Reconocimiento de patrones auditivos en ambientes ruidosos (Dic 2009)

Belmer Alberto Cordoba Diaz

***Resumen*— Este proyecto tiene como objetivo aportar una solución al reconocimiento de patrones fonéticos en un entorno ruidoso. El cálculo de la trasformada discreta de Fourier a corto plazo da como resultado coeficientes que son organizados por su densidad espectral de energía, que permite la obtención de su envolvente espectral para así generar su espectrograma, el cual muestra su patrón característico, logrando así poder realizar la comparación de dichos patrones y concluir la existencia o no del patrón fonético.**

I. INTRODUCCIÓN

En los sistemas de reconocimiento de voz no se intenta reconocer el sonido del fonema, sino identificar una serie de características principales para saber si el locutor dijo lo que se presume. El tamaño de la "frase" en el reconocimiento de voz afecta su complejidad.

El comportamiento de los sistemas de reconocimiento del habla se degrada rápidamente debido a la presencia del ruido de fondo, recientemente se ha propuesto una técnica de representación de la señal de voz basada en predicción lineal. Que ha mostrado ser atractiva para el reconocimiento de señales de audio en condiciones severas de ruido gracias a su simplicidad computacional.

El problema del reconocimiento de voz permanece sin resolver aun en caso de palabras aisladas y vocabularios pequeños. Por esta razón se ha propuesto diversas técnicas de reducción de ruido en cada una de las etapas del proceso de reconocimiento especialmente en extracción de parámetros fonéticos y medidas de similitud.

La etapa de parametrización es dada por un código de predicción lineal. Usado ampliamente en reconocimiento de sonidos fonéticos, este código es sensible al ruido blanco, pero aun así esta técnica es favorable respecto a otras técnicas de reconocimiento auditivos como lo son, bancos de filtros y vector de cuantización. Con esto la técnica de predicción lineal en combinación con el procedimiento matemático de Autocorrelación permite una aproximación a un buen proceso de reconocimiento de patrones de audio.

El procedimiento de desarrollo del proyecto se llevará a cabo inicialmente con un estudio de las técnicas de reconocimiento observando ventajas que pueda tener alguna sobre los demás, además si se pueden mezclar para obtener mejores resultados en el proceso de reconocimiento.

II. CONCEPTOS BÁSICOS

A. *Marco teórico*

La voz es una mezcla de sonidos y se realizan a través de un proceso donde se intervienen el tracto vocal y el tracto nasal. El primero está compuesto por la apertura de las cuerdas vocales y finaliza en los labios, y consiste en la conexión del esófago con la boca donde el área seccional cruzada del tracto vocal determina en qué posición están la lengua, la mandíbula, los dientes y los labios. El segundo inicia en el velo y finaliza en la nariz. Cuando el velo (mecanismo situado en la parte posterior de la cavidad vocal) es reducido, el tracto nasal esta acústicamente acoplado al tracto vocal y produce el sonido nasal de la voz [6].

Cuando el flujo del aire es expelido por los pulmones a través de la tráquea, la elasticidad de las cuerdas vocales dentro de la laringe vibra por el flujo del aire, el cual es fraccionado en pulsos casi periódicos los cuales luego son modulados en frecuencia y pasados a través de la faringe, la cavidad vocal y posiblemente la cavidad nasal [6].

B. *Código de predicción lineal (LPC)*

Dado que LPC es capaz de extraer la información lingüística y eliminar la correspondiente a la persona particular. La predicción lineal modela la zona vocal humana como una respuesta al impulso infinita, que produzca la señal de voz [6].

El término predicción lineal se refiere al método para predecir ó aproximar una muestra de una señal en el dominio del tiempo s[n] basada en varias muestras anteriores s[n - 1], s[n - 2], s[n - M] (ecuación 1) [6].

$$s[n] \approx \hat{s}[n] = -\sum_{i=1}^{M} a_i s[n-1] \qquad (1)$$

Donde s[n] es llamada señal muestreada, y a_i, i = 1,2,…, M son los predictores ó coeficientes LPC. Un pequeño número de coeficientes LPC $a_1, a_2, \ldots, a_M$ pueden ser usados para representar eficientemente una señal s[n]. Los valores a_1 $a_2, \ldots, a_M$ son la base para la realización de este trabajo debido a que nos ayudan a modelar los parámetros de la voz de cada uno de los hablantes que se emplean en este sistema propuesto [6].

