Identificación de personas por medio de la voz utilizando LPC's y reconocimiento por KNN (K-Nearest Neighbors)

Jesús-Gustavo Medrano-Romero[1], Yesenia Nohemí González-Meneses[2], José Federico Ramírez-Cruz[3] y Blanca Estela Pedroza-Méndez[4]

[1, 2, 3, 4] Instituto Tecnológico de Apizaco, Apizaco, Tlaxcala, México
[1]didier9_10@hotmail.com, [2]yeseniaglez@hotmail.com, [3]federico_ramirez@yahoo.com
[4]thelismedina@hotmail.com
Paper received on 08/08/12 , Accepted on 10/08/12

Resumen. El reconocimiento de voz es un área de la inteligencia computacional que sigue siendo explotada y muy utilizada en la actualidad por diferentes empresas e instituciones, logrando cada vez mejores resultados en cuanto al porcentaje de reconocimiento; dicho reconocimiento podemos dividirlo en dos grandes ramas, reconocimiento de habla (que intenta identificar lo que la persona dice) y reconocimiento de hablante (que intenta encontrar a la persona que habla). Este trabajo está enfocado directamente al reconocimiento de hablantes por lo que se presenta un modelo para el procesamiento de la señal obtenida de la voz y otro para el entrenamiento, al final se hace el reconocimiento de dicha señal con el objetivo de identificar al hablante. Esto se lleva a cabo a través de la aplicación de LPC´s para el procesamiento de la señal y el reconocimiento se hace mediante la combinación de KNN para encontrar elementos parecidos, probabilidad para determinar la mejor solución y actualización de pesos para garantizar un mejor funcionamiento.

Palabras Clave: Reconocimiento de Voz, Identificación de hablante**.**

1 Introducción.

El proceso de reconocimiento automático del habla permite a las máquinas recibir mensajes hablados u orales. Tomando como entrada la señal acústica recogida por un micrófono, el proceso de reconocimiento automático del habla tiene como objetivo final decodificar el mensaje contenido en la señal acústica para realizar las acciones pertinentes. Para lograr este fin, una herramienta de cómputo necesita tener una gran cantidad de conocimientos sobre del sistema auditivo humano, sobre la estructura del lenguaje, la representación del significado de los mensajes y sobre todo el auto aprendizaje de la experiencia o el uso diario. Actualmente estamos lejos de lograr un sistema completo que pueda comprender cualquier mensaje oral en cualquier contexto tal y como lo podría hacer un ser humano. Sin embargo, la tecnología

Advances in Intelligent and Information Technologies.
Research in Computing Science 60, 2012, pp. 77 -86.

actual sí permite realizar sistemas de reconocimiento del habla que pueden trabajar, con un error aceptable.

Existe gran cantidad de trabajos en donde se recopilan los avances que se han dado dentro del área del reconocimiento automático del habla, ya que no es un problema nuevo, un ejemplo de ello lo tenemos en [2], este trabajo es una recopilación de varios artículos cada uno dedicado al análisis de una rama específica dentro de esta área, componiendo así una colección de documentos dedicados a explicar los avances que se han tenido desde hace algunos años. En [4] se explican los fundamentos necesarios para iniciarse dentro del área del reconocimiento del habla, desde definiciones hasta aplicaciones funcionales; y así como se mencionan estos dos trabajos, existen muchos más que explican el porqué utilizar tecnologías inteligentes y como usarlas para solucionar el problema del reconocimiento del habla. En los siguientes apartados se decribe como se ha abordado en este trabajo dicho problema, presentando primero la metodología general que se siguio, posteriormente las conclusiones y resultados obtenidos y al final se lista una serie de trabajos futuros.

2 Metodología.

Las principales etapas de la solución planteada se puede observar en la Fig. 1, cada uno de los paso implica un proceso en sí, en las siguientes secciones se presenta una breve explicación de cada uno de estos.

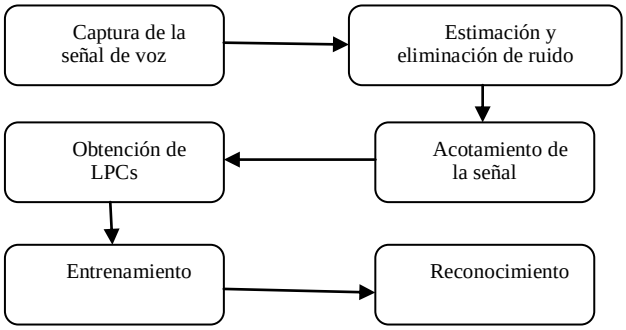

Figura 1. Metodología general.

