Conciencia Tecnológica
ISSN: 1405-5597
contec@mail.ita.mx
Instituto Tecnológico de Aguascalientes
México

Esparza Arellano, María Elena; Avalos Briseño, J. Benito
Reconocimiento de voz
Conciencia Tecnológica, núm. 22, 2003
Instituto Tecnológico de Aguascalientes
Aguascalientes, México

RECONOCIMIENTO DE VOZ.

Instituto Tecnológico de Aguascalientes
Av. Adolfo Lopez Mateos 1801 Ote.
María Elena Esparza Arellano. Depto. de Ciencias Básicas
J. Benito Avalos Briseño. Depto. de Eléctrica y Electrónica
Tel: 01(449)9-105-002 Ext. 143/106 E-mail: jbenitomx@yahoo.com.mx

INTRODUCCION

Una interfase de lenguaje hablado a la computadora es un tema que ha atraído y fascinado a ingenieros científicos del lenguaje.

Avances en la tecnología del lenguaje humano son necesarios para el común de los ciudadanos, quienes para comunicarse con redes, usando las habilidades naturales, ocupan dispositivos de uso diario, como son el teléfono y la televisión. Sin que exista un fundamental avance en las interfaces centradas en el usuario (Computadora), gran parte de la sociedad se preparará para participar en la era de la información, mas sin embargo, otros no lo harán obteniéndose con esto una completa estratificación de la sociedad, resultando en la trágica pérdida del potencial humano.

En el presente trabajo se visualiza un panorama general, no así menos importante, del presente y futuro del Reconocimiento de Voz, para su uso en Aplicaciones Industriales y de Servicios , así como un pequeño Análisis Matemático básico, como soporte ***introductorio*** hacia la fundamentación matemática avanzada.

En la ***figura 1***, se muestra un diagrama esquemático de las etapas básicas de un sistema de este tipo considerando, fundamentalmente :

Entradas de señales analógicas, etapa de comparación señal a ruido (S/N), sección de filtrado, conversión análogo – digital y salida digital para etapas de proceso posteriores.

Dentro del mismo trabajo se muestra la fortaleza del sistema de muestreo y retención de la señal analógica bajo análisis, considerando el enfoque desde un sistema de adquisición, control y proceso de datos , con las etapas posteriores.

La relación entre los principios básicos de relación y soporte entre redes neuronales y lo básico de inteligencia artificial permite que el lector de este artículo se forme un a idea básica e interesante de los temas mencionados para que este artículo cumpla con el objetivo deseado que es el de la ***Divulgación Científica*** de nuestro quehacer

ANTECEDENTES HISTORICOS

En 1952 Davis, Bidulph y Balashek, de los laboratorios Bell fabricaron el primer reconocedor capaz de discriminar con cierta precisión los diez dígitos ingleses pronunciados de forma aislada por un único lector. El dispositivo era totalmente electrónico. Los primeros trabajos que hacen uso de tecnología informática, comienzan a aparecer en 1959/1960; Deves y Mathews introducen el concepto de normalización temporal no lineal, que permite la comparación de parámetros de palabras iguales pronunciadas a distinta velocidad.

A partir de estas fechas comienza la explosión de trabajos, principalmente de reconocimiento de palabras aisladas, con la

extrapolación optimista, por parte de investigadores y organismos financiadores, de llegar, en poco tiempo, a sistemas capaces de reconocer de forma precisa frases cualesquiera, pronunciadas por un lector cualquiera, de forma continua.

Con este objeto más o menos en mente, se lanzan grandes proyectos de investigación en los que se pretende llegar a las menores restricciones gramaticales posibles de las frases a reconocer, así como del léxico utilizado. Son varios los países en los que se comienza a trabajar en proyectos de ésta índole (Japón, Francia, etc.), pero es en EE.UU. donde se lanza, en 1971, el mayor proyecto conocido en la historia del reconocimiento del habla. Se trata del <<ARPA-SUR>> (Advanced Research Projects Agency – Speech Understanding Research), con un presupuesto de quince millones de dólares y una duración de cinco años.

Aunque los ambiciosos objetivos pretendidos en éste y otros proyectos no llegaron realmente a alcanzarse. Las aportaciones derivadas de ellos, contribuyeron de forma notable a un mejor conocimiento de los mecanismos del habla y de las limitaciones de los sistemas automáticos de reconocimiento.