Este modelamiento permite simular el tracto vocal de una persona por medio de una función de transferencia la cual está compuesta solo por polos, es decir hace el numerador igual a 1 como muestra la ecuación 2, donde $u(n)$ es una es una excitación normalizada y G la ganancia de esa excitación [6].

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1-\sum_{i=1}^{p} a_i z^{-1}} = \frac{1}{A(z)} \quad (2)$$

Esta función de transferencia se ve expresada en la figura 1 por medio de un diagrama de bloques.

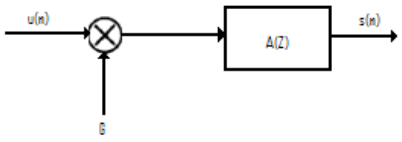

Figura 1. Representación de la función de transferencia que simula el tracto vocal.

C. Banco de filtros

Los bancos de filtros son un conjunto de filtros utilizados en reconstrucción perfecta de señales, los cuales están vinculados por muestreo de operadores y algunas veces por retardos. El bajo muestreo de los operadores es reducido en décadas, mientras que los altos muestreos son expandidos. En un banco de filtros de dos canales, el análisis de los filtros es pasa bajos o pasa altos respectivamente [8] tal y como se observa en la figura 2.

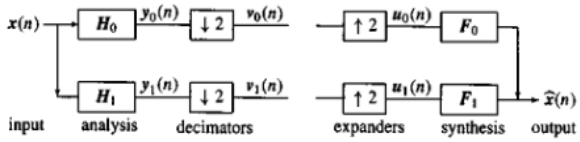

Figura 2. Banco de filtros (análisis y síntesis). [8]

Los bancos de filtros se implementan para realizar descomposiciones de la señal original discretizada sobre un mismo segmento del espectro, es decir separar las frecuencias altas de las bajas, con el fin de buscar información dentro de las nuevas frecuencias obtenidas que permitan un reconocimiento del patrón auditivo. Cuando se somete la señal a un banco de filtros se obtiene (ecuación 3).

$$y_i(n) = x(n) * h_i(n) \quad (3)$$

Donde

$$y_i(n) = \sum_{m=0}^{M_i-1} h_i(m)s(n-m) \quad (4)$$

D. Filtros MEL

Existe una amplia variedad de parámetros que pueden utilizarse para tratar de capturar las características mas sobresalientes de la señal: los coeficientes de la FFT de la señal, los coeficientes del espectro, los coeficientes espectrales en la escala de MEL, la energía de la señal, etc. Todos ellos se han empleado individualmente o en combinación con otros sistemas de reconocimiento de características de fonemas [7].

Escala MEL: es bien conocido que el oído humano presenta una escala perceptual logarítmica en frecuencias, esto motiva a que algunos sistemas de reconocimiento utilicen una transformación del eje de frecuencias para adecuarlo a la escala perceptual. La escala MEL es una aproximación a la escala perceptual humana [7].

$$.Mel(f) = 2595 * log\left(1 + \frac{f}{700}\right) \quad (5)$$

MEL-Cepstrum: los coeficientes cepstrum en escala MEL han mostrado en el reconocimiento de patrones auditivos unas buenas prestaciones respecto a otras técnicas de parametrizacion. Para el cálculo se utiliza normalmente un número de filtros triangulares. Estos filtros están igualmente espaciados en la escala MEL de frecuencias la idea que da origen a esta familia de parámetros es la obtención de vectores de coeficientes cepstrum en los cuales es espaciamiento de frecuencias no es lineal, sino que se distribuye en la escala perceptual MEL [7] tal y como se ve en la figura 3.

