Sólo le falta hablar…?

Reconocimiento automático del habla, su desarrollo, particularidades y aplicaciones.

Investigación Matemática en el Siglo XXI
Profesor a cargo, Pedro Willging
Facultad de Ciencias Exactas y Naturales
Universidad Nacional de La Pampa

Marina Roldán
Año 2014

Introducción

La matemática está presente en todo lo que nos rodea, desde los pequeños milagros de la naturaleza hasta en las grandes construcciones humanas. Su vinculación con cada área de estudio no sólo revela su importancia sino su versatilidad para adaptarse en sus distintas formas, ya sea desde la abstracción a la aplicación.

Una de las áreas de la ciencia donde la matemática ha constituido una herramienta de aportes continuos como un soporte para su progreso y mejora es la inteligencia artificial. El presente trabajo hace referencia a una de las múltiples ramas de la inteligencia artificial como lo es el procesamiento de señales para el reconocimiento de voz.

De gran auge en la actualidad y con variadas aplicaciones, el reconocimiento del habla tiene como objetivo permitir la comunicación hablada entre seres humanos y computadoras. Contradiciendo al viejo refrán popular *"Sólo le falta hablar"*, la inteligencia artificial lleva paso a paso a la realidad humana a aquel escenario de película ya antes imaginado en que los seres humanos y las computadoras socializan. Un ejemplo más en que la realidad no sólo alcanza a la ficción sino que se nutre con ella.

Un sistema de reconocimiento de voz es una herramienta computacional capaz de procesar la señal de voz emitida por el ser humano y reconocer la información contenida en ésta, convirtiéndola en texto o emitiendo órdenes que actúan sobre un proceso. En su desarrollo intervienen diversas disciplinas, tales como: la fisiología, la acústica, el procesamiento de señales, la inteligencia artificial y la ciencia de la computación.

Cada día nos encontramos más, con infinidad de aplicaciones de los sistemas de cómputo, con capacidad de interactuar con los usuarios mediante el reconocimiento y síntesis de voz. Estos van desde aplicaciones simples en el reconocimiento de comandos aislados (palabras), hasta el reconocimiento de frases para ejecutar acciones a manos libres: teléfonos celulares, control por voz de instrumental de ayuda en la cirugía, acceso a servicios de compra por teléfonos, llenado de solicitudes, reservaciones de pasaje, entre otros; la búsqueda en Internet es una de las muestras más impactantes de estas aplicaciones.

El informe presentado a continuación encuentra su raíz en el trabajo especial de grado realizado por los alumnos Dayana Karina Salcedo Cherubini y Alejandro Patricio Teixeira Gómez: "Diseño de un sistema de reconocimiento del habla para controlar dispositivos eléctricos" en el año 2006 en Caracas;

quienes exponen en el marco teórico del mismo qué se entiende por reconocimiento automático del habla, su desarrollo, particularidades y aplicaciones, con el objetivo de construir un dispositivo de reconocimiento de la voz.

1. Señales: definición, características, aplicaciones.

"Las señales pueden describir una variedad muy amplia de fenómenos físicos, y aunque se pueden representar de muchas formas, en todo caso la información dentro de una señal está contenida en un patrón de variaciones de alguna forma.

Por ejemplo, el mecanismo vocal humano produce el habla mediante la creación de fluctuaciones de la presión acústica. Así, los diferentes sonidos corresponden a diferentes patrones en las variaciones de la presión acústica, y el sistema vocal humano produce una voz inteligible generando secuencias diferentes de esos patrones. Las variaciones de presión acústica se convertirán después en señal eléctrica.

Es decir, una señal no va a ser más que una función de una o unas variables independientes que contiene información acerca de la naturaleza o comportamiento de algún fenómeno." [1]

Una de las principales cuestiones a tener en cuenta es la diferencia existente entre señal y modelo matemático de dicha señal. La primera siempre está dada por una magnitud física mientras que el modelo matemático de la señal atiende más precisamente a cómo varía la señal con el tiempo, lo que comúnmente se denomina representación en el dominio del tiempo.

Generalmente, el modelo matemático de una señal suele venir denominado por el tipo de función que la representa. Claramente la variable que representa nuestra magnitud será la variable dependiente, mientras que, en el caso concreto de las señales que dependen del tiempo, se denominará al mismo variable independiente.

Existen dos parámetros básicos que caracterizan a una señal en el dominio del tiempo, los mismos son:

- *Amplitud:* es el valor que tomará la señal en cada instante de tiempo.
- *Periodo:* se dice que una señal es periódica si se repite cada cierto intervalo de tiempo. Se denomina periodo al tiempo T_0 tal que a partir de un tiempo dado el valor de la señal se repite cada T_0 segundos. Matemáticamente hablando, una señal es periódica si $y(t)=y(t+mT_0)$ con m un número natural y t un número real finito.

[1] http://es.scribd.com/doc/145994765/Capitulo-06-Discretizacion-de-Senales

Ahora bien, existen dos formas de clasificar a las señales, una de ellas es a partir de la variable independiente y se clasifican como:

1. Señales continuas: En el caso de las señales de tiempo continuo la variable independiente es continua y entonces estas señales están definidas para una sucesión continua de valores de la variable independiente.

2. Señales discretas: Una señal o secuencia de tiempo discreto puede representar un fenómeno para el cual la variable independiente es sustancialmente discreta. También puede representar muestras sucesivas de un fenómeno para el cual la variable independiente es continua. Por ejemplo, el procesamiento de la voz en una computadora digital requiere del uso de una secuencia discreta que represente los valores de la señal de voz de tiempo continuo en puntos discretos en el tiempo.

Otra de las formas de clasificar una señal es a partir de la variable dependiente y se clasificarán en:

1. Señales Analógicas: Se dice que una señal es Analógica si la misma es continua y su amplitud puede tomar cualquier valor.

2. Señales digitales: Se dice que una señal es Digital si es discreta y su amplitud sólo puede tomar valores determinados.

Todos los sistemas que trabajan procesando señales de voz necesitan una primera etapa en la cual segmentos consecutivos de la señal de voz son convertidos a secuencias temporales de vectores de parámetros, una manera de disminuir la cantidad de parámetros la parametrización se combina con técnicas discriminativas, las cuales toman los parámetros que mayor información nos puedan proporcionar.

Íntimamente ligado al concepto de señal se encuentra la idea de espectro de la misma. El espectro de frecuencia se caracteriza por la distribución de amplitudes para cada frecuencia de un fenómeno ondulatorio (sonoro, luminoso o electromagnético) que sea superposición de ondas de varias frecuencias. El espectro de frecuencias o descomposición espectral de frecuencias puede aplicarse a cualquier concepto asociado con frecuencia o movimientos ondulatorios como son los colores, las notas musicales, entre otras.

