Reconocedor de voz mediante el uso de la *FFT*

Andrés Fuentes Hernández (*)
M. en C. Álvaro Anzueto Ríos (*)
**Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas*
UPIITA-IPN

Resumen – *En el presente trabajo se describe la implementación de un reconocedor de voz con el software MATLAB, el cual llega a distinguir hasta cuatro palabras. El proceso inicia cuando el sonido de un vocablo es capturado por el ordenador, acto seguido la señal es procesada y se segmenta la parte de interés, para después obtener de ella la Transformada Rápida de Fourier (por sus siglas en inglés: FFT) y, finalmente, mediante el promediado de señal, el resultado es comparado en la base de datos que contiene los patrones a reconocer*.

I.- Introducción

Hoy en día los sistemas de reconocimiento de voz buscan sustituir la gran mayoría de los dispositivos de entrada —teclado, *mouse*, etc.—. Representan una gran ventaja en movilidad, sin embargo, su empleo requiere un entrenamiento previo —calibración—, con el fin de que puedan distinguir los vocablos utilizados por el usuario [1].

El reconocimiento de voz no es un tema nuevo, y los algoritmos para su implementación han variado junto con la capacidad de procesamiento de los computadores. Esto se puede encontrar en el libro escrito por Tod Loofbourrow en 1978, en el cual, se describe un algoritmo rudimentario para el reconocimiento de 2 instrucciones. Este método consiste en dividir la voz en frecuencias bajas y frecuencias altas, y cuantificar los cruces por cero realizados por cada señal para un procesamiento posterior. Este método no es del todo eficaz, ya que como él autor describe, es imperativo que la palabra sea pronunciada con el mismo tono y timbre en cada ocasión para que el sistema la reconozca [2]. Este trabajo presenta un sistema robusto a los cambios de tono y timbre.

Actualmente se usan métodos estadísticos —LPC—, además de la implementación de reconocedores basados en redes neuronales [3]. El método de obtención de patrones presentado en este artículo es la *FFT*, aplicada a la señal de voz segmentada y filtrada. El reconocedor propuesto es la correlación entre señales, el cual arroja datos de: -1 a 1, dependiendo de qué tan parecida es una señal con la otra; -1 es la misma señal, pero, invertida y, 1 es una señal similar al patrón.

II.- Metodología

Haciendo uso de un programa implementado en MATLAB: se graban dos segundos de audio con una frecuencia de muestreo de 20KHz. La grabación da como resultado un vector de 40 mil datos, de los que se discriminarán los datos significativos mediante un umbral de 0.1. Con base en el vector de datos obtenidos, se realiza el siguiente procesamiento:

1.- Aplicar el filtro de preénfasis para acentuar las frecuencias altas de la señal de voz, debido a que el modelo de tracto vocal utilizado no filtra adecuadamente las señales de frecuencia alta —no sonoras: las consonantes— a diferencia de las de baja frecuencia —sonoras: las vocales—. El filtro de preénfasis está dado por la ecuación 1.

$$\tilde{s}(n) = s(n) - a * s(n-1) \text{.....} \textbf{(1)}$$

n= 1,2,3.....N
a: es un coeficiente usualmente de 0.95 a 0.97
N: total de datos en el vector

2.- Segmentar la señal en tramos de 20 a 30ms, puesto que en este período de tiempo se considera a la señal de voz como estacionaria. En este proyecto, se fracciona en tramos de 24ms, por lo tanto, cada vector resultante tiene 480 datos y, se analiza utilizando un *overlap* de 80 datos. Lo anterior se expresa matemáticamente en la ecuación 2 y, se ejemplifica en la figura 1.