2.1 Captura de la señal de voz.

Las muestras de voz con las que se estuvo trabajando durante toda la investigación se obtuvieron de dos maneras distintas:

- Base de datos proporcionada por el INAOE. Esta base de datos consta de comandos de voz cortos como “adelante” o “atrás”, así como números pronunciados por hombres y mujeres utilizados en otras áreas del reconocimiento de voz.
- Capturadas directamente con micrófono a través de una laptop Dell studio, igualmente consta de muestras de hombres y mujeres con la diferencia de que se prestó mayor atención en la frase “buenos días”

La frase "buenos días" fue elegida debido a que al contener las cinco vocales nos proporciona mayor información al obtener los componentes de estas señales, las imágenes que se muestran en todo el artículo pertenecen a este tipo de frases Fig. 2.

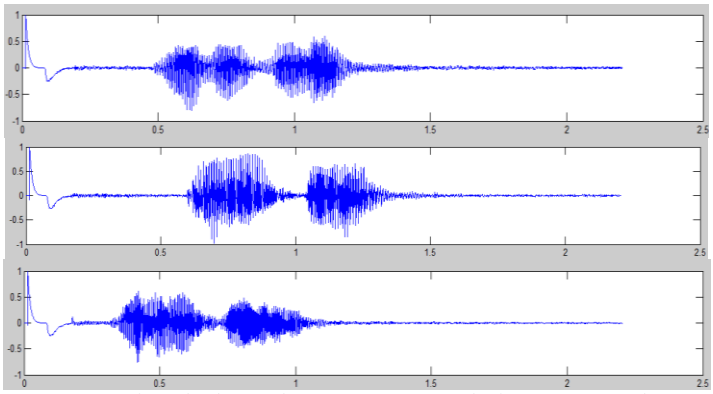

Figura 2. Ejemplo de señales que representan la frase "Buenos días".

Cabe mencionar que todo el proceso de reconocimiento de voz se realizó mediante el software Matlab lo que facilitó la implementación de algunos algoritmos y sobre todo la graficación de las señales.

2.2 Estimación y eliminación de ruido.

El ruido es uno de los principales problemas con el que nos encontramos al analizar señales de voz, el ruido son básicamente datos irrelevantes que pueden afectar a los datos que si nos interesan y que difícilmente podemos evitar durante la captura de audio, así pues es necesario comprar dispositivos que filtren estos datos innecesarios o bien encontrar algún método que nos permita eliminar o minimizar el ruido después de haber sido capturado.

Intencionalmente todas las muestras con las que se trabajó tienen ruido moderado (también conocido como ruido blanco), esto con la finalidad de que reducir el ruido de la señal sea parte de todo el proceso de reconocimiento

Para eliminar el ruido blanco se analizaron principalmente dos trabajos [5] y [1] en donde se proponen métodos que atenúan este ruido de los cuales se utilizaron dos para realizar pruebas:

- Método de Wiener. Que básicamente hace una estimación del ruido calculado de los momentos de menor intensidad o silencios, este método se aplico aumentando gradualmente el ruido para evaluar su desempeño frente al ocultamiento de los datos relevantes.
- Método de Boll. Éste método busca siempre reducir a cero el ruido estimado dejando únicamente fragmentos de sonido puros.

Una vez que se analizaron ambos métodos se pudo determinar que la mejor solución para eliminar este tipo de ruido es el método de Wiener, ya que se presenta me-

nor pérdida de información aún en sectores donde podría haber datos importantes.

2.3 Acotamiento de la señal.

Antes de poder trabajar realmente sobre los datos de los que se compone cada señal es necesario encontrar el punto exacto donde inicia y termina realmente el sonido que es útil, así un primer paso es eliminar el salto que existe al inicio de cada señal, este salto se origina al encender el micrófono y por lo tanto se considera un ruido constante con lo que basta una resta al inicio de cada señal para eliminarlo, una vez realizado esto las señales nos quedan como se muestra en la Fig. 3.

