox, llamado sptoolbox[2] que permite obtener los coeficientes LPC con el fragmento de la señal elegida, la frecuencia de muestreo (11025 Hz) y el orden del filtro que varía por múltiplos de 2, desde el 2 hasta el 36, después de esto, además puede modificarse el tipo de ventana que se desea para realizar el análisis y elegir el tamaño de la muestra sobre la que se realizará el análisis, y con estos valores, se obtiene la gráfica respectiva, acompañada de la gráfica del fragmento a analizar.

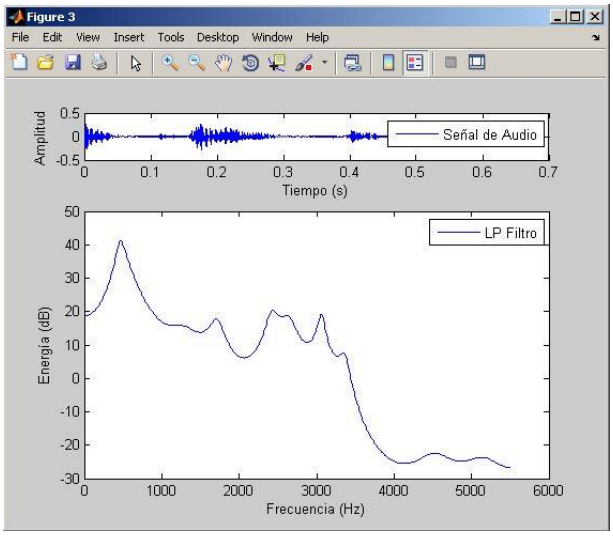


Fig.4 Gráfica LPC

Para el funcionamiento del Cepstrum, se tienen en cuenta las características mencionadas anteriormente en el funcionamiento de la FFT, debido a que la función de cepstrum parte de la transformada inversa de Fourier del logaritmo del módulo espectral. Para calcular el cepstrum, se tomó de igual manera un fragmento de una señal con una vocal, a la que posteriormente se realizó el pre-énfasis, y al igual que en la FFT, se obtuvieron los valores de la transformada rápida de Fourier, aplicando igualmente la opción de ventaneo de la señal. A los valores obtenidos de este cálculo, se aplica el logaritmo de la señal para posteriormente aplicar de nuevo la transformada rápida de Fourier, así se obtienen los valores del cepstrum. Como el cepstrum está dado en valores de "quefrencia" que son valores pseudotemporales en milisegundos, se obtiene el valor menor y el mayor al que se va a graficar a partir de los valores de las frecuencias en las cuales se analizará la voz y la frecuencia de muestreo. Con el valor de la "quefrencia" y el valor del cepstrum, es posible realizar la gráfica de este parámetro.

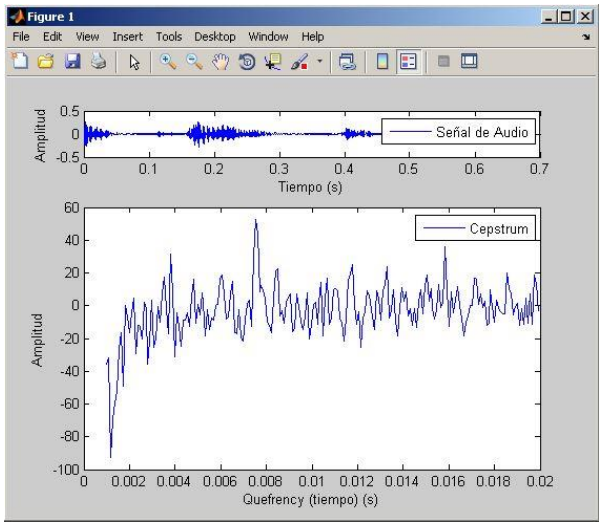


Fig. 5 Gráfica del Cepstrum

[2] Seo, Naotoshi, <<Project: Pitch Detection>> http://note.sonots.com/SciSoftware/Pitch.html (consultada en Septiembre de 2012)