Matemáticamente hablando, el análisis espectral está relacionado con una herramienta llamada transformada de Fourier o análisis de Fourier. Dada una señal o fenómeno ondulatorio de amplitud s(t), esta se pude escribir matemáticamente como la siguiente combinación lineal generalizada:

$$s(t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} A(v) e^{-ivt} dv$$

Es decir, la señal puede ser concebida como la transformada de Fourier de la amplitud A = A(*v*).[2] Ese análisis puede llevarse a cabo para pequeños intervalos de tiempo, o menos frecuentemente para intervalos largos. Asimismo la transformada de Fourier de una función no sólo permite hacer una descomposición espectral de los formantes de una onda o señal oscilatoria, sino que con el espectro generado por el análisis de Fourier incluso se puede reconstruir (*sintetizar*) la función original mediante la transformada inversa. Para poder hacer eso, la transformada no solamente contiene información sobre la intensidad de determinada frecuencia, sino también sobre su fase. Esta información se puede representar como un vector bidimensional o como un número complejo. En las representaciones gráficas, frecuentemente sólo se representa el módulo al cuadrado de ese número, y el gráfico resultante se conoce como espectro de potencia o densidad espectral de potencia (SP):

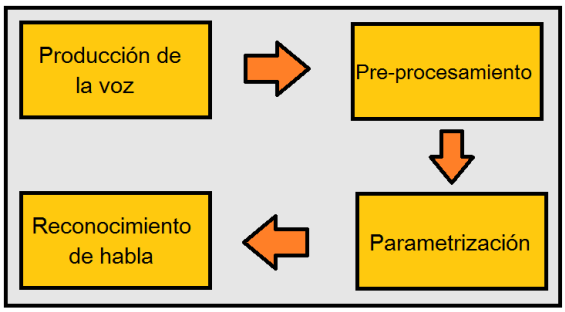
$$SP_v \propto |A(v)|^2$$

Es importante recordar que la transformada de Fourier de una onda aleatoria, mejor dicho estocástica, es también aleatoria. Un ejemplo de este tipo de onda es el ruido ambiental. Por tanto para representar una onda de ese tipo se requiere cierto tipo de promediado para representar adecuadamente la distribución frecuencial. Para señales estocásticas digitalizadas de ese tipo se emplea con frecuencia la transformada de Fourier discreta. Cuando el resultado de ese análisis espectral es una línea plana la señal que generó el espectro se denomina ruido blanco.[3]

El Procesamiento de Voz es el estudio de las señales de voz y de todos aquellos métodos para procesar estas señales. En reiteradas ocasiones se lo denomina digital ya que las señales de voz son llevadas a una computadora para su análisis y procesamiento. Así, el procesamiento de voz es un caso especial del Procesamiento Digital de Señales que se aplica en particular a las señales de voz.

[2] DeVries, Paul L. A first course in computational physics. Ed. John Wiley & Sons, INC. 1994.
[3] Wikipedia, 2014, http://es.wikipedia.org/wiki/Espectro_de_frecuencias

Una de las principales cuestiones a tener en cuenta es que para la realización de cualquier tarea que involucre el procesamiento de una señal, en particular de las señales de voz, es fundamental el reconocimiento de ciertos aspectos que ayuden a crear las bases necesarias para la comprensión de las características de la voz y del tratamiento que se le dará de acuerdo al fin deseado. En el caso del procesamiento del habla, el esquema de procesamiento es similar al siguiente:

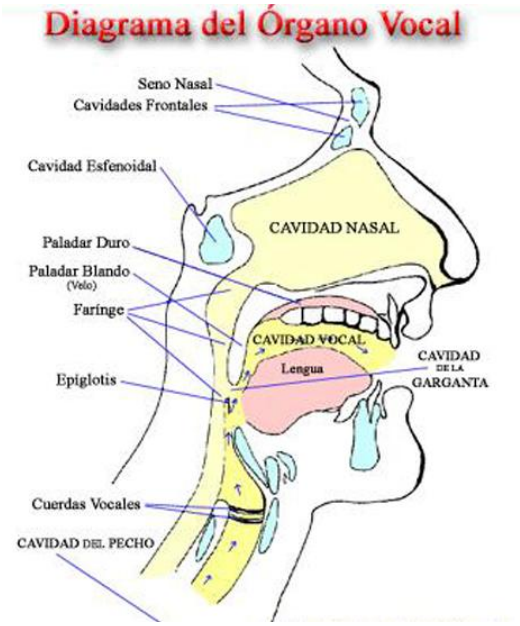


2. Naturaleza y Señal de la voz.

Para comenzar a abordar el tema de interés, una de las principales herramientas que debemos tener a favor es el conocimiento de la naturaleza de la voz.

La voz se obtiene por la acción conjunta de varias regiones de órganos:

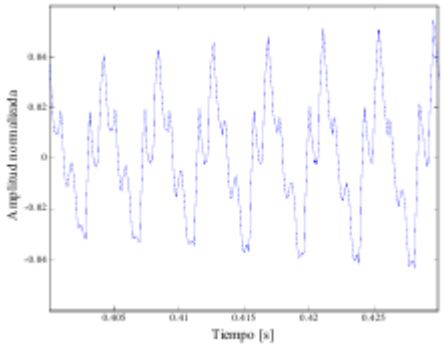
- El tracto pulmonar, formado por los pulmones y la tráquea, los cuales controlan la amplitud de los sonidos.
- La laringe, donde se sitúan las cuerdas vocales, que controlan la entrada de los sonidos y cuya tensión afecta la frecuencia (tono) de la señal de voz.
- El tracto vocal, que se encarga de la articulación de la voz, principalmente por la lengua, los labios y la mandíbula baja.



Fuente: http://es.slideshare.net/dennisepatito/guia-didactica-deidiomaidocx1

De acuerdo a la configuración y el comportamiento de la cavidad bucal asociada al habla, la señal que se genera puede clasificarse en dos grandes grupos:

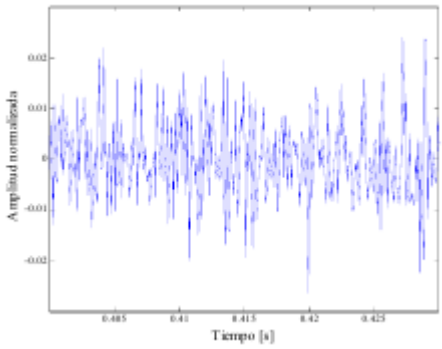
- *Señal sonora:* la misma es producida por la vibración de las cuerdas vocales, las cuales se abren y cierran modificando el área de la tráquea , produciendo una señal modulada y cuasi periódica, con un período o frecuencia llamado pitch. Este tipo de señal se caracteriza por tener alta energía y su rango de frecuencias está entre los 300Hz y los 4000Hz. Por ejemplo la señal sonora correspondiente a la vocal “u” emitida por un niño de 9 años será:



Fuente: DISEÑO DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA PARA
CONTROLAR DISPOSITIVOS ELÉCTRICOS - Dayana Karina Salcedo Cherubini - Alejandro Patricio Teixeira Gómez

- *Señal no sonora:* en este caso el aire fluye libremente, al permanecer abiertas las cuerdas vocales hasta alcanzar el tracto vocal, por lo que la señal generada es una contribución desordenada de componentes frecuenciales. Se caracteriza por tener baja energía y

componente frecuencial uniforme, además debemos considerar que el análisis de esta señal muestra que carece en su totalidad de periodicidad. Por ejemplo, la señal no sonora correspondiente a la letra “s” emitida por un hombre de 22 años será:



Fuente: DISEÑO DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA PARA CONTROLAR DISPOSITIVOS ELÉCTRICOS - Dayana Karina Salcedo Cherubini - Alejandro Patricio Teixeira Gómez

Ahora bien, en el caso de las señales sonoras, el tracto vocal actúa como una cavidad resonante, estando centradas las frecuencias de resonancia generalmente en 500 Hz y sus armónicos pares. Esta resonancia produce grandes picos en el espectro resultante, a los cuales se les llama formantes. Por su parte, para las señales no sonoras, el tracto vocal presenta una estructura ruidosa y aleatoria tanto en el dominio temporal como en el frecuencial, por lo que no se tienen formantes. Un formante es el pico de intensidad en el espectro de un sonido; se trata de concentración de energía (amplitud de onda) que se da en una determinada frecuencia y viene determinado por el proceso de filtrado que se produce en el tracto vocal, dependiendo de la configuración de los articuladores como lo son la lengua, la mandíbula, los labios y el velo del paladar.

Los formantes permiten distinguir los sonidos del habla humana, sobre todo las vocales y otros sonidos sonorantes. También sirven para los sistemas de reconocimiento de voz y las transposiciones de altura del audio digital. Esto es posible porque cada sonido del habla humana tiene una marca característica de formantes, es decir, hace un reparto diferente de la energía sonora entre los diferentes formantes, lo cual permite clasificarlos o categorizarlos. El oído humano puede hacer ese análisis de formantes de manera inconsciente, y es por eso que podemos distinguir los sonidos de nuestra lengua materna. El formante de frecuencia más baja se denomina F_1; el segundo F_2; y así sucesivamente. En las vocales, los dos primeros formantes se determinan principalmente por la posición de la lengua. F_1 tiene una frecuencia más alta cuanto más baja está la lengua; es decir, cuanta mayor abertura tenga

una vocal, mayor es la frecuencia en que aparece el F_1. El F_2 tiene mayor frecuencia cuanto más hacia delante está posicionada la lengua, es decir, cuanto más anterior es una vocal, mayor es el F_2.[4]

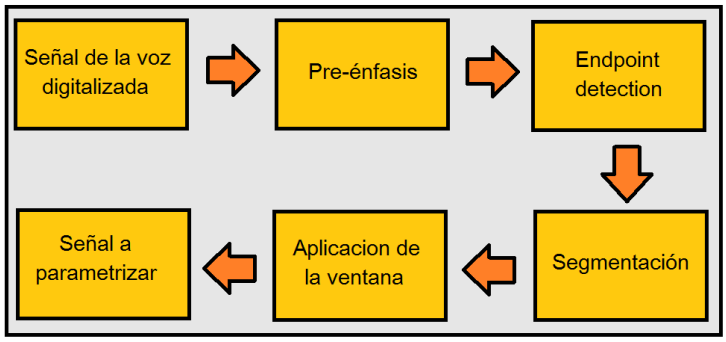
Una vez ya en conocimiento de estas cuestiones, es necesario destacar un hecho de relevancia matemática, y es que la Transformada de Fourier resulta una herramienta de gran importancia para el análisis de señales de voz en el dominio de la frecuencia, ya que permite obtener información de las mismas que no es evidente en el dominio del tiempo. Una de las principales características de la Transformada de Fourier es que permite extraer el contenido frecuencial de la señal, lo cual es muy importante para establecer el muestreo adecuado para la misma. La importancia que se le adjudica se debe al Teorema de Nyquist, que establece que la frecuencia de muestreo óptima para cualquier señal debe ser por lo menos dos veces el valor de su frecuencia máxima.

3. Procesamiento de la voz.

A continuación se llevará adelante un breve desarrollo de los procesos implicados en el procesamiento y reconocimiento de la voz humana, ya citados con anterioridad, a saber: pre-procesamiento y parametrización.

3.1. Pre-procesamiento de la voz

Esta etapa corresponde a los pasos previos a la etapa de parametrización y es la que permite resaltar las características más importantes de habla que luego serán parametrizadas. Esta etapa del proceso se desarrollará teniendo en cuenta algunos puntos básicos entre los que se encuentran:



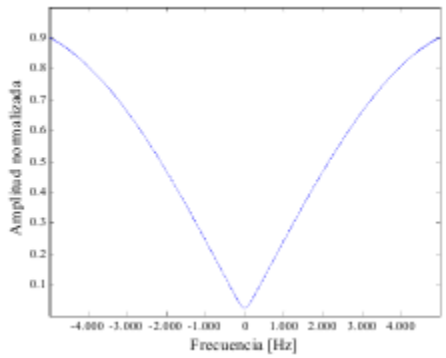
[4] Wikipedia, 2014, http://es.wikipedia.org/wiki/Formante.

Previo a esto, la voz ha sido convertida en una onda sonora que ingresa por el micrófono y bajo una conversión analógica-digital fue llevada al computador de manera que la señal ahora ya se encuentra digitalizada. A continuación se describirán las fases siguientes:

- *Preénfasis:* una vez que la señal ya ha sido digitalizada, la etapa de preénfasis se realizará con el objetivo de hacer el procesamiento de la señal menos susceptible a truncamientos y aplanarla espectralmente. Generalmente se usa un filtro digital de primer orden cuya función de transferencia será:

  $$H(z) = 1 - a.z^{-1}$$

  Donde $0,9 \leq a \leq 0,95$, es un valor que se escoge cercano a la unidad a fin de que la estructura de los formantes mayores sean acentuadas. La representación en frecuencia del filtro de preénfasis mencionado se muestra a continuación.