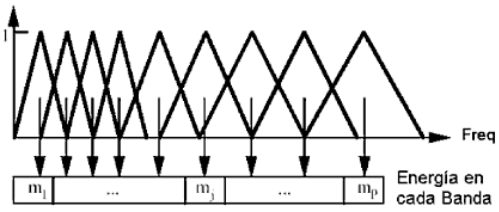

Figura 3. Banco de filtros en la escala MEL. [7]

III. DESARROLLO

A. Captación de la señal de audio

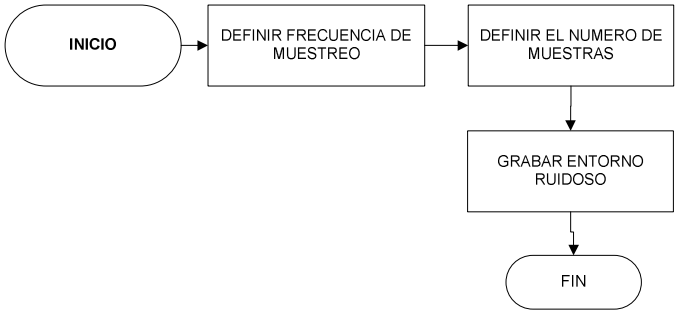

Figura 4. Secuencia captura de voz.

Dentro del proceso de captura de voz en Matlab como primera medida hay que definir la frecuencia de muestreo Fs, para esto es necesario conocer unos rangos de frecuencia o anchos de banda como:

- Ancho de banda de la voz.
- Ancho de banda del oído humano.
- Espectro electromagnético del audio.

Para el desarrollo del proyecto requirió dimensionar que rango de frecuencias no fueron tenidas en cuenta. De esta forma se debió conocer cuáles son los respectivos anchos de banda, vale la pena aclarar que estos rangos se encuentran dentro de VLF (Very Low Frecuency).

Espectro electromagnético del audio: está compuesto por frecuencias que varían desde los 3 Hz y los 30 KHz.

Ancho de banda del oído humano: de una forma ideal se dice que el oído humano percibe frecuencias entre 20 Hz y los 20 KHz.

Ancho de banda de la voz: este rango de frecuencias se encuentra entre los 200 Hz y los 4 KHz.

Ya cuando se obtuvo esta información se definió una frecuencia de muestreo Fs=20000 para cumplir con el ancho de banda del oído así sea de una forma ideal. Posteriormente a la selección de la frecuencia de muestreo fue necesario definir un tiempo de captura, el cual depende de la duración del archivo de audio que deseemos obtener.

Cuando se habla de la definición del número de muestras simplemente se realizó como el producto entre la frecuencia de muestreo con el tiempo de captura definido. Y finalmente con el comando wavrecord en Matlab proceder a la grabación del archivo de voz.

En cuanto a la creación y lectura de un archivo .WAV simplemente fue posible gracias a la implementación de dos comandos que ofrece el toolbox de procesamiento digital de señales.

Para la creación de archivo .WAV se planteó el comando wavwrite el cual junto con los datos recolectados con wavrecord y la frecuencia de muestreo genera un archivo con el nombre que se deseo, con extensión .WAV. Para este caso el archivo generado es EntornoRuidoso.wav como se ve en la figura 5.

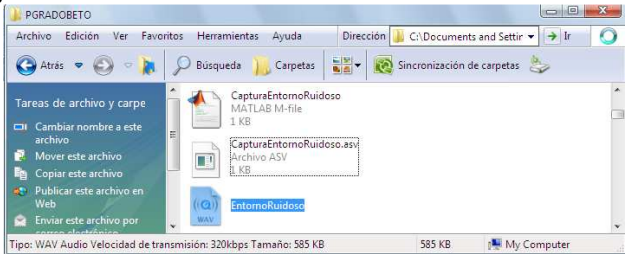

Figura 5. Pantallazo creación del archivo de audio.

Luego de haber creado el archivo, se procedió con la lectura del mismo, esto se realizó con el comando wavread el cual guarda cada muestra de la señal en una posición de un vector.