Para el cálculo del pitch, se utiliza la función spPitchTrackCepstrum del sptoolbox, en el que se requiere inicialmente un fragmento del audio, la frecuencia de muestreo (11025 Hz), la longitud de la trama en milisegundos, la superposición de la trama y el tipo de ventana utilizada para calcular el tono. Esta función permite hacer un cálculo del tono que prevalece en la señal, a través del cepstrum, detectando la frecuencia principal que se percibe, de acuerdo a la "quefrencia" para cada trama y que posteriormente es graficado, esto debido a que el cepstrum calcula la frecuencia fundamental en la que se presenta la voz. Debe tenerse en cuenta que si se toma una palabra completa, que requiere de análisis, se verá una mayor variación en el tono, de acuerdo a la consonante que se pronuncia y su modo de pronunciación (oclusiva, fricativa, africada, nasal, espirante, lateral o vibrante), y que el tono se mantiene de cierto modo constante, en una vocal determinada.

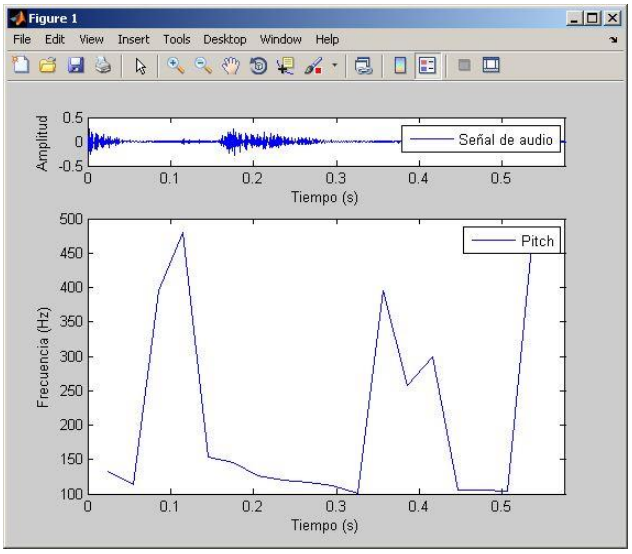


Fig. 6 Gráfica del Pitch por Cepstrum

Para el cálculo del pitch por auto-correlación, se utiliza la función spPitchTrackCorr del sptoolbox, mencionado anteriormente, y que para su ejecución requiere el fragmento de la señal a analizar, la frecuencia de muestreo (11025 Hz), el tamaño de la trama y la superposición de la trama. Este método permite obtener la información de cuan repetitiva es la señal respecto a si misma, por lo cual el valor del pitch al repetirse, no con valores exactamente iguales para las vocales, pero si cercanos, genera un mayor valor en la correlación, se toman los mayores valores y se calculan las distancias relativas entre los valores y luego el máximo común divisor de estas distancias corresponde con el valor del periodo fundamental. Este también tiene periodos de grandes variaciones, por las consonantes y las características en su emisión.

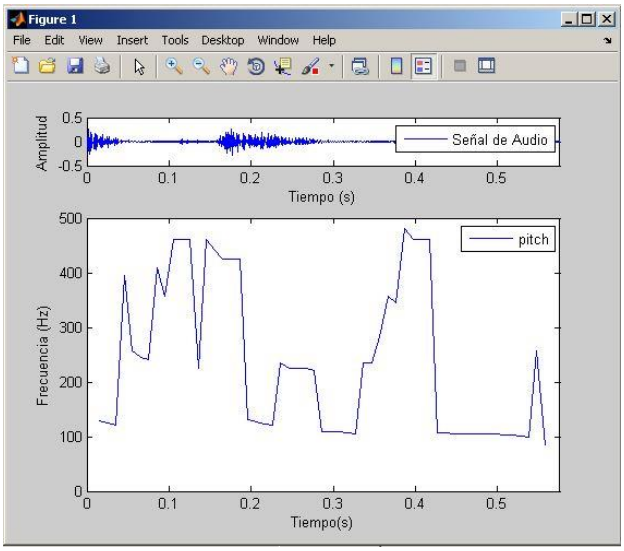


Fig. 7 Gráfica del pitch por Auto-Correlación

Para calcular el pitch robusto, se ha realizado un filtro robusto de estimación del tono para altos niveles de ruido (PEFAC) y se ha realizado el cálculo, mediante la función fxpefac, perteneciente al toolbox llamado voicebox, mencionado anteriormente. Este tipo de estimación únicamente se hace presente en las partes en las cuales se encuentran vocales, y no se muestra el comportamiento de las consonantes, por lo que puede observarse más claramente el comportamiento del tono.