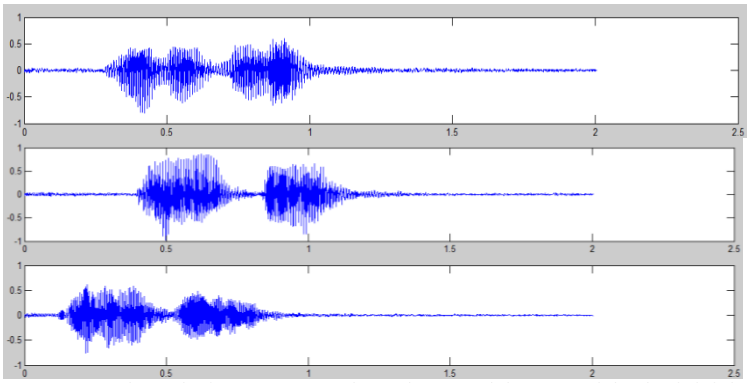

Figura 3. Señal con la frase "Buenos días" después del recorte del salto inicial.

Lo siguiente que hay que realizar es encontrar el punto en donde inicia y termina realmente cada señal de voz, eliminando todo lo demás ya que esto se consideran silencios que no aportan información útil. Para lograr esto se utiliza el cálculo de la energía de una señal, lo que da como resultado un espectro donde se puede notar con mayor facilidad los puntos donde inicia y termina la señal de voz mediante las siguientes formulas:

- La energía debe ser un vector que cumpla la condición de ser mayor a cero y menor a infinito, esto es:

$$0 < E < \infty$$

- La energía se calcula con:

$$E(n) = [(1 - \Upsilon) * E_{n-1}] + [\Upsilon * y^2_n]$$

Donde:
E=Es el vector de energía obtenido cuya primera posición vale 0
y=Es el vector de la señal original
γ=Es una constante definida como:

$$\gamma = 1000/(fr * 16)$$

Una vez calculados estos valores se obtiene gráficamente lo que se muestra en la Fig. 4.

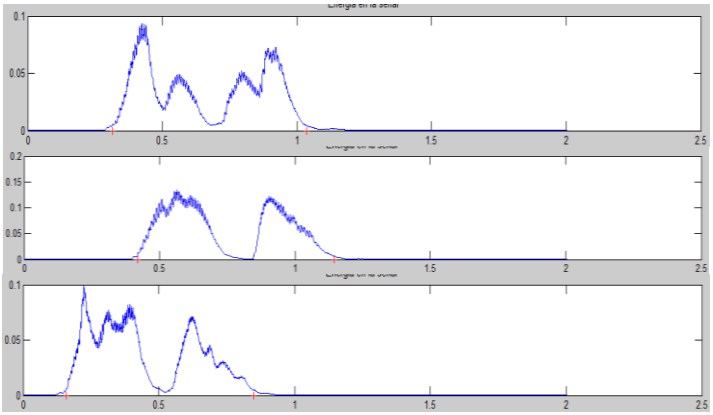

Figura 4. Espectro que muestra la energía de cada señal.

En base a estos datos es posible encontrar más fácilmente a partir de qué puntos se pueden considerar los datos de la señal como útiles o no ya que se logra ver más claramente que los valores innecesarios se aproximan mucho a cero y puesto que la energía de la señal son los valores respectivos a la señal original de voz se puede simplemente recortar los extremos de silencio y reflejarlos directamente en las señales de audio, lo que da como resultado lo que se observa en la Fig. 5.

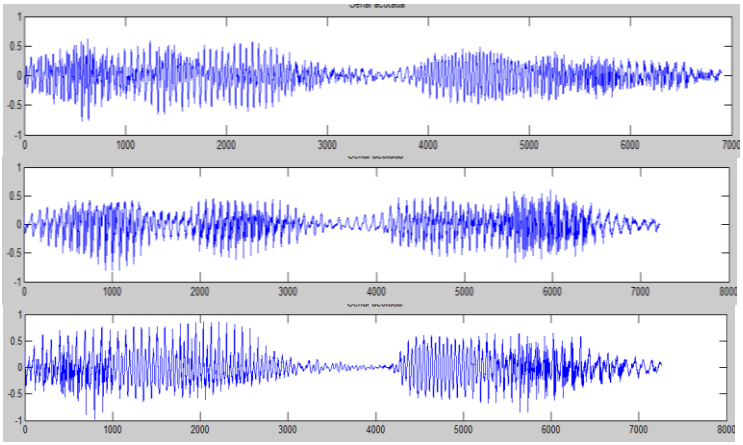

Figura 5. Señales después del corte de los extremos de silencio.