Para graficar la señal se tiene que por defecto Matlab asignará valores máximo entre 1 y -1 para las amplitudes de cada una de las muestras (Figura 6). Con esto el trabajo restante para graficar la información del archivo de audio fue asignar el valor de la muestra con respecto al número total de las mismas.

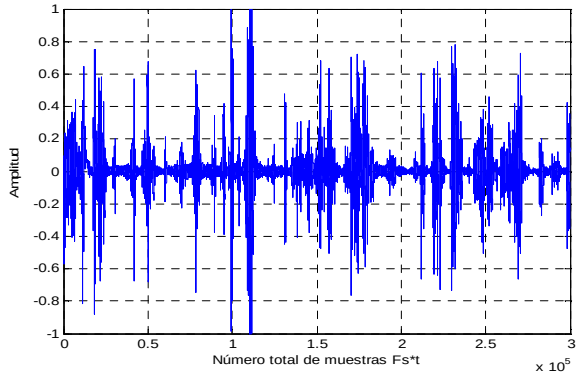

Figura 6. Señal de voz en tiempo discreto.

B. Filtrado de la señal de audio

Debido a que en el proyecto la captura de voz se realizó en un entorno donde no solo existía esta señal, esto produce que ruido aditivo generado por todas las perturbaciones alrededor del transductor o micrófono sean inherentes y afecten las señales de voz y el reconocimiento de las mismas.

Es por eso que inicialmente para solucionar este problema se diseño un filtro cuyas frecuencias de corte fuesen lo más similares a las del ancho de banda de la voz (200Hz – 4KHz). Con el fin de suprimir todas aquellas frecuencias que no se encontraran dentro de este rango y que pudiesen afectar posteriormente el reconocimiento de los patrones de voz.

Gracias a que Matlab ofrece diversas herramientas de diseño para múltiples necesidades se recurrió al uso del toolbox de diseño de filtros. Esta herramienta tiene como nombre Filter Desing & Analysis Tool. La ventana que ofrece la herramienta de diseño del filtro se observa en la figura 7:

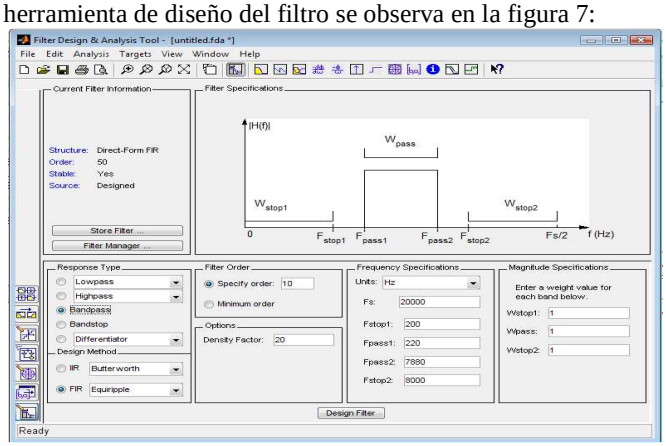

Figura 7. Pantallazo toolbox de diseño de filtros analógicos y digitales.

Para diseñar el filtro se definen las frecuencias de corte, se especifica el orden, si se quiere pasa banda o rechaza banda, si se quiere el filtro analógico o digital, además se puede observar que esta herramienta permite mostrar la respuesta en magnitud, la respuesta en fase, diagrama de polos y ceros, la respuesta al pulso y la respuesta al paso.

El filtro que se diseño para suprimir las frecuencias fuera del ancho de banda de la voz tiene los siguientes parámetros:

- Filtro digital FIR Equiripple.
- Pasa banda.
- Orden 10.
- Factor de densidad 20.
- Frecuencia Stop1 200Hz.
- Frecuencia Pass1 220Hz.
- Frecuencia Pass2 7880Hz
- Frecuencia Stop2 8000Hz
- Frecuencia de muestreo 20000Hz

Con esto el resultado respectivo del filtrado para una señal de audio se observa en la figura 8 donde se representa la señal de audio original por medio del color amarillo y por medio del color morado se ve la señal filtrada.