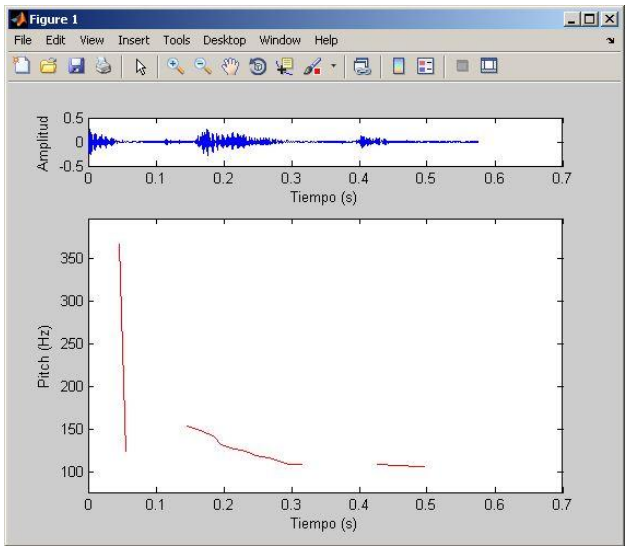


Fig. 8 Grafica de pitch robusto

Para graficar los formantes, fue necesario utilizar la función spFormantsTrackLpc, que pertenece al sptoolbox. Esta función se realiza utilizando el método LPC. Para su funcionamiento, requiere del fragmento de señal a analizar, la frecuencia de muestreo (11025 kHz.), el orden del filtro, el tamaño de la trama, la superposición y el tipo de ventana. Inicialmente se obtienen los valores de los formantes, utilizando el LPC, posteriormente se divide el rango de frecuencias en cinco partes, calculando así cada formante separadamente y permitiendo modificar los puntos,

para observar los colores diferentes por formante. Así como en la estimación del pitch por correlación y cepstrum, en el caso de los formantes, también pueden verse inclusive los que se hacen presentes en el espacio de las consonantes, y que tienen una forma irregular, no definida, por la razón aclarada anteriormente, por eso es necesario fijarse en su mayoría en las formantes pertenecientes a las vocales.

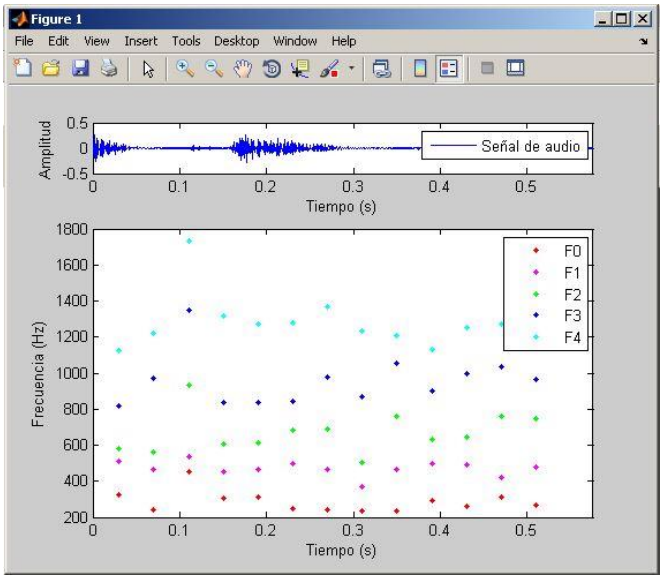
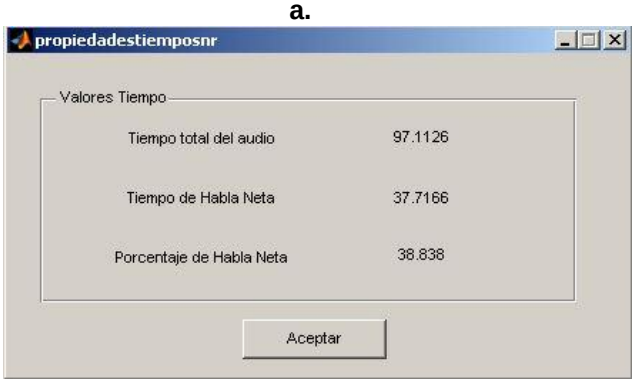