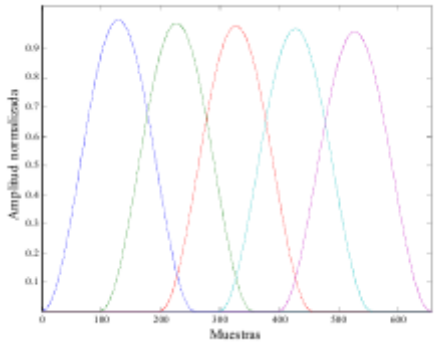
Fuente: DISEÑO DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA PARA CONTROLAR DISPOSITIVOS ELÉCTRICOS - Dayana Karina Salcedo Cherubini - Alejandro Patricio Teixeira Gómez

  La salida del sistema de preénfasis S(n) está relacionada a la entrada del sistema s(n) mediante la ecuación

  $$S(n) = s(n) - a.s(n-1)$$

- *Endpoint detection:* una vez que la señal de voz ha sido ya grabada en la PC, resulta de importancia determinar el comienzo y el final de las partes útiles de la señal, lo cual se realiza mediante una técnica, conocida como endpoint detection que, mediante un algoritmo de búsqueda, encuentra los puntos de la señal donde se concentra la energía y extrae sólo ese fragmento.
- *Segmentación:* llegado este momento la voz ha sido ya digitalizada y se han reconocido las principales partes del habla, es decir la información de interés que la misma contiene. Por ello el siguiente paso a seguir es dividir la señal en N muestras, donde N es un valor arbitrario que

se escoge tomando en cuenta que la señal de la voz es estacionaria "a trozos"; lo cual resulta una condición necesaria para poder realizar el análisis de Fourier en tiempo corto. El intervalo de tiempo que se considerará estacionaria a la señal dependerá de la velocidad de los cambios del tracto vocal y de las cuerdas vocales y comúnmente se establece un valor entre los 20 y 40 ms. La figura siguiente muestra



Fuente: DISEÑO DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA PARA CONTROLAR DISPOSITIVOS ELÉCTRICOS - Dayana Karina Salcedo Cherubini - Alejandro Patricio Teixeira Gómez

La figura anterior muestra un ejemplo de segmentación utilizado comúnmente en las aplicaciones que involucran el procesamiento de la voz. Por lo general se utilizan tramos de 256 muestras por ser un valor que establece un equilibrio entre la resolución en tiempo y en frecuencia para una señal de voz. En el caso particular de la figura esta cantidad de muestras corresponde a 25,6 ms de señal, ya que la frecuencia de muestreo es de 10 KHz.

- *Aplicación de la ventana:* las ventanas son funciones matemáticas usadas para evitar las discontinuidades al principio y al final de los bloques analizados. Las mismas se utilizan cuando nos interesa una señal de longitud voluntariamente limitada. Para observar una señal en un tiempo finito, la multiplicamos por una función ventana. La más simple es la ventana rectangular, que se define como:

$$h(t) = \begin{cases} 1 & si\ \ t \in [0,T] \\ 0 & en\ otro\ caso \end{cases}$$

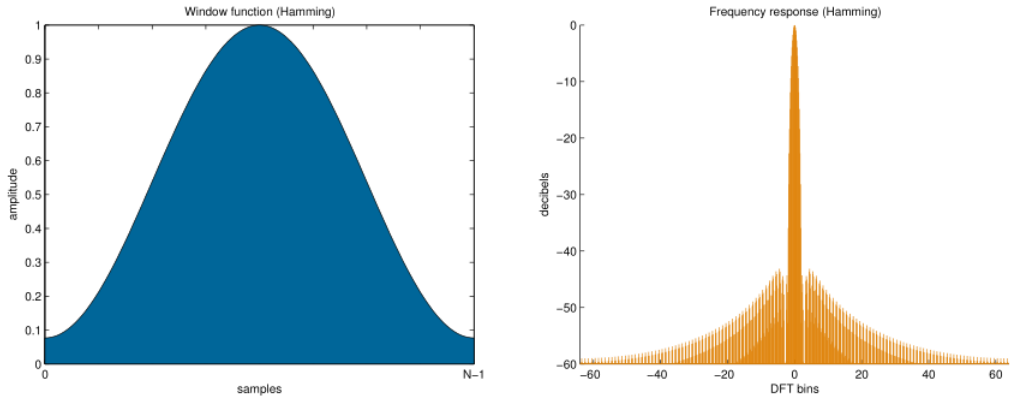
Así, al multiplicar una señal s(t) por esta ventana, se obtendrán únicamente los T primeros segundos de la señal y se podrá observar la señal en un intervalo T. Es decir, en lugar de estudiar la señal s(t), se estudia la señal truncada: $s_h(t) = s(t)\ .\ h(t)$.

Si pasamos al dominio de la frecuencia, mediante una transformada de Fourier, obtenemos el producto de convolución $s_h(f) = s(f) \cdot H(f)$, donde H(f) es la transformada de Fourier de la ventana. [5]

Así, esta etapa del pre-procesamiento corresponde a la elección del tipo de ventana que se le aplica a cada tramo. Por lo cual esta parte del trabajo resulta de gran importancia para analizar el efecto de la ventana sobre la resolución espectral de la señal, la cual dependerá del ancho de la onda principal de la ventana y la atenuación de las ondas secundarias respecto a la principal. Lo ideal es que la onda principal sea muy estrecha respecto de las secundarias que deberían ser lo más pequeñas posibles, de modo tal que resulte esta configuración similar a un Delta de Dirac. Tanto las ventanas rectangulares como las no rectangulares tienen sus ventajas y desventajas, pero en el caso específico de la ventana de Hamming se cumple que la atenuación de las ondas secundarias sea brusca respecto de la principal lo que permite tener una mejor resolución espectral.

La ventana Hamming se define matemáticamente por la ecuación

$$W(nt) = 0{,}54 - 0{,}56.\cos\left(\frac{2\pi}{N}\right) \;\; donde \;\; 0 < n < N$$



Fuente: wikipedia, 2014, http://es.wikipedia.org/wiki/Ventana_(funci%C3%B3n).

[5] Wikipedia, 2014, http://es.wikipedia.org/wiki/Ventana_(funci%C3%B3n)

3.2. Parametrización de la Señal de Voz.

Ahora bien, una vez realizado el pre-procesamiento de la voz, se continuará, de acuerdo al esquema inicial con la parametrización de la misma. La selección de la mejor representación paramétrica de la voz es una tarea importante en el diseño de cualquier sistema de reconocimiento del habla. El objetivo principal que persigue esta representación es comprimir los datos correspondientes a la señal de la voz eliminando la información no pertinente al análisis fonético de la información y extraer asimismo las características de la señal de la voz que contribuyen con la detección de las diferencias fonéticas que no son apreciables mediante un simple análisis en tiempo o en frecuencia, sino usando un análisis más exhaustivo.