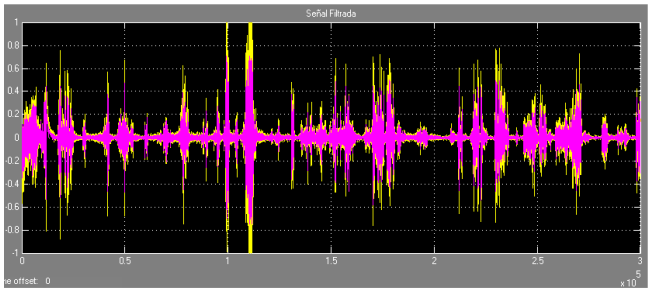

Figura 8. Filtrado de la señal de audio.

Luego del respectivo filtrado de la señal capturada se procedió a un análisis sinusoidal el cual está compuesto de una serie de modelos matemáticos los cuales representan la implementación del algoritmo para el reconocimiento de patrones de fonéticos.

C. Análisis sinusoidal de la voz

Para un buen reconocimiento de los patrones de audio se debe tener en cuenta que para la implementación del algoritmo existen factores que pueden implicar que el reconocimiento sea más complejo. A continuación de mencionarán algunos de estos factores:

- Tamaño de la frase: implica que entre más larga sea la frase más difícil es el reconocimiento.
- Locutor: esto debido a que uno no pronuncia las palabras siempre de la misma forma. Esto incluye si la palabra va al inicio, en medio, o al final de la oración.
- Entorno físico: esto se debe al hecho de que no es lo mismo un sistema que funciona en un ambiente poco ruidoso, o por el contrario en un ambiente ruidoso.

Debido a que la señal de voz no posee frecuencias sinusoidales fijas, esto las hace no estacionaria y no lineal. Para esto existen métodos que permiten representar las señales no estacionarias y no lineales como la suma de componentes que intervienen en la señal, generalmente segmentos de señal que no están afectados o que no presentan algún tipo de degeneración. En la figura 9 se ve representado la secuencia de trabajo para la obtención de los espectrogramas.

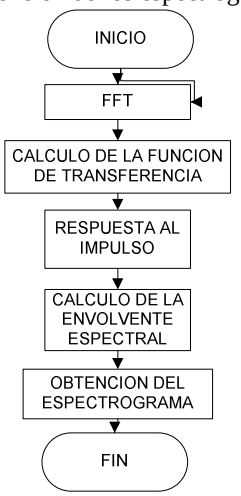

Figura 9. Obtención del espectrograma.

Inicialmente la señal de voz en entorno ruidoso y en entorno sin ruido fue muestreada pasándola del dominio del tiempo continuo al dominio del tiempo discreto, cumpliendo con la teoría del muestreo de Nyquist, la cual dice que para que una señal sea muestreada con una buena cantidad de información la frecuencia de muestreo debe ser por lo menos del doble en comparación de la señal en el tiempo continuo, por ejemplo el ancho de banda de la voz es de 4KHz por tanto la frecuencia de muestreo debe ser de mínimo 8KHz. Esto se hace para que el cálculo de DTF (Transformada Discreta de Fourier) tenga un resultado coherente.

Ahora bien, prosiguiendo con el desarrollo se creó el archivo de audio tanto para la muestra de voz en un entorno ruidoso, como en un entorno sin ruido esto para poder realizarle los procesamientos necesarios dentro de los cuales se encuentran la lectura de los valores que representan la amplitud por cada instante de tiempo, el filtrado de la voz para garantizar el trabajo solo con las características de la señal que se encuentren entre los 200Hz y los 4Hz y poder suprimir algunas degradaciones que se presenten en la señal muestreada que se encuentren por fuera del ancho de banda, claro que este filtrado no garantiza que la señal resultante ya esté libre de degradaciones ya que se pueden presentar perturbaciones de baja frecuencia al interior del rango de frecuencias que maneja la voz.