Fig. 9 Gráfica de los formantes

Para obtener la relación señal-ruido, se utilizó una función llamada snreval[3], hace parte de una colección de funciones, que calculan algunas medidas objetivas para la calidad del habla. Esta función permite observar gráficamente en toda la señal, los puntos en los cuales se presenta el habla y los puntos en los que no se encuentra información relevante. Además de esto, obtiene los valores de relación señal ruido Wada y Stnr. Lo único que requiere esta función es ingresar el nombre y la dirección del archivo elegido para realizar el cálculo. Esta función requirió de ciertas modificaciones para funcionar de acuerdo a las necesidades de APAVOIX, además de permitir el cálculo de habla neta en la señal.

a.



b.

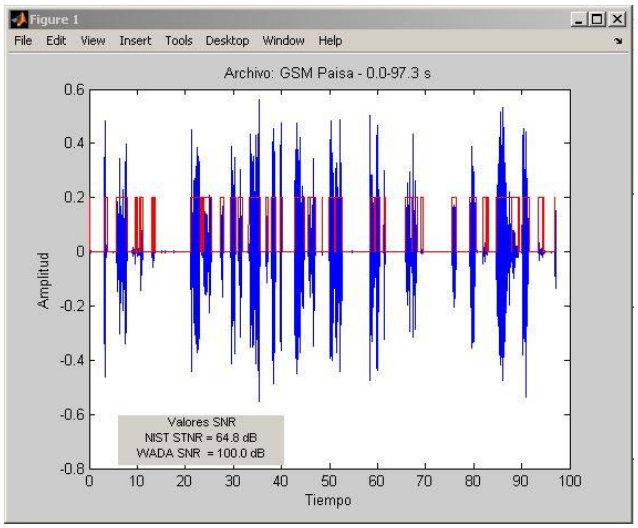
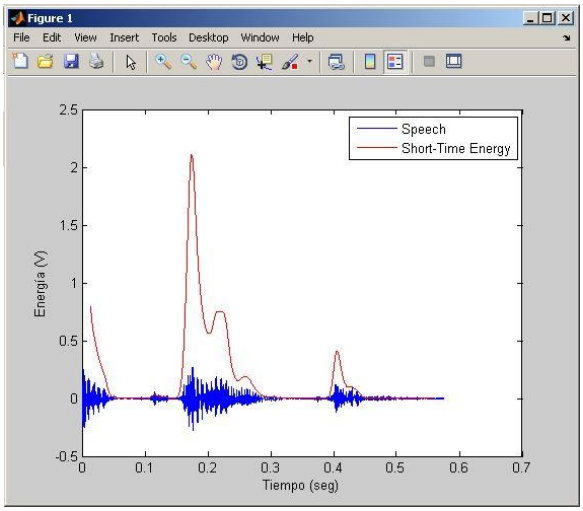


Fig. 10 (a, b) Resultados S/N a. Propiedades del tiempo de habla neta b. Gráfica de actividad de voz

En el momento de obtener la energía de la señal, inicialmente se toma el fragmento de voz q se ha elegido, y se procesa por medio de ventaneo utilizando Hamming, luego de esto se realiza la sumatoria del valor absoluto de la señal, elevado al cuadrado, q es el cálculo de la energía y a partir de allí se obtienen las gráficas que muestran el comportamiento de la energía en la señal de audio, junto con la señal a ser analizada.

a.



[3] Ellis, Dan, <<Objective Measures of Speech Quality/SNR>> http://labrosa.ee.columbia.edu/projects/snreval/ (consultada en Octubre de 2012)

b.