$$x_i(k) = \tilde{s}(Ml + k) \text{.....} \textbf{(2)}$$

donde **k=**1,2,3.....K-1 y **l=**1,2,3....L-1

K: total de datos por cada tramo
L= total de tramos
M= *overlap*

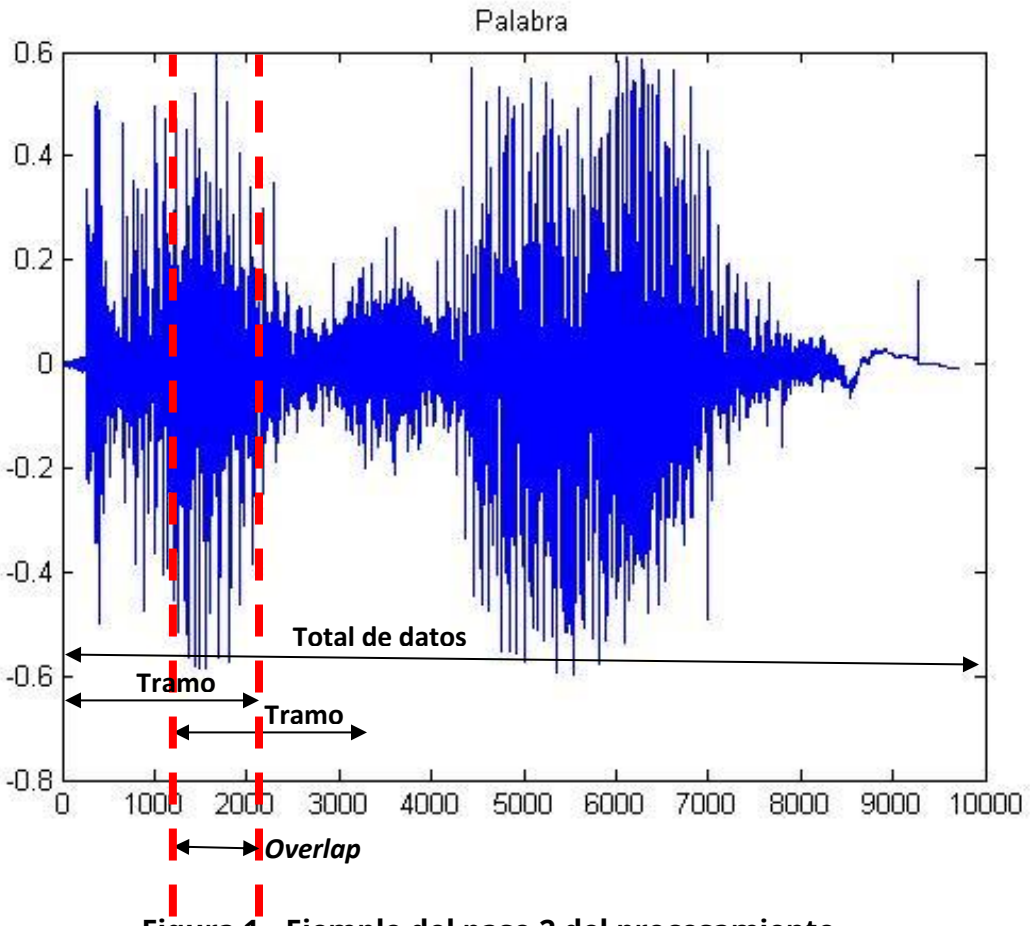

Figura 1.- Ejemplo del paso 2 del procesamiento.

3.- Aplicar una ventana *Hamming* a cada tramo de datos, para así eliminar las discontinuidades provocadas al segmentar, ya que éstas podrían llegar a interpretarse como altas frecuencias. En la ecuación 3 se resume lo antes mencionado.

$$\widetilde{x_i}(k) = x_i(k)w_i(k) \text{.....} \mathbf{(3)}$$

i= 1,2,3....L y **k=**1,2,3....K

donde **w=**ventana de *Hamming*

4.- Obtener la *FFT* de cada tramo—ecuación 4— [4], con el objetivo de generar una superficie en la que se pueda observar las frecuencias y su variación en el tiempo —figura 2—. Se promedian las **FFT** de cada tramo, para obtener un patrón de la palabra pronunciada —figura 3—.

$$X(k) = \sum_{n=0}^{N-1} X(n)e^{-j2\pi(\frac{Kn}{N})} \text{.....} \mathbf{(4)}$$

