ISSN: 1692-7257 Volumen 2 Número 6 año - 2005

Revista Colombiana de
Tecnologías de Avanzada

AUTOMATIC SPEECH RECOGNITION USING FOURIER TRANSFORM AND NEURAL NETWORK

RECONOCIMIENTO AUTOMÁTICO DEL HABLA UTILIZANDO LA TRANSFORMADA DE FOURIER Y REDES NEURONALES.

Ph.D. Cesar Torres Moreno, Mg. Lorenzo Mattos, Ing. Gilberto Perpiñan Iseda, Ing. José Ángel Castro, Ing. José David Pardo

Universidad Popular del Cesar,
Laboratorio de Óptica Sede Balneario Hurtado,
torres.cesar@caramail.comgperpignan29@yahoo.com,jac_castro@walla.com,
parduz@hotmail.com
Valledupar

Abstract: In the present article it is do the automatic recognition of the signal of speech using for this the Fourier's transform and the Neural. Networks. The characteristics of the speech will be discriminated in the domain of the frequency and we will proceed to train a Neural. Networks for to classify the different patterns and to use them as commands to realize functions of control.

Resumen: En el presente artículo se realiza el reconocimiento automático de las señales de voz utilizando la transformada de Fourier y las redes neuronales. Se discriminará las características del habla en el dominio de la frecuencia y se procede a entrenar una red neuronal para clasificar los diversos patrones y utilizarlos como comandos para realizar funciones de control, como encender y apagar las luces.

Keywords: Automatic Speech Recognition, Fourier Transform, Neural Network.

1. INTRODUCCIÓN

Los sistemas de reconocimiento automático de voz tienen gran auge en la actualidad dada la amplia comodidad y versatilidad que brindan. Entre sus virtudes está el ofrecer la posibilidad de que el hombre interactúe en su lenguaje natural con equipos electrónicos y eléctricos, además es un proceso casi intuitivo el operar estos equipos por la sencillez de los comandos que se utilizan, que por lo general son palabras cortas como "encender", "apagar", entre otras instrucciones que describen explícitamente su función. Esto permite realizar actividades como encender el televisor o las luces de una habitación con solo pronunciar una orden, sin importarnos si estamos ocupados. El proyecto que se propone es vanguardista a nivel regional, porque en la costa atlántica el trabajo en esta área es poco.

El reconocimiento automático de la voz dota a las máquinas de la capacidad de recibir mensajes orales. Tomando como entrada la señal acústica

recogida por un micrófono, este proceso de reconocimiento del habla tiene como objetivo final decodificar el mensaje contenido en la onda acústica para realizar las acciones pertinentes.

2. PROCEDIMIENTO

2.1. Componentes de un sistema de Reconocimiento Automático de Voz (RAV)

El Estos sistemas se pueden describir como la concatenación de cuatro subsistemas Ver Fig. 2: La adquisición de la entrada, la representación de la misma, una ordenación local y un decodificador general. En el primer subsistema se muestra la relevancia del micrófono como un transductor que convierte las ondas sonoras en señales eléctricas, introduciendo al mismo tiempo ruido, que sumado al generado por el salón acústico donde se instala el sistema y a la posición del micrófono afectan el espectro de la señal translúcida.

La extracción de las características es la forma como se representa la señal de voz mediante modelos robustos de la variación acústica, con lo que se busca que cada señal pueda ser representada inequívocamente como un conjunto de valores que la distinga de las demás. La correcta diferenciación de las señales depende de las técnicas de clasificación que se utilicen. El subsistema de Ordenación o igualación local, establece las distancias que existen entre los patrones de referencias que previamente se guardaron en diccionarios y los patrones debidos a la señal en cuestión. Por último tenemos un decodificador general, que toma el patrón más parecido al de la señal en estudio y le da un significado que viene a ser la palabra reconocida (MACMILLAN, 1993).