Ya podemos considerar que tenemos de las señales los datos que realmente nos aportan información pero como para cada señal los extremos que se eliminaron fueron cortados de diferentes puntos las señales que ahora tenemos son de diferente tamaño así pues realizamos una normalización en longitud para obtener señales de un mismo tamaño.

Las señales resultantes después de la normalización están compuestas de 6000 muestras cada una, con lo que ya es más sencillo realizar una buena comparación entre señales. Fig. 6.

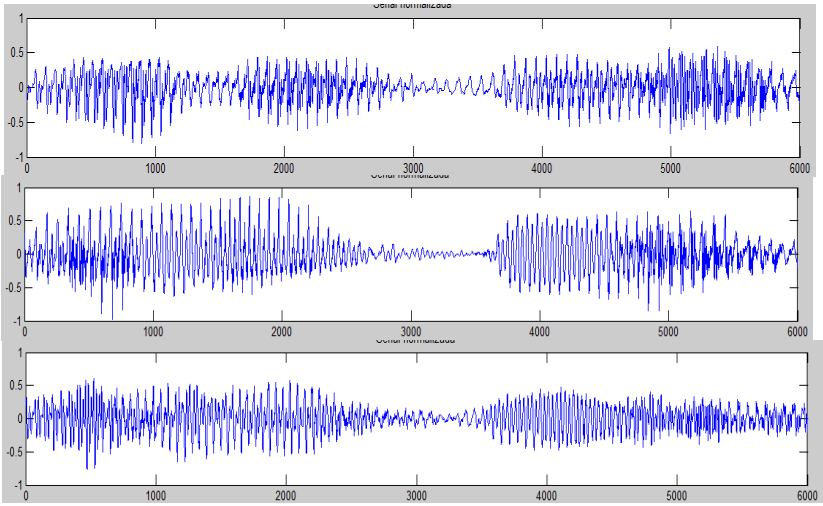

Figura. 6. Señales después de la normalización en longitud.

2.4 Obtención de LPC´s.

Muchos trabajos de investigación parten de la idea de que para realizar un buen reconocimiento de voz se deben calcular primero coeficientes LPC de cada señal y posteriormente usando estos, calcular coeficientes cepstrales lo que garantiza un mejor reconocimiento [3], aquí se plantea que es posible realizar de igual manera un buen reconocimiento utilizando únicamente los coeficientes LPC lo que también ahorra tiempo de computo que bien puede eliminarse o utilizarse posteriormente en la fase de entrenamiento y/o reconocimiento.

Es posible calcular un número N de coeficientes LPC para cada señal de voz pero aquí vamos a dividir primero cada señal en varias partes y después calcularemos LPCs para cada parte de cada señal lo que nos da no solo un arreglo de coeficientes sino una matriz. El cálculo de los LPC´s, en la implementación de una aplicación debe hacerse en tiempo real.

Para dividir cada señal se utilizar una técnica llamada enventanamiento (Fig. 7), la cual no solo divide la señal en varios fragmentos, sino que la multiplica a través de una función (en este caso es una gaussiana) de tal manera que resalte los elementos que se encuentren más al centro para cada ventana y también utilizaremos un solapamiento de ventanas (Fig. 8) para que los elementos que se encuentren sobre las orillas en una ventana puedan encontrarse en el centro de otras evitando así que perdamos información relevante que pueda estar presente en estos lugares.

Con estos dos métodos obtenemos para cada señal veintiocho ventanas de las cuales calcularemos 20 coeficientes para cada una, esto mediante el comando A=lpc(x, P) en donde x es el número de coeficientes que queremos obtener y P la ventana que está siendo evaluada en ese momento.

Finalmente tenemos para cada señal una matriz de valores que podemos utilizar para entrenar y clasificar nuevas entradas.