El algoritmo parte del cálculo de la función de trasferencia del cuantizador de la voz (contador de muestras de la señal de voz), con el fin de poder calcular la respuesta al impulso para determinar la envolvente del espectro y por último a partir de estos, graficar el espectro de voz. La figura 10 muestra el cálculo de la FFT.

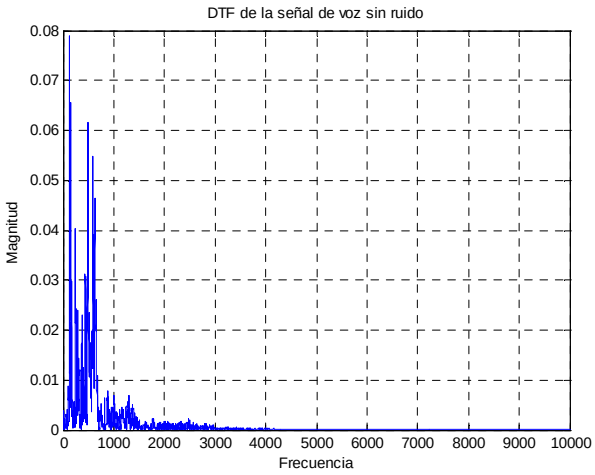

Figura 10. FFT de la señal de audio.

Es aquí donde se determina el comportamiento de la envolvente del espectro que se generará, ya que se puede determinar la posición en frecuencia de las componentes formantes de la voz (estas son las que tienen mayor valor de amplitud) y son las que rigen el comportamiento espectral de la voz.

El espectrograma generado a través del espectro nos indica que en ciertas porciones de la señal se encuentra la mayor concentración de energía, es decir, los espacios en que las formantes son más grandes obteniendo valores pronunciados de magnitud donde la mayor cantidad de densidad espectral de

energía se denota con el color rojo, por el contrario en el lugar donde se representa la menor cantidad de densidad espectral de energía se representa con un color azul (figura 11).

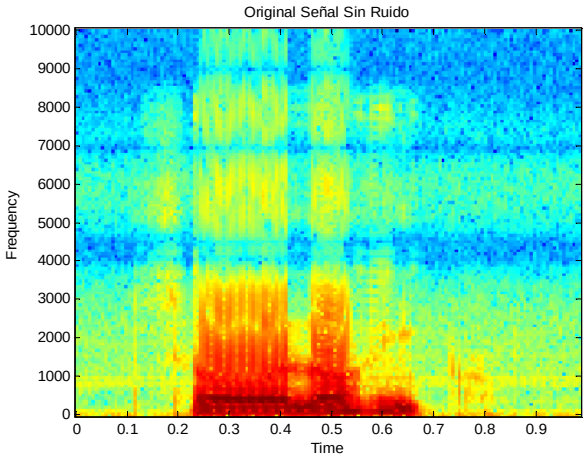

Figura 11. Espectrograma de la señal de audio.

Vale la pena resaltar que un espectrograma de frecuencias es el resultado de calcular el espectro en tramas de ventaneo de una señal, esta se representa en una grafica tridimensional en donde se muestra la energía del contenido en frecuencias de la señal según va variando esta a lo largo del tiempo.

El espectrograma sirve para analizar la sonoridad, la duración, la estructura de los formantes (timbre), la intensidad, las pausas, y el ritmo.

Puede observarse que en el espectrograma mostrado en la figura 11 existen dos bandas de frecuencia fundamentales que es donde se presenta la mayor cantidad de energía. Esta banda de frecuencia se encuentra entre los 0Hz y los 4000Hz.

D. Identificación del patrón de audio

Una vez obtenido el patrón se procede a desplazarlo por toda la señal, muestra a muestra, calculando la correlación cruzada en un instante dado entre el patrón y un segmento de señal de su misma longitud. De esta forma, calculamos en cada instante la similitud entre un trozo de señal y el patrón determinado previamente con poca degradación por el ruido. Lo que quiere decir que el resultado de la comparación presentará valores pequeños en lugares donde el parecido es mayor (zonas donde presumiblemente hay poca degradación) y presentara valores grandes donde el parecido es menor (zonas de una mayor degradación). Este proceso se observa en la figura 12.