2.2. Adquisición de señales de voz

El procedimiento desarrollado involucra la captura de la señal con la ayuda de un micrófono y posteriormente es digitalizada por medio de la tarjeta de sonido del computador; en la figura 1 se muestra un diagrama de bloques de este proceso, y es expresado matemáticamente de la siguiente forma: x(t), que es la forma de onda original, es decir la señal de voz, y la versión que entrega el micrófono es $y(t) = x(t) * \tilde{h}(t)$, luego muestreando con un intervalo T, a una frecuencia que es igual a dos veces la componente frecuencial de la señal de mayor orden de acuerdo con el teorema de Nyquist, fc=2f, para el caso de las señales de voz la frecuencia de muestreo es igual a 8000 muestras por segundo, la salida y(nT) es dada por,

$$y(nT) = \int_{-\infty}^{\infty} h(t - nT)x(t)dt \qquad (1)$$

Donde, $\tilde{h}(t)$ es la respuesta impulsional del micrófono; y(t) es la salida del micrófono; T es el intervalo de muestreo; $y_s(t)$ es la versión muestreada de y(t); y(nT) es la salida, $n \in Z$ y son los valores de las muestras. El bloque del C/D es un conversor análogo a digital. La adquisición de las señales se realiza utilizando el DAQ (data acquisition) de Matlab, ver Fig. 3.

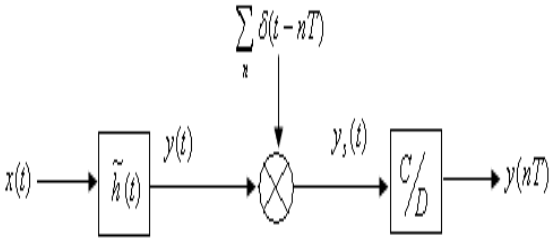

Fig. 1. Muestreo de una señal: x(t) es la señal de voz continua en el tiempo.

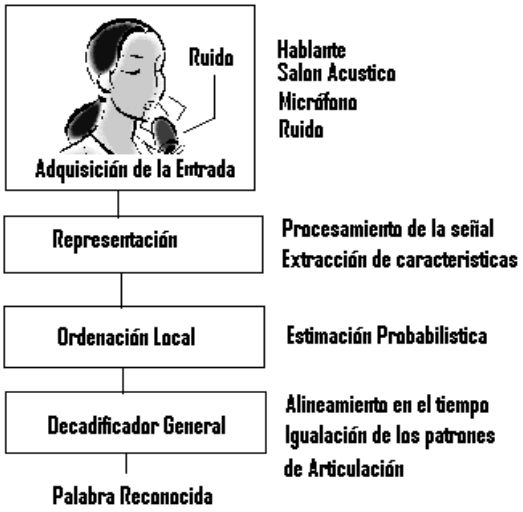

Fig. 2. Diagrama de bloques de un sistema de reconocimiento de voz.

En esta etapa se tiene en cuenta que la amplitud de la señal no sobre pase los límites de la tarjeta de sonido, que sólo visualiza señales de 1 Voltio (V),

si el locutor al hablar supera los umbrales, la señal no se procesa y se pide una nueva señal, de igual manera si se tienen señales inferiores a | 150 mV| se toma como ruido, pues se comprobó que cuando no se está hablando se visualizan señales entre 150 mV, y no se procesa.

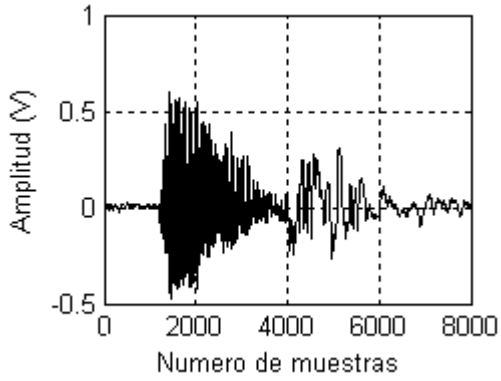

Fig. 3. Adquisición de la señal de voz, utilizando el Matlab.