El análisis por preferencia a usar en el reconocimiento del habla es el llamado Cepstrum. El mismo es una de las herramientas más utilizadas para la representación paramétrica de las señales de voz, y se define como la Transformada de Fourier del espectro logarítmico de la señal. Esto nos lleva claramente a la existencia de un Cepstrum real y uno imaginario dependiendo de si la función logarítmica está definida para valores reales o complejos. La principal diferencia entre uno y otro radica en el hecho de que el Cepstrum complejo permite reconstruir la señal y el real no, ya que se pierde la información correspondiente a la fase. Para entender las bases matemáticas del cálculo del Cepstrum es necesario considerar una secuencia estable x(n), cuya Transformada Z se puede expresar en coordenadas polares según la ecuación:

$$X(z) = |X(z)|.\, e^{j\propto X(z)}$$

Donde el lado izquierdo de la ecuación representa la magnitud y αX(z) el ángulo de la Transformada Z de x(n). Como la señal x(n) es estable, la región de convergencia para X(z) incluye el círculo unitario y la Transformada de Fourier de x(n) existe y es igual a $X(e^{j\omega})$. El Cepstrum complejo correspondiente a x(n) se define como una secuencia estable $\hat{x}(n)$ cuya transformada Z se puede definir mediante la ecuación:

$$\hat{X}(z) = log(X(z))$$

Como se requiere que $\hat{x}(n)$ sea estable, la región de convergencia incluye el círculo unitario, por lo que el Cepstrum complejo se puede representar usando la Transformada Inversa de Fourier, tal y como se observa en la ecuación:

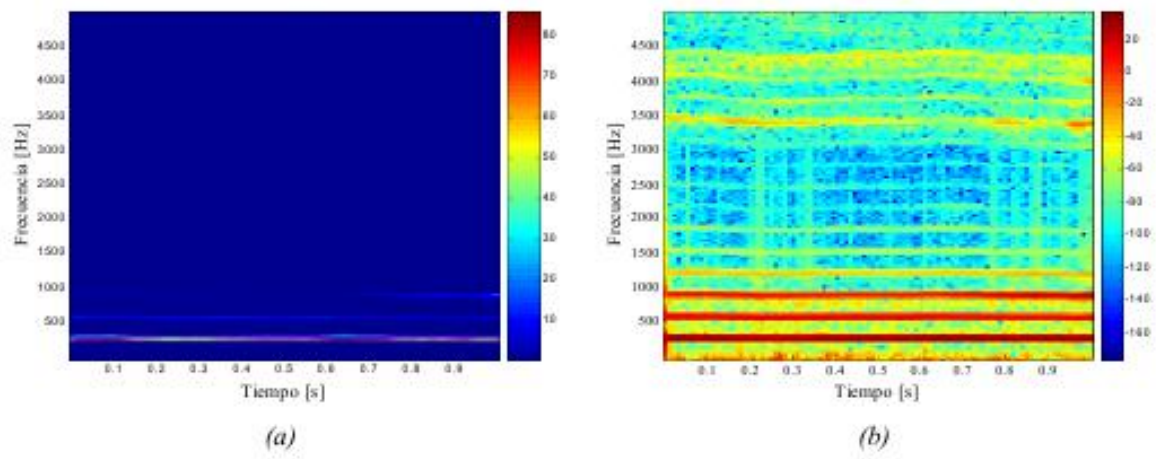
$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\mathrm{X}\left(\mathrm{e}^{\mathrm{j}\omega}\right)\right).\, \mathrm{e}^{\mathrm{j}\omega\mathrm{n}}\mathrm{d}\omega$$

En contraste con el Cepstrum complejo, el $x_c(n)$ o Cepstrum real de una señal, es definido como la Transformada Inversa de Fourier del logaritmo de la magnitud de la Transformada de Fourier, tal y como se muestra en la ecuación:

$$c_x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X\left(e^{j\omega}\right)\right|. e^{j\omega n} d\omega$$

El Cepstrum real se utiliza en muchas aplicaciones y como no depende de la fase de $X(e^{j\omega})$ es mucho más fácil de calcular que el Cepstrum complejo, aunque x(n) no puede ser recuperada a partir de $x_c(n)$.

El análisis Cepstral es comúnmente utilizado para obtener información de la señal de voz que permita parametrizarla para luego ser usada en la fase de reconocimiento. Mediante el Cepstrum se puede separar la señal de excitación del sistema que modela la producción de la voz y la función de transferencia que modela el tracto vocal. Por esta razón es que al analizar el espectrograma mostrado en la figura *b* (señal de voz original) se pueden observar componentes frecuenciales que no se aprecian en la figura *a* (señal de voz con Cepstrum).



Fuente: DISEÑO DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA PARA CONTROLAR DISPOSITIVOS ELÉCTRICOS - Dayana Karina Salcedo Cherubini - Alejandro Patricio Teixeira Gómez

Ahora bien, las representaciones paramétricas más utilizadas para trabajar con señales de voz y que están basadas en el análisis Cepstral de la misma, pueden ser divididas en dos grandes grupos: aquellas basadas en la predicción lineal del espectro y aquellas basadas en el espectro de Fourier.

En el presente trabajo se hará hincapié en las basadas en el espectro de Fourier. Para esta técnica de parametrización, las características espectrales de la señal de voz se derivan del análisis de Fourier de tiempo corto, definido por la ecuación:

$$S_x(t,f) = \int_{-\infty}^{\infty} x(\tau + t).w(\tau).e^{-j2\mu f\tau}\, d\tau$$

Donde x(τ) es la señal de voz y w(τ) representa la función de la ventana de análisis (por ejemplo, la ventana Hamming). De esta forma se realiza un análisis localizado de la señal mediante la aplicación de una ventana w(τ) a la señal, alrededor del instante de tiempo t, analizada a todas las frecuencias consideradas f.

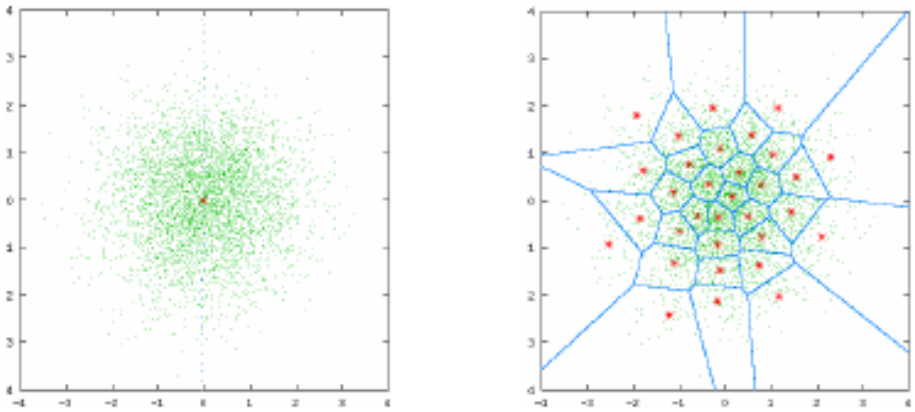
Una vez realizado el análisis Cepstral el cual permite reconocer las principales características de la señal de voz de importancia, es momento de parametrizarla. Existen diferentes técnicas de parametrización aunque todas ellas se basan en la determinación de una matriz que represente los componentes principales de la señal dada. Una de las técnicas más usadas en el procesamiento de las señales de voz es el análisis de predicción lineal. La misma ha probado ser eficiente debido a que permite parametrizar a la señal con un número pequeño de patrones con los cuales es posible, entre otras cosas, reconstruir la señal dada de manera eficiente. Asimismo, mediante esta técnica se puede representar a la señal de voz usando parámetros que varían en el tiempo y que están relacionados con la función de transferencia del tracto vocal y las características de la fuente sonora.