n=0,1....N, k=0,1....N

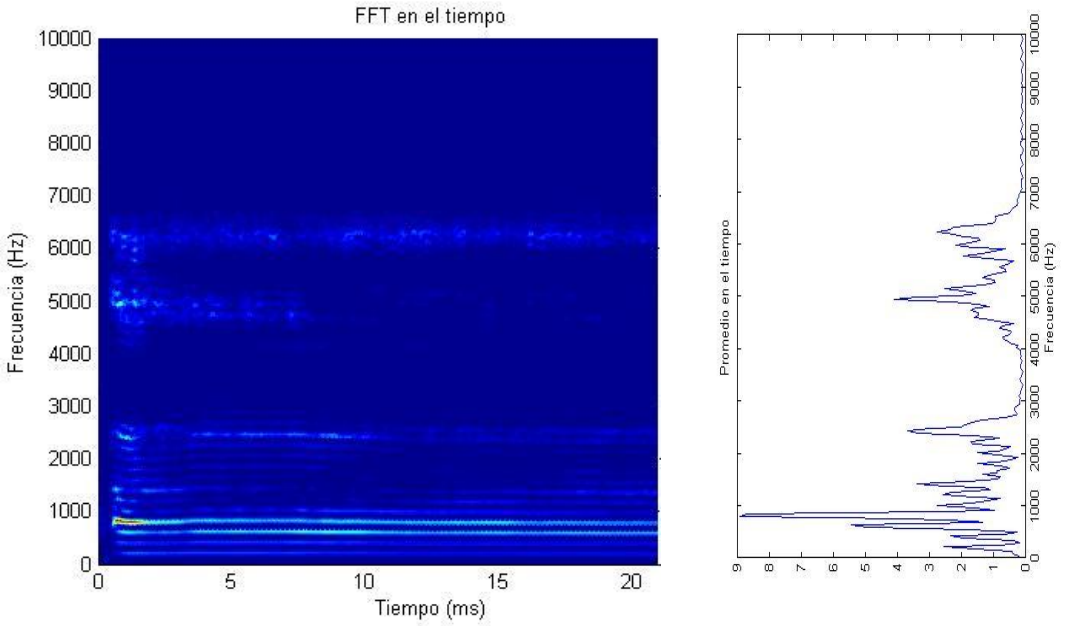

Figura 2.- *FFT* en el tiempo de la letra 'a' pronunciada por 21ms y su promedio en el tiempo

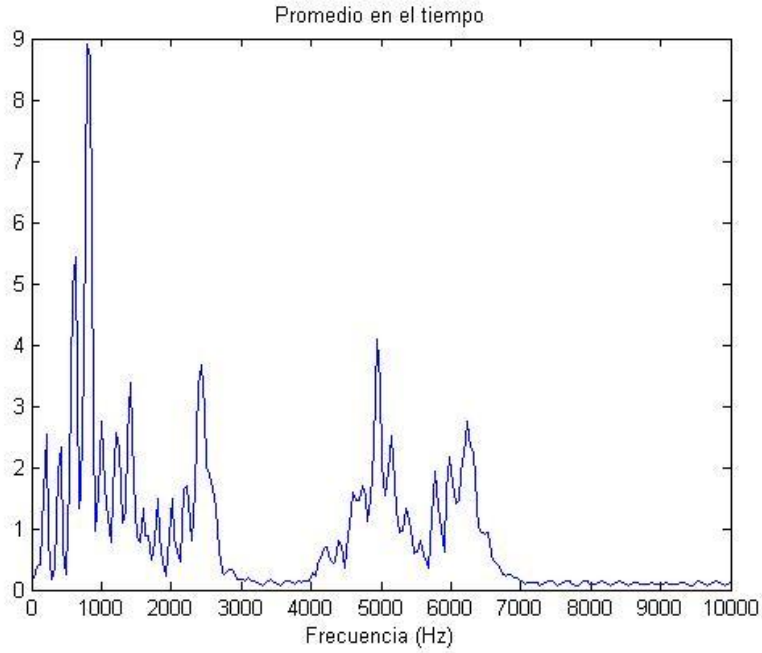

Figura 3.- Patrón de la letra 'a'

En este proyecto se decidió que los vocablos que el sistema reconocerá, son: ***agua***, ***leche***, ***café*** y ***jugo***. Los patrones de reconocimiento responden a las fortalezas del

sistema —identificar sonidos vocales—, por lo que se forman seis patrones base, mismos que tienen las cinco vocales y el fonema /sh/, que caracteriza a la pronunciación de la palabra ***leche***.

Los patrones son:

***Agua=** patrón ‘a’

***Café=** patrón ‘a’ + patrón ‘e’

***Leche=** patrón ’e’ + patrón ‘sh’

***Jugo=** patrón 'u' + patrón 'o'

Con el fin de calibrar el sistema de reconocimiento de voz, se le solicitará al usuario pronunciar cualquiera de las cuatro palabras —agua, café, leche, jugo—, para así obtener ciertos patrones que el ordenador correlacionará e identificará como el vocablo emitido.