2.3. Tratamiento de las señales

Las herramientas más utilizadas para el análisis y tratamiento de las señales de voz son los bancos de filtros y la transformada de Fourier, entre otras(Gold, B.2000).

Los filtros se utilizan par separar unas señales de otras, el primer filtro que se utilizó fue el de traslado de promedios (moving average) es el filtro mas común en procesamiento digital de señales (DSP) en el dominio del tiempo (Smith, 1999), utilizado para reducir el ruido aleatorio mientras impide los cambios abruptos de la respuesta impulsional. Este filtro se describe matemáticamente en la ecuación (2)

$$y[i] = \frac{1}{M} = \sum_{j=0}^{M>1} x[i+j] \qquad (2)$$

Donde x[] es la señal de entrada, y[i] es la señal de salida y M es el número de puntos que se van a promediar de la señal de entrada. Por ejemplo si en un filtro de traslado de promedios M = 4 el punto 50 de la señal será el resultado de:

$$y[50] = \frac{x[50]+x[51]+x[52]+x[53]}{4} \qquad (3)$$

En este trabajo se utilizó un filtro de traslado de promedio con un M = 2, el cual se le aplicó a la señal que se ubicó de tal manera que la información relevante iniciara en las primeras muestras. En la Fig. 3. se observa que las muestras desde 0-1500 sólo es ruido. Después del filtrado se normalizó la señal, de modo que a cada muestra característica de la señal corresponda una potencia similar. Esto se muestra en la Fig. 4.

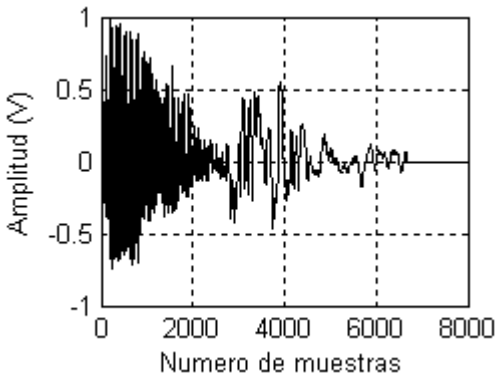

Fig. 4. Señal de voz filtrada y normalizada.

La transformada discreta de Fourier unidimensional es ampliamente utilizada en el estudio de las señales, sin embargo debido a su gran costo computacional hasta hace algunas décadas se ha implementado masivamente, cuando Cooley y Tukey (Cooley, J 1965) introdujeron la transformada rápida de Fourier (FFT). La FFT es una técnica basada en la descomposición de las señales en sinusoidales, a demás la FFT es una versión eficiente de la DTF (transformada discreta de Fourier). La DTF de una secuencia de duración finita

$x(n) \le n \le N-1$ está definida por:

$$X(k) = \sum_{n=0}^{N-1} x(n) W^{nk} \qquad (4)$$

Donde $W = e^{-j(2\pi/N)}$, la DTF calcula el espectro de una secuencia finita.

Haciendo uso de esta herramienta (FFT), se extraen características particulares de cada señal de voz, es decir se discriminan las frecuencias que poseen dichas señales. Utilizando el algoritmo de la FFT que trae Matlab se visualiza la señal mostrada en la Fig. 5.

El espectro de la señal de voz se normaliza y

posteriormente se realiza una ponderación de las muestras de modo que los datos más característicos estén incluidos en un vector de tamaño reducido, para disminuir el costo computacional y aumentar la velocidad de procesamiento. Esto se muestra en la Fig. 6.