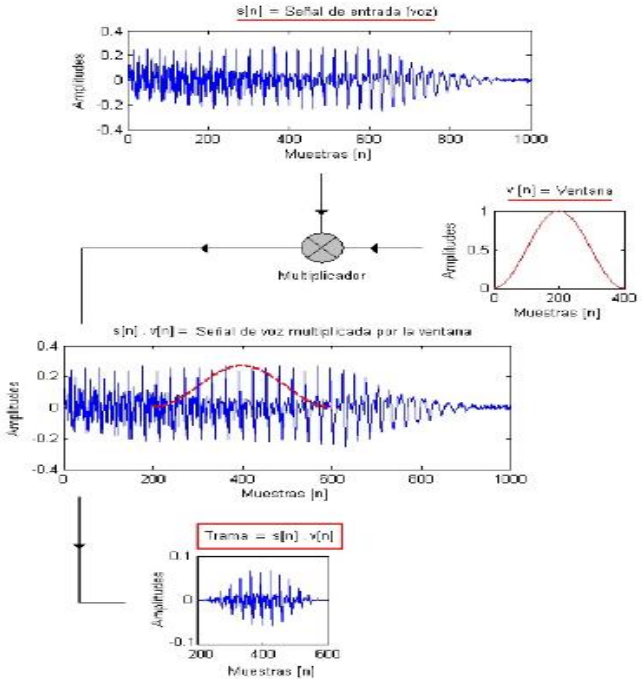

Figura 7. Diagrama representativo de enventanamiento.

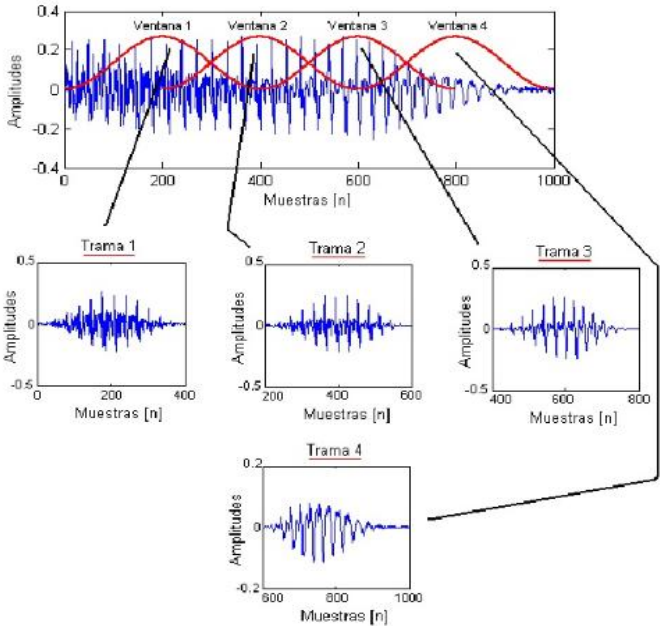

Figura 8. Diagrama representativo de solapamiento.

2.5 Entrenamiento y reconocimiento.

El reconocimiento y entrenamiento son dos etapas que van muy ligadas, es por eso que se utiliza un solo diagrama para explicar estas dos funciones (Fig. 9).

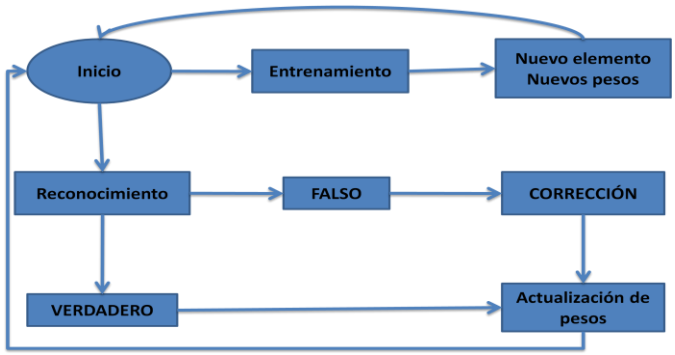

Figura 9. Diagrama general de entrenamiento y reconocimiento.

El entrenamiento se realiza únicamente para elementos nuevos y por lo tanto sus entradas son directas ya que de no ser así no podrá reconocer una nueva entrada que no se encuentre guardada anteriormente, o bien reconocerá a la más parecida cometiendo así un error.

El reconocimiento funciona de la siguiente manera:

- Primero la nueva entrada es descompuesta en una matriz de coeficientes tal y como se describió en secciones anteriores con el fin de tener elementos que comparar.
- La matriz de coeficientes de la nueva entrada es comparada contra las matrices guardadas mediante el algoritmo de KNN obteniendo los cinco elementos que más se parezcan.
- Mediante una probabilidad se elige uno de entre los cinco posibles elementos y este es el que arroja como salida.

Ahora bien una vez elegida la salida esta es evaluada y para ambos casos (Cierto o Falso) la matriz de coeficientes es considerada ahora como una matriz de pesos que se actualizaran tal y como lo harían los pesos de una red neuronal de tal modo que si es correcta la salida sea aún más fácil de identificar la próxima vez y en caso de que sea falsa para que pueda garantizar que la próxima vez acierte.