La mencionada técnica de Predicción Lineal de los Coeficientes Cepstrales (LPCC) permite estimar los coeficientes cepstrales mediante el uso del algoritmo LPC el cual establece un modelo que permite calcular la próxima muestra de la señal sonora.

Continuando con esta línea de trabajo, una vez parametrizada la señal de la voz se habrá generado una serie de coeficientes que representan las características de la señal dada que pueden ser utilizadas en la fase de reconocimiento del habla.

El tamaño de la matriz obtenida en el proceso de parametrización depende directamente de la longitud de señal de la voz, la cual resulta variable pues tiene relación con la palabra en sí y con el hablante. Por este motivo resulta de gran importancia la estandarización de la matriz que contiene los coeficientes cepstrales calculados, para que el tamaño de las matrices usadas en cada caso, para el reconocimiento del habla, sea el mismo.

La estandarización de la matriz de coeficientes cepstrales constituye el primer paso en el proceso de reconocimiento del habla y se denomina cuantización vectorial. La esencia de esta técnica aplicada a; caso particular del reconocimiento del habla, es la de obtener una matriz cualquiera de coeficientes cepstrales cuyo tamaño sea fijo y se parezca lo más posible a la matriz dada originalmente. Para generarla, el espacio dado por los coeficientes cepstrales es dividido en un conjunto de regiones convexas mutuamente excluyentes y para cada una de las mismas se calcula el centroide; el cual pensado en dos dimensiones resultaría ser un punto que se encuentra a la menor distancia del resto de los puntos que están en esa región. El conjunto de centroides obtenidos se denomina codebook. Por lo tanto, caracterizar a una palabra significa calcular su codebook correspondiente.



Fuente: DISEÑO DE UN SISTEMA DE RECONOCIMIENTO DEL HABLA PARA CONTROLAR DISPOSITIVOS ELÉCTRICOS - Dayana Karina Salcedo Cherubini - Alejandro Patricio Teixeira Gómez

En la figura de la izquierda se puede observar la ubicación del primer centroide calculado, mientras que en la figura de la derecha se muestra el codebook correspondiente a todos los centroides generados mediante la cuantización vectorial.

Como puede observarse entonces, la finalidad de la cuantización vectorial es la de obtener una matriz que represente de la manera más exacta posible a la matriz original y que además permita hacer una comparación razonable entre la señal de voz que proviene del hablante y las señales contenidas en la base de datos del sistema de reconocimiento del habla. Así, dado que todas las matrices tienen el mismo tamaño será posible aplicar alguna técnica comparativa entre ellas.

La principal técnica de comparación de matrices resulta ser la distancia euclideana. La misma se emplea para el reconocimiento del habla como método para calcular la distancia existente entre el codebook obtenido de la palabra dicha por el hablante y el codebook de la palabra almacenada en la base de datos del sistema de reconocimiento. El resultado de la comparación es un valor numérico que

representa la mencionada distancia entre ambas matrices de igual dimensión. El codebook perteneciente a la base de datos de entrenamiento del sistema, cuya distancia al codebook generado para representar la palabra dicha por el hablante sea menor que el nivel de comparación establecido de la aplicación, identifica la palabra con la que existe mayor semejanza.

4. Aplicaciones del Procesamiento de Señales del habla

Hasta aquí se ha visto cómo a partir de una señal de voz, puede ser digitalizada, procesada y parametrizada, mecanismo mediante en el cual se basa el reconocimiento del habla. Sin embargo, y una de los lados más interesantes de este enfoque, no es el procesamiento de la voz sino las aplicaciones que el reconocimiento del habla tiene en la vida actual.

El ser humano se encuentra inmerso en un mundo actualizado en el cual la tecnología se ha hecho presente en todos los aspectos de la vida diaria, el reconocimiento del habla no marca la diferencia y se encuentra en múltiples facetas de la cotidianidad humana.

Sin duda los principales aportes son aquellos que además de novedosos realizan un trabajo de ayuda a la comunidad en algún punto de su existencia. En el caso particular del reconocimiento del habla sorprenden sus aportes tanto a la educación como al cuidado de la salud. Entre los principales y más interesantes aportes del reconocimiento del habla se encuentran los mencionados a continuación.

4.1. El reconocimiento del habla a favor del aprendizaje y la medicina.

“La posibilidad de ver lo que se dice, ha resultado muy novedosa como método para la implantación y rehabilitación de la Voz y el 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 y en Escuelas Especiales. En protocolos de investigación realizados durante un año en Escuelas Especiales, se ha notado un adelanto sustancial en el aprendizaje de la correcta dicción, en aquellos alumnos que adicionalmente al método tradicional, utilizaron un sistema de extracción y visualización de perfiles acústicos, representación de imágenes asociada al sonido y realimentación auditiva del sonido patrón y del producido por el usuario durante la sesión de trabajo” afirma Sergio Suárez Guerra

del Centro de Investigación en Computación de México. Uno de los principales matemáticos involucrados en el trabajo con el "Sistema para la extracción y análisis de parámetros de la voz" EXPARAM V.1.2, programa que ha sido incorporado en la República de Cuba que busca apoyar la gestión en consultas de foniatría mediante el análisis de la voz y que involucra de igual manera la representación de perfiles acústicos, así como la inclusión de una base de datos clínicos de los pacientes.

El mencionado programa se instaló inicialmente y de manera experimental, en una escuela de niños sordos y con trastornos del lenguaje en la República de Cuba, realizándose una evaluación del mismo en las actividades de enseñanza. Los resultados fueron alentadores, los profesores se familiarizaron con el uso de computadoras personales estándares y el software aplicado, los niños aceptaron el producto como un elemento de aprendizaje que les posibilitaba además el acceso a las computadoras.