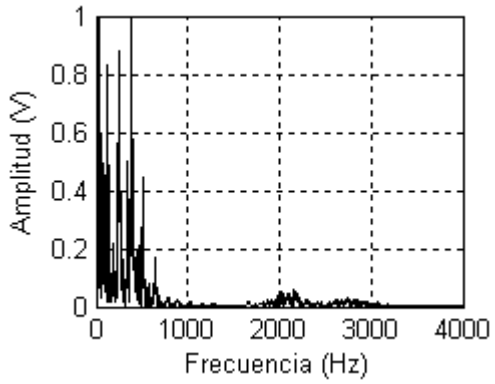

Fig. 5: Espectro de la señal.

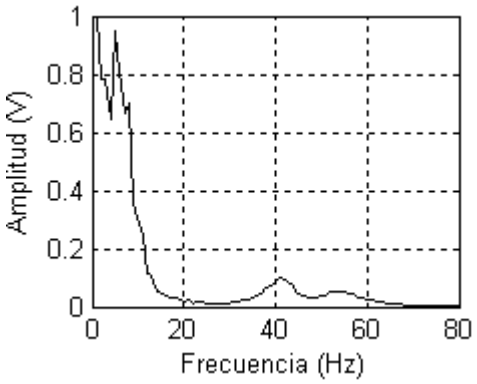

Fig. 6: Espectro normalizado y promediado.

2.4. Implementación de redes neuronales

Al aplicar las técnicas de análisis de señales unidimensionales discretas en el tiempo, se extraen las características predominantes en cada una de las palabras que formaran la base de datos del reconocedor automático del habla.

Las redes neuronales son los algoritmos encargados de realizar la ordenación local o decir si la señal de entrada es similar a las señales que hacen parte de la base de datos con que se entrena la red neuronal. (DELGADO 1998)

Para entrenar la red neuronal se tomaron 2400 muestras de 30 palabras de un mismo locutor, 10 palabras 'tele', 10 'canal' y 10 'volumen'. La red que se entrenó fue una perceptrón multicapa con retropropagación de error. Se utilizó el toolbox de redes neuronales de matlab

Para una red de tres capas, la capa de entrada con 80 neuronas cada una para procesar las 80 muestras del vector de entrada que corresponde al espectro normalizado y promediado de la señal en cuestión, 5 neuronas en su capa oculta y 3 neuronas en la capa de salida, donde cada neurona de la capa de salida corresponde a una palabra del vocabulario (tele, canal, volumen).

3. RESULTADOS

	Tele % O.K	Canal % O.K	Volumen % O.K
Tele	91.3	3.5	4.1
Canal	3.4	92.5	5.3
Volumen	6.4	2.8	90.2

Con la red neuronal entrenada satisfactoriamente, se concatenó cada una de las etapas; adquisición de la señal, procesamiento de la misma y extracción de características, resultando un reconocedor de tres palabras (tele, canal y volumen).

4. CONCLUSIONES

La transformada discreta unidimensional de Fourier brinda una herramienta útil para la extracción de características de las señales de voz. Para que esa extracción mejore se debe realizar un pre-procesamiento.

Las redes neuronales proveen una excelente ayuda en la diferenciación de las señales, su implementación en Matlab se hace de manera sencilla, sin embargo se debe estudiar otras topologías para optimizar los resultados obtenidos.

REFERENCIAS

MACMILLAN, 1993. Signal Processing of Speech. F. J. Owens. - Ed.

Vetterli M, June 2002 "Sampling signals with finite rateo f innovation" IEEE transactions on signal processing, Vol. 50, NO. 6,

Gold, B.2000, Morgan N, "Speech and Audio Signal Processing ". - Ed. JOHN WILEY & SON,. Cap. 4

Smith, 1999 W. Steven, the Scientist and Engineer's Guide

Digital Signal Processing Second Edition. California Technical Publishing. San Diego, California,. Cap 15

Cooley, J 1965. W., y Tukey, J., "An algorithm for machine computation of complex Fourier series", Math comput. 19: 297-301,

DELGADO 1998, Alberto. Inteligencia Artificial y Mini robots. ECOE Ediciones.