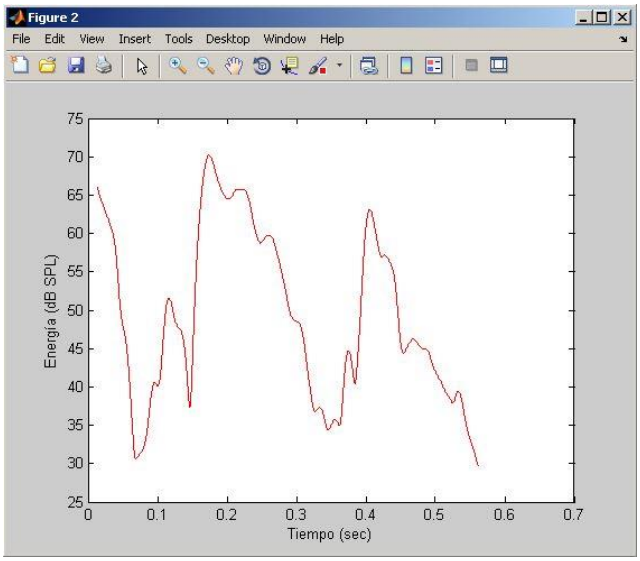


Fig. 11 (a,b) Gráfica Energía de la señal a. Voltaje b. dB SPL

Para obtener el género, el método utilizado es la autocorrelación, mediante la cual, se toma una parte de la señal, se limitan las frecuencias máximas y mínimas del habla, que se encuentran entre 50 y 500 Hz, y se realiza la autocorrelación de la señal, a partir de ella se busca la región que contenga los puntos máximos y se encuentra la frecuencia principal (como se realizaría normalmente en el procesamiento del tono visto anteriormente). Si el valor que se obtiene se encuentra en un rango de 80 a 190 Hz, detectará el género masculino, si se encuentra de 190 a 255 Hz, detectará el género femenino, y si el valor no se encuentra en ninguno de los dos parámetros, mostrará que no es posible reconocer y que debe elegirse otra parte del audio.

APAVOIX cuenta con una interfaz gráfica que permite manejar fácilmente las funciones requeridas en el análisis del habla. Cuenta con un manual de instrucciones (Anexo A), que explica detalladamente su funcionamiento. Además de esto, permite cargar el archivo deseado, o grabarlo y guardarlo, permite ubicar cursores para elegir la parte de la señal que se desea analizar. Estos cursores se obtuvieron mediante la función dualcursor[4], que además pueden ser actualizados después de ubicarse en la posición deseada y pueden ser guardadas sus posiciones como archivos .mat, para ser cargados y utilizados posteriormente cuando se requiera continuar el análisis.

Finalmente, cuenta con una barra de herramientas, con las que puede hacerse zoom, mover la señal, poner los cursores, reproducir toda la señal y reproducir la señal que se encuentra entre los cursores.

[4] Hirsch, Scott, <<Dualcursor>> http://www.mathworks.com/matlabcentral/fileexchange/2875 (consultada en Mayo de 2012){

4. CONCLUSIONES

El software APAVOIX diseñado y tratado a lo largo de este proyecto, es un buen comienzo para comprender el procesamiento digital de las señales de voz, como se planteó en el inicio del desarrollo del proyecto, el software es capaz de mostrar las características suficientes y necesarias a la hora de trabajar en cotejos de voz, como son:

- La frecuencia fundamental, cuyos resultados, según se pudo comprobar, varían de acuerdo a la emisión vocal que se esté evaluando, ya sea la pronunciación de una consonante o una vocal.
- La intensidad de la señal.
- El espectro y la densidad espectral de energía de la señal, mediante periodograma y método de Welch.
- El código lineal predictivo (LPC), el cual no es adecuado para la predicción de consonantes (a diferencia de las vocales), debido a la naturaleza aleatoria de las mismas.
- Los formantes, que, de la misma manera que la frecuencia fundamental, varían de acuerdo al sonido emitido y se mantienen presentes todo el tiempo en la señal analizada.
- La relación señal ruido utilizando dos algoritmos diferentes para obtener un resultado óptimo, teniendo en cuenta además que el algoritmo WADA parte del algoritmo NIST inicial, y permite visualizar también el tiempo en el que la señal de voz se hace presente en la grabación
- El espectrograma de la señal, elemento indispensable en el cotejo de voz para observar claramente las formantes y el comportamiento de la señal.
- El porcentaje de tiempo que se hace presente la voz en la grabación (Tiempo de habla neta)
- El género, evaluado mediante el método de correlación.