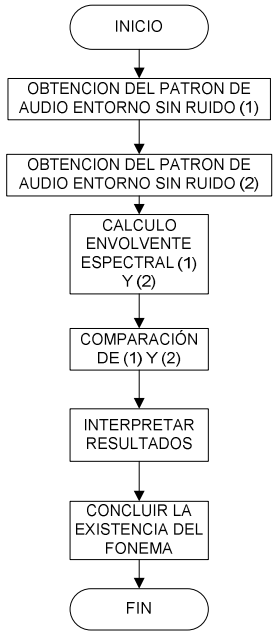

Figura 12. Secuencia de identificación del patrón de audio.

A. Interfaz gráfica

La figura 13 muestra la interfaz grafica que se diseño y se implemento por medio de una herramienta llamada Guide Quick Start.

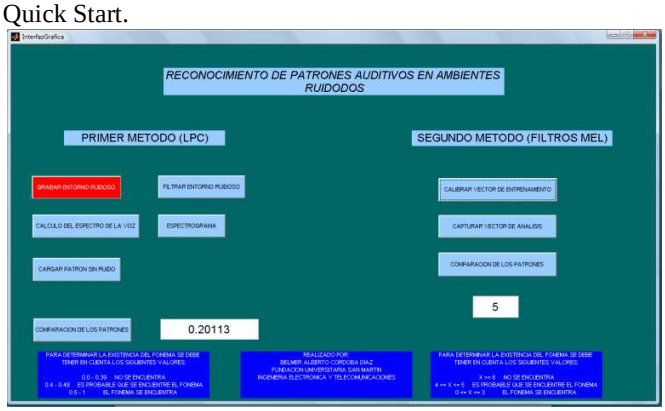

Figura 13. Interfaz gráfica del proyecto.

En esta interfaz gráfica se realizan los procesos de captura de voz, filtrado de la señal de voz, cálculo del espectro de la señal, generar el espectrograma, comparar los patrones para demostrar la existencia del patrón fonético. Esto para el método LPC.

Para el método de Filtros MEL la interfaz grafica permite generar un vector de entrenamiento mediante la obtención de 10 muestras en un entorno sin ruido, capturar una señal de audio en un entorno ruidoso para generar un vector de análisis y por último calcular la existencia del patrón fonético por medio de la comparación de los dos vectores (entrenamiento y análisis).

IV. CONCLUSIONES

- El software de grabación de voz, como el que se encuentra predeterminado por Windows, maneja una extensión para los archivos de audio capturados conocido como .wma (Windows Media Audio). El problema de

esta extensión es que ya trae un códec de compresión de audio que genera pérdidas de información, la cual puede ser relevante al momento del cálculo del espectro para la obtención de las formantes características. Por el contrario la extensión .wav (Waveform Audio Format) es la representación con menos pérdidas en la captura de sonidos en el dominio del tiempo discreto.