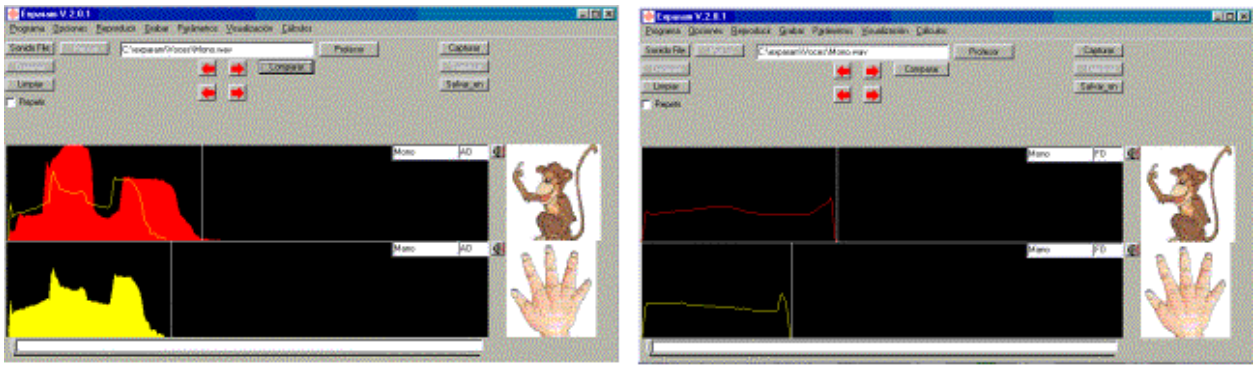
En la actualidad se continúa el desarrollo de aplicaciones para la educación y se inicia el diseño y programación de sistemas para el análisis de voz en el área médica de consultas de foniatría. Más particularmente, el objetivo central para los nuevos desarrollos se presenta en dos líneas: los sistemas educacionales y las aplicaciones médicas. Así, los sistemas presentados son el resultado de varios años de trabajo en la línea de procesamiento de voz para aplicaciones de educación y medicina, en ambos casos los software están instalados en centros de atención y con ellos se llevan a cabo trabajos de análisis, entrenamiento y rehabilitación de problemas de voz y habla.

A nivel educación, las principales características del programa residen en dos puntos básicos:

• Se incorporan representaciones de imágenes para cada sonido de voces que se recibe con el sistema, de forma tal que además de poder ver la señal acústica de la voz y los perfiles paramétricos que se extraen de cada sonido, el usuario puede ver el significado del sonido en una figura.

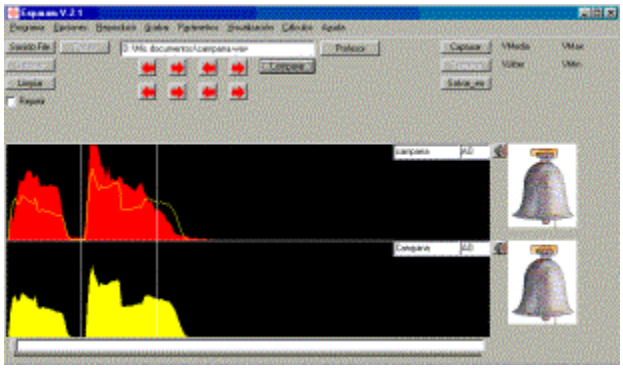
• El usuario puede incorporar nuevos archivos de voces. Si desea ver la representación en forma de imagen, tiene que incorporar la misma en la carpeta correspondiente, en el formato JPG.

La siguiente figura es un ejemplo de lo antes mencionado en relación al uso del programa en la misma se visualizan las palabras "mano" y "mono", en la figura del lado izquierdo de la figura se da una representación de los perfiles de intensidades de ambas mientras que la figura de la derecha muestra una representación de los perfiles del Tono Fundamental para cada palabra.



En cuanto al área médica el sistema para la gestión y análisis acústico en consultas de foniatría es el llamado FONAVOZ 1.0. El mismo fue creado producto de un trabajo de maestría y está capacitado para el llenado de una base de datos de la información personal de los pacientes atendidos adicionando a la misma, archivos de voces que permiten realizar análisis de la voz de los mismos. Igualmente, la funcionalidad del software es análoga a la de EXPARAM ya que se tiene la posibilidad de observar la señal real, el parámetro de frecuencia y la representación gráfica de la palabra mediante el uso de imágenes. Con estas representaciones el especialista puede comparar los gráficos correspondientes a una voz normal y la del paciente bajo estudio y, además de diagnosticar, proponer sesiones de entrenamiento, donde el paciente mejore su dicción, de ser necesario.

Como ya se lo mencionó en el caso del software educacional y en el cuál el programa de soporte médico no es una excepción, se es posible dentro de estos sistemas comparar, mediante la superposición de gráficas los perfiles acústicos que representan la voz. Esta herramienta permite visualizar de manera más objetiva que tan diferentes son los mismos sonidos pronunciados por dos locutores distintos, al tiempo que permite ver que tanto se aleja de lo esperado, la fonación de un paciente enfermo. Asimismo, para trabajos de rehabilitación, el uso de FONAVOZ 1.0 ha resultado de gran importancia ya que el usuario puede detectar con mayor precisión en qué parte de la fonación está incurriendo una falta, omisión o pronunciación equivocada. En la figura siguiente se muestra una ilustración representativa de esta situación, en la cual la palabra “campana” ha sido pronunciada por dos personas diferentes y si bien los perfiles acústicos de la energía son similares, el perfil superior tiene menos definida la vocal “a” terminal.



Ahora bien, es de importancia notar que los programas por sí solos son insuficientes para el tratamiento de pacientes con enfermedades del habla, pero constituyen una herramienta auxiliar de aportes muy significativos para el tratamiento y el establecimiento de las posibles causas del problema presente en el locutor.

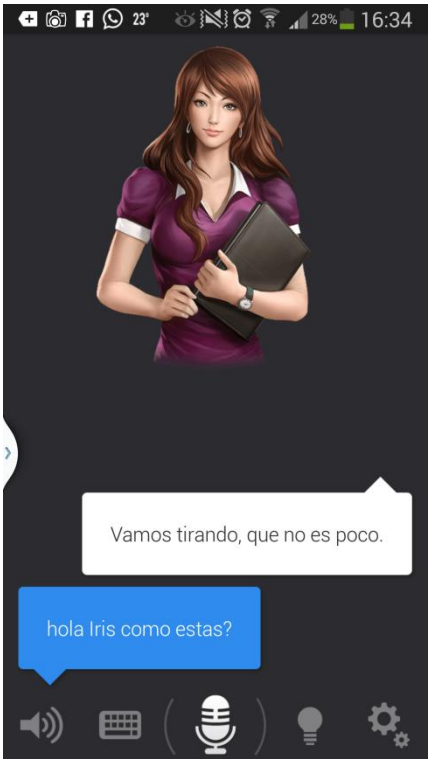
4.2. El reconocimiento del habla en el uso de los Smartphones

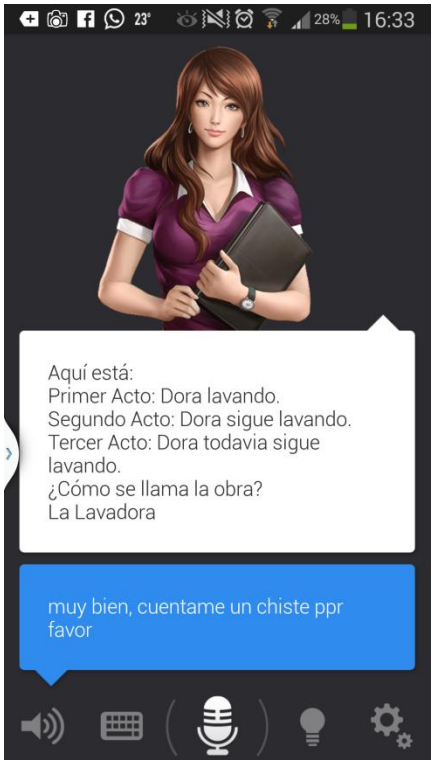
Las mejoras en el hardware y los avances de los sistemas operativos han posibilitado una nueva generación de aplicaciones que pueden funcionar como asistentes inteligentes de los usuarios de smartphones y tabletas que permiten realizar acciones y tareas que sólo eran posibles en la ciencia ficción. Estos asistentes van más allá de programar comandos de vos para realizar tareas específicas, sino que son aplicaciones que están orientadas a responder preguntas basadas en datos accesibles a través de internet o guardadas en el propio teléfono mediante una conversación amigable. Se presentan a continuación dos ejemplos de cómo estos asistentes han revolucionado el mundo de los dispositivos móviles.