El software permite observar también la cantidad de energía por frecuencia a través de la transformada rápida de Fourier (FFT) y el cepstrum de la señal que permite identificar la frecuencia fundamental en términos de “quefrencia”.

Además, el software cuenta con diferentes características que pueden ser modificables de acuerdo al parámetro de análisis, para obtener los resultados más precisos y más fáciles de visualizar para las personas que realizan los cotejos de voz.

Adicional a esto, APAVOIX tiene herramientas que permiten trabajar fácilmente con la señal, como son los cursores, los botones de reproducción, el zoom y el paneo, además de contar con un menú de archivo que permite cargar y guardar datos de los cursores para continuar trabajando en el mismo caso, en el momento que se desee, y además permite grabar la señal de audio y almacenarla en la ubicación deseada.

Por otro lado, se debe aclarar que para verificar la influencia de agentes externos en la captura de la señal, como el ruido de fondo o el espacio que se realiza, es necesario que un experto evalúe las condiciones de la voz a ser analizada, es decir, el software procesa sin ningún inconveniente las señales con o sin ruido y sin importar el lugar donde sean grabadas, lo cual puede verse evidenciado en los audios evaluados por el software, uno en un ambiente cerrado, con cierto nivel de ruido, cercano a superficies reflejantes y sin ningún tipo de aislamiento, y el otro, grabado vía GSM, con ruido debido a la compresión de la señal transmitida, que aprovecha las características del oído para transportar por la red, únicamente la información necesaria para llevar a cabo la comunicación de acuerdo a la capacidad del canal. Sin embargo estas características únicamente pueden ser evaluadas por expertos competentes, refiriéndose a la calidad del audio y no del software.

Así mismo es necesario aclarar también que el software está diseñado para trabajar únicamente con una frecuencia de muestreo de 11025 Hz, lo que quiere decir que no es posible comparar el desempeño del software a otras frecuencias de muestreo. Además de esto, el programa únicamente trabaja con audios en formato WAV, por lo que no puede compararse el funcionamiento con otros formatos. Esto no quiere decir que el software no pueda modificarse para hacerlo apto para otras frecuencias de muestreo o algunos otros formatos. La configuración se realizó de esta manera, debido a que las exigencias mínimas para utilizar los audios para cotejo de voz son audios de buena calidad, en formato WAV. PCM a una frecuencia de muestreo de 11025 Hz y 16 bits monocanal.

5. REFERENCIAS

[1] Canuyt Georges, “La Voz; Técnica Vocal, La voz Hablada – el Arte de la Dicción, la Palabra en Público”. Buenos Aires. Hachette, 1955.

[2] Llisterri, Joaquim. “Aparato Fonador” Departament de Filologia Espanyola, Universitat Autònoma de Barcelona. http://liceu.uab.es/~joaquim/phonetics/fon_produccio/aparato_fonador.html

[3] Campo, María Mercedes, Campos, Alexander, Otero, Jorge. “Diseño de un software de tipo VST, mediante un algoritmo matemático, que convolucione una señal de audio, con la respuesta al impulso de recintos cerrados es la ciudad de Bogotá”. Universidad de San Buenaventura, Facultad de Ingeniería, Ingeniería de Sonido 2009, pp. 40.

[4] Agnitio. “Manual de Usuario Batvox 2.2”, Agnitio, SL 2006

[5] Jones, Douglas, Swaroop Appadwedula, Matthew Berry, Mark Haun, Jake Janevitz, Michael Kramer, Dima Moussa, Daniel Sachs, and Brian Wade. “Speech Processing: Theory of LPC Analysis and Synthesis” Connexions. June 1, 2009. http://cnx.org/content/m10482/2.19/

[6] Alzate, Ricardo. “Estimación de contornos del Pitch en Línea sobre DSP”. Universidad Nacional de Colombia Sede Manizales, Facultad de Ingeniería y Arquitectura. http://wpage.unina.it/r.alzate/Support_files/BSc.pdf

[7] Pérez, Constantino. “Ruido”, en: Sistemas de Telecomunicación. Servicio de publicaciones de Universidad Cantabria 2007. http://personales.unican.es/perezvr/pdf/CH8ST_Web.pdf