Resumiendo estos pasos tenemos un algoritmo que para reconocer utiliza: KNN para encontrar elementos parecidos, probabilidad para determinar la mejor solución y actualización de pesos para garantizar un mejor funcionamiento.

3 Conclusiones y Resultados.

Para poder realizar una buena comparación en cuanto a resultados no sólo se utilizó el método propuesto, también se hicieron pruebas con una red neuronal (ART) y con un mapa auto organizado (SOM).

A continuación se muestran los valores obtenidos para la palabra "nueve" del primer grupo de datos y para la frase "buenos días" del segundo conjunto.

TABLA DE RESULTADOS PARA LA PALABRA "NUEVE"

	SOM	ART	KNN
TIEMPO DE ENTRENAMIENTO	0.4 s	0.4 s	0.4 s
TIEMPO DE RECONOCIMIENTO	7 s	0.2 s	0.5 s
NO. DE ITERACIONES PARA RECONOCER CORRECTAMENTE	5	8	3
PORCENTAJE DE ACIERTO EN EL PRIMER INTENTO	60%	70%	85%
PORCENTAJE DE ACIERTO EN EL SEGUNDO INTENTO	70%	75%	92%

TABLA DE RESULTADOS PARA LA FRASE "BUENOS DIAS"

	SOM	ART	KNN
TIEMPO DE ENTRENAMIENTO	0.4 s	0.4 s	0.4 s
TIEMPO DE RECONOCIMIENTO	8 s	0.2s	0.5 s
NO. DE ITERACIONES PARA RECONOCER CORRECTAMENTE	3	5	2
PORCENTAJE DE ACIERTO EN EL PRIMER INTENTO	60%	60%	90%
PORCENTAJE DE ACIERTO EN EL SEGUNDO INTENTO	65%	70%	95%

Con estos resultados se puede observar que el primer método (SOM) tarda demasiado en reconocer una nueva señal, esto es debido a la separación que hace entre todos los datos de entrenamiento y más la competición que se realiza entre los elementos más cercanos al que se busca y aún así no realiza un correcto reconocimiento para ninguno de los dos casos.

En el caso de la ART se observa una disminución considerable en el tiempo de reconocimiento pero un aumento en el número de iteraciones necesarias para el mejor reconocimiento obtenido, que también es mejor en comparación con el obtenido en la prueba anterior.

Finalmente con el método propuesto con KNN se tiene un pequeño aumento en el tiempo de reconocimiento debido principalmente al cálculo de distancias pero que a fin de cuentas arroja un mejor resultado y en menos iteraciones que con cualquiera de los métodos anteriores.

Con todo esto se concluye que es posible realizar un reconocimiento eficiente utilizando únicamente los coeficientes LPCs, sin realizarles ningún otro procedimiento

posterior y que de los métodos propuestos el de KNN con actualización de pesos fue el mejor.

4 Trabajo futuro.

Puesto que la voz no es la única característica que identifica a una persona la metodología que aquí se presento puede ser utilizada para otro tipo de rasgo como podría ser el rostro, la retina, la huella digital, etc; aprovechando el mismo método de reconocimiento que aquí se presento. Por otro lado, siguiendo el área de reconocimiento de hablante es posible que los resultados puedan ser mejorados si las muestras que se adquieren son diferentes para cada persona, logrando una mayor diferencia entre individuos si aparte de las características únicas para cada individuo se toman en cuanta también diferentes frases para cada uno como podría ser el nombre hablado de cada quien.

Referencias

1. G. O. Borrás, "Reductor de ruido mediante resta espectral en entorno Matlab". EUIT Telecomunicación. 2006.
2. K. Daoudi, "State of the art in speech and audio processing". 2004
3. L. M. Hernández, A. E. Simancas, M. M Nakano, M. H. Pérez. "Reconocimiento de hablantes con dependencia del texto basado en LPC-Cepstral y red neuronal de retropropagación". SEPI ESIME Culhuacan, Instituto Politécnico Nacional. 2003.
4. C. A. Reyes, "Conceptos sobre Reconocimiento Automático del Habla". 2005
5. D. R. Tomassi, L. Aronson, C. E. Martínez, D. H. Milone, M. E. Torres y H. L. Rufiner, "Evaluación de técnicas clásicas de reducción de ruido en señales de voz". XV Congreso argentino de bioingeniería. Facultad de Ingeniería Universidad Naciónal de Entre Ríos. 2004.