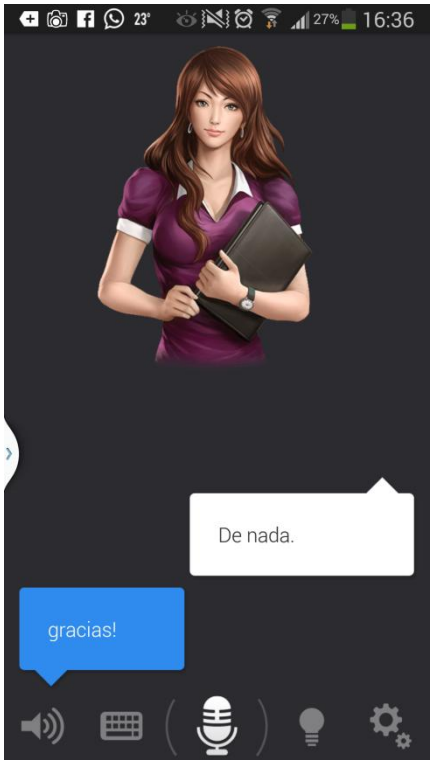
- Siri. Siri es una de las novedades del último sistema operativo iOS 5 de Apple y está presente en los dispositivos móviles de dicha marca. Es un asistente personal que funciona por reconocimiento de voz permitiendo realizar tareas mediante comandos de voz que comprenden el lenguaje natural de forma análoga a la que nos comunicamos entre las personas de nuestro entorno con frases como "¿Qué tiempo va a hacer mañana por aquí?" o "¿Cuál es mi agenda para hoy?". En el primer caso, Siri interactúa con el GPS para saber la posición del usuario y luego busca en la Red el pronóstico para, en consecuencia, ofrecérselo. En el segundo, muestra la agenda que tiene el usuario guardada en su móvil. Asimismo el asistente está capacitado para escribir mensajes de texto o correos electrónicos, organizar reuniones, tomar notas, realizar llamadas telefónicas, ente otras.
- Iris. Siri, el nuevo asistente de voz del iPhone ha encontrado rápidamente un equivalente para terminales Android. Se trata de una aplicación llamada Iris la cual posee un software de reconocimiento de voz, creado por un grupo de desarrolladores en un concurso en tan solo 8 horas. Permite realizar búsquedas de voz sobre varios temas como conversiones, arte,

literatura, historia y biología, además de desempeñar tareas simples como escribir mensajes de texto o correos electrónicos, organizar reuniones, tomar notas, realizar llamadas telefónicas, ente otras. De gran versatilidad y adaptado al lenguaje natural este nuevo asistente permite elegir el avatar que lo representará, su voz y el lenguaje del mismo.

Se muestra a continuación una breve conversación con Iris desde un terminal Samsung con Sistema operativo Android. La aplicación fue descargada desde el sitio de descargas “Play Store”, una vez conocida la misma motivo del presente trabajo.







Conclusiones

Para concluir y en base a lo expuesto en páginas anteriores se puede afirmar que el reconocimiento del habla resulta una herramienta de múltiples aplicaciones cuya existencia se ha universalizado en las diferentes áreas de la ciencia generando grandes aportes a las mismas.

Desde aplicaciones simpáticas para el uso diario en tabletas, computadoras o teléfonos móviles, hasta la creación de software para el tratamiento de enfermedades en general y del habla en particular, usadas en escuelas o centros de salud, el reconocimiento de la voz se adapta con versatilidad generando un aporte de gran importancia a la vida actual de las personas.

El enfoque del trabajo es básicamente desde el área de la matemática pero los ejemplos muestran el potencial que esta nueva herramienta envuelve, lo cual se manifiesta en su presencia en el cotidianidad que rodea al ser humano.

Ideado por la ciencia ficción y anhelado por la sociedad, diversas ramas de la ciencia trabajaron en conjunto para acercarnos en la actualidad aquello que antes era sólo una idea absurda: la comunicación entre el hombre y los dispositivos electrónicos, dejando al ser humano a tan sólo un paso del “hombre bicentenario”.

Bibliografía

- Salcedo Cherubini, Dayana. Teixeira Gómez, Patricio. (2006). "Diseño de un sistema de reconocimiento del habla para controlar dispositivos eléctricos". Universidad Católica Andrés Bello. Caracas, Venezuela.
- Casacuberta Nolla, F. (2003). "La lengua española y las nuevas tecnologías: Análisis y síntesis de la señal Acústica". Universidad Politécnica de Valencia. España. Véase en: http://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/mesaredon_casacuberta.htm.
- DeVries, Paul L. (1994). "A first course in computational physics". Ed. John Wiley & Sons, INC. Oxford, Ohio.
- http://profesores.fi-b.unam.mx/jareyc/Voz/Tema1.pdf
- http://www.cic.ipn.mx/sitioCIC/images/seminarios/b11/Material/proyectos.pdf
- http://elastixtech.com/fundamentos-de-telefonia/transmision-de-la-voz/
- http://www.soyentrepreneur.com/24633-3-apps-para-convertir-voz-en-texto.html
- http://www.tesis.uchile.cl/handle/2250/104320
- http://imarrero.webs.ull.es/sctm04/modulo2/15/jsanrosa.pdf
- http://huitoto.udea.edu.co/SistemasDiscretos/contenido/capitulo_07.html
- http://www.rcim.sld.cu/revista_6/articulo_htm/perfiles.htm
- http://es.scribd.com/doc/145994765/Capitulo-06-Discretizacion-de-Senales
- http://www.smartblog.es/2012/03/iris-vs-cloe-y-el-reconocimiento-de-voz-en-android/
- http://cofretecnologico.com/los-smartphones-android-tiene-su-propio-siri-iris.html
- http://es.wikipedia.org/wiki/Reconocimiento_del_habla
- http://www.consumer.es/web/es/tecnologia/software/2012/01/03/205630.php