Resultados

Con una vez que se calibre, el reconocer de voz podrá ser usado por distintos usuarios, sin ser relevante el timbre, tono y rapidez con la que pronuncien las palabras, aunque, tiene como limitante reconocer las expresiones cuando éstas son dichas a bajo volumen.

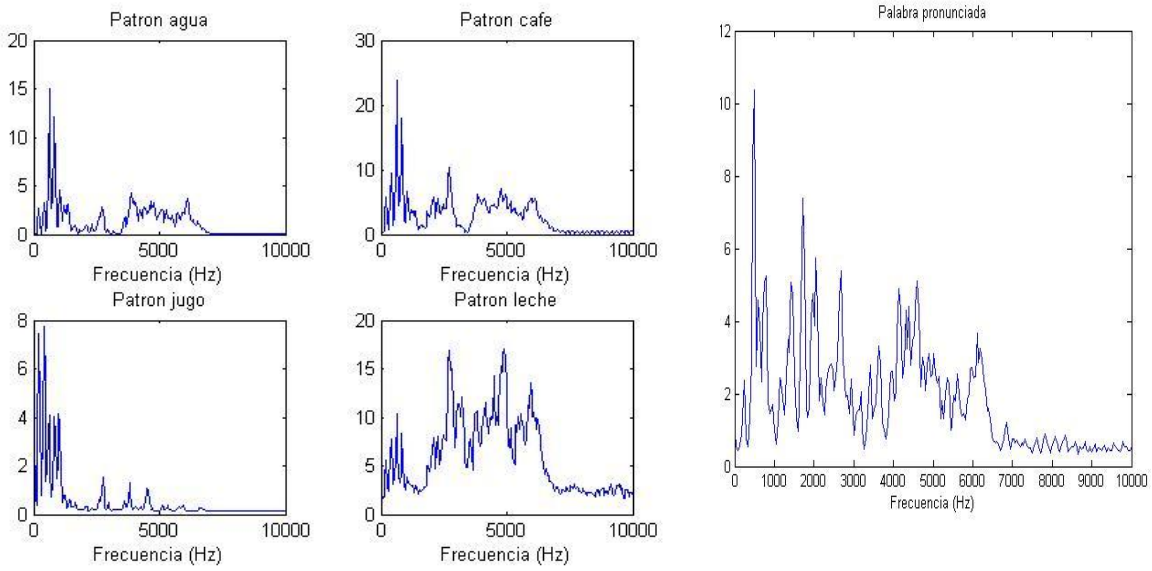

Figura 4.- Patrones comparados contra la palabra pronunciada

La figura 4 expone los patrones utilizados para reconocer la palabra pronunciada, la cual en este caso fue: ***café***. En ella se puede observar que el patrón de ***café*** y ***agua*** son parecidos, sólo los diferencia una señal entre 1000 y 3000 Hz —el aporte de la vocal "e"—; esto mismo se puede apreciar en comparación con la palabra pronunciada, que presenta más atenuados los componentes de 0 a 1 HKz.

El coeficiente de correlación entre el patrón que resulta ser la palabra y el vocablo pronunciado, fue de 0.6283. Y si bien, en la señal del vocablo pronunciado no se tomaron en cuenta todos los sonidos presentes en ***café —****como* /k/ y /f/—, puesto que las vocales son dominantes es posible identificar la palabra sólo utilizando la suma de patrones 'a' + 'e' .

Finalmente, se ha de tomar en cuenta que si se agregaran más palabras al reconocedor, las cuales contuvieran las mismas vocales que las palabras antes mencionadas: disminuiría la exactitud del sistema, aumentando el margen de error. Si se quisiera evitar, se tendría que identificar la aportación de las consonantes a la señal patrón en cada palabra.

Referencias

1.- "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot", Thiang, Suryo Wijoyo, Electrical

Engineering Department, Petra Christian University, Jalan Siwalankerto 121-131, Surabaya 60236, Indonesia.

2.- "How to buil a computer controlled robot", Tod Loofbourrouw, HAYDEN, 1978.

3.- Y.M. Lam, M.W. Mak, and P.H.W. Leong. Fixed point implementations of Speech Recognition Systems. *Proceedings of the International Signal Processing Conference*. Dallas. 2003

4.- "Digital Signal Processing", John G. Proakis, Diminitris G. Manolakis, , Prentice Hall, New Jersey 1996.