- La selección de una adecuada frecuencia de muestreo es de vital importancia, ya que de esta depende la cantidad de información que tenga la señal capturada por esto es necesario el cumplir con el teorema de muestreo de Nyquist.
- La implementación de filtros digitales es mucho más efectiva que la implementación de filtros analógicos por dos razones. La primera es que la respuesta transitoria de los filtros análogos es más lenta que la de los filtros digitales y esto se ve reflejado en la respuesta al impulso y la respuesta al paso del filtro digital. La segunda por que los filtros análogos presentan un cambio de fase en algún instante de tiempo perjudicando la información que posea la señal después del filtrado.
- La transformada discreta de Fourier es una herramienta muy necesaria para poder obtener la distribución espectral de energía, debido a que esta permite encontrar en que sectores de la señal de audio en el dominio de la frecuencia se encuentran los coeficientes más representativos o formantes, por medio de los cuales se genera la envolvente espectral.
- A pesar de que la transformada discreta de Fourier a corto plazo es una buena herramienta para identificar los picos representativos de la característica, esta posee una serie de limitaciones, ya que cuando hay formantes que cambian abruptamente en intervalos cortos de frecuencia la predicción de la envolvente del espectro no es tan efectiva. Por lo que se recomienda implementar otro tipo de herramientas matemáticas que se apliquen al procesamiento digital de señales en el reconocimiento de los patrones de audio.
- Otro método de reconocimiento propuesto pero con deficiencias es el método de Vector de cuantización, en el cual existe una distorsión espectral inherente en el vector de análisis o vector donde se encuentra la muestra de audio en el entorno ruidoso.
- Debido al trabajo con los códigos de predicción lineal LPC los cuales se encargan de la simulación del tracto vocal por medio de un filtro llamado todo polos, fue posible la obtención del espectrograma, el cual es la representación del resultado de calcular el espectro en tramas de una señal, es decir básicamente representar el contenido en frecuencias de la señal conforme varia el tiempo, esta representación esta codificada en colores, donde el color rojo representa la mayor cantidad de densidad espectral de energía.
- Aunque todos los métodos de reconocimiento de patrones auditivos buscan realizar la comparación de dos vectores característicos (análisis y entrenamiento), la implementación del método de filtros MEL es más efectivo al nivel del reconocimiento, ya que este representa la fusión de dos métodos (bancos de filtros y la variación logarítmica de la escala MEL) haciéndolo más robusto.

TRABAJO FUTURO

- Analizar de una forma correcta las formantes del espectro de la señal de audio para así, poder determinar la huella digital que poseen cada uno de los seres humanos en el tracto vocal a través de un espectrograma de frecuencias.
- Debido a que este proyecto se encarga de reconocer palabras, una recomendación es implementar este reconocimiento de fonemas para generar una señal de control en la maniobrabilidad de ciertos dispositivos.
- Gracias a que este proyecto logró reconocer fonemas satisfactoriamente a través de un procesamiento digital de señales donde se implementaron una serie de algoritmos matemáticos, este mismo reconocimiento puede ser realizado a través de un procesador digital de señales (DSP) para lograr su implementación en hardware.
- Para un mejor cálculo de las formantes en la obtención de la envolvente espectral de una señal de audio este puede ser efectuado por medio de análisis Wavelet y autómatas celulares.

REFERENCES

[1] ERICSSON TEMS AB, AQM in TEMS Automatic-PESQ Technical Paper. Suecia: Trademark owned by Telefonaktiebolaget L M Ericsson, 2006.

[2] FERNANDEZ Rubio, Juan A. Comunicaciones Analógicas. Bogotá: Ediciones UPC, 1999.

[3] GABIOLA, Francisco J. y AL-HADITHI, Basil M. Análisis y Diseño de Circuitos Analógicos. Madrid: Editorial Visión Libros, 2007.

[4] GARCIA Luz, DE LA TORRE Angel, BENITEZ Carmen y RUBIO Antonio J. Speech Recognition, Technologies and Applications. Vienna, Printed in Croatia: Edited By France Mihelič and Janez Žibert, 2008.

[5] MARTI, M. Antonia y LLISTERRI, Joaquim. Tecnologías del Texto y Habla. Barcelona: Edicions Universitat Barcelona, 1996.

[6] RABINER, Lawrence y JUANG, Biing-Hwang. Fundamentals of Speech Recognition. New Jersey: Prentice-Hall, 1993.

[7] SACEDO, Francisco. Modelos ocultos de Markov: del reconocimiento de la voz a la música. Valencia: universidad de Valencia press, 2002.

[8] TRUONG, Nguyen y STRANG, Gilbert. Wavelets and Filter Bank. Wisconsin: Wellesley-Cambridge press, 1996.

Belmer Alberto Cordoba Diaz: Nació en Bogotá el 1 de junio de 1985. Realizó su bachillerato en el Colegio Militar Simón Bolívar donde obtuvo su título de bachiller académico en el año 2001, actualmente es estudiante de ingeniería electrónica y telecomunicaciones de la Fundación Universitaria San Martin.