APLICACIÓN TIPICA

El Software DragonDictate. Este producto está disponible para Windows y permite al usuario interactuar con muchas aplicaciones diferentes en su PC. Permite la entrada de datos en Excel sin la utilización de las manos, un sistema de dictado en Word, así como muchos otros programas de aplicación. Otro uso que está siendo desarrollado es para el sistema de reservaciones de las aerolíneas. El usuario será capaz de marcar un número y contestar las siguientes preguntas:

C: Este es el sistema de información de vuelos. ¿En qué puedo ayudarle?.
U: Me gustaría hacer una reservación.

C: Por favor especifique su plan de vuelo.
U: Quisiera ir de New York a Chicago el sábado por la mañana.

DEFINICIONES DEL PROBLEMA

Algunas definiciones del problema que existe en el reconocimiento del habla son las siguientes:

1° Definición. Hacer cooperar un conjunto de informaciones plagadas de ambigüedades, incertidumbres y errores inevitables, para llegar a una interpretación aceptable del mensaje acústico recibido.

2° Definición. Encontrar la mejor estrategia en el reconocimiento de formas que posee la señal vocal procedente de algún locutor humano y el algoritmo capaz de identificar qué formas específicas componen determinado fonema.

VARIABLES DEL PROBLEMA

Los obstáculos con los que se lucha en el reconocimiento de voz se describen a continuación:

- *Bidireccionalidad.* La comunicación oral, comporta generalmente un intercambio

bidireccional de información entre dos locutores – auditores o más.

- *Incomplenitud.* La información intercambiada es siempre mayor que la estrictamente contenida en el mensaje oral (gestos, énfasis, contexto, etc.).

- *Multiinteractividad.* Existen varios niveles de comprensión, que interaccionan dinámicamente entre sí y en combinación con otros sistemas perceptivos y motores. Cada uno de estos niveles aplica la fuente de conocimiento sobre el lenguaje que le es propia y extrae su parte correspondiente de la información total necesaria para la comprensión del mensaje.

- *Continuidad.* A pesar de que se tenga la impresión contraria, ni los fonemas ni las sílabas, ni las palabras se pueden separar fácilmente de forma automática.

- *Variabilidad.* Es imposible que un locutor pronuncie dos veces exactamente igual una misma sílaba, palabra o frase.

- *Transitoriedad.* Sólo las variaciones de una señal permiten transmitir información. El tipo de parámetros que diferencian las transiciones no es aún suficientemente conocido.

- *Incertidumbre e inexactitud.* Tanto la propia señal como las fuentes de conocimiento asociadas a los distintos niveles de percepción, constituyen informaciones <<ruidosas>>, en el doble sentido de que, en general, son incompletas y con <<artefactos>> superpuestos.

TÉCNICAS USADAS

Por medio de las técnicas actuales de reconocimiento de formas de señales acústicas como la FFT (" Fast Fourier Transform ") y a través de los métodos que de I.A. (Inteligencia Artificial) hasta la fecha conocidas, podemos resolver en parte la mayoría de las variables del problema que en el Reconocimiento del habla existen.

En el habla, el universo físico de los objetos a reconocer está constituido por las ondas de presión producidas por el aparato fonador humano. Los objetos externos de este universo los constituyen las diferentes formas acústicas del habla.

La parte inicial de todo subsistema de preproceso de la señal vocal estará siempre constituida por:

- *Un micrófono*, que convertirá la onda sonora de presión en una señal eléctrica.

- *Un amplificador*, que extenderá hasta nivel manejable la débil señal que proporciona el micrófono.

- *Un filtro activo pasa bajas,* que eliminará la altas frecuencias indispensables según el teorema de muestreo de Nyquist.

MUESTREO Y CUANTIFICACIÓN

A partir de la señal eléctrica que produce el amplificador sería teóricamente posible construir un sistema de reconocimiento por medios totalmente analógicos. Sin embargo, en el estado actual de la tecnología, resulta más conveniente utilizar técnicas digitales; sobre todo para las partes del sistema involucradas en la decisión.

Básicamente un convertidor A/D debe realizar dos tareas:

- Muestrear la señal analógica; es decir, medir la amplitud de dicha señal, cierto intervalo de tiempo.

- Cuantificar la señal muestreada; es decir, codificar numéricamente el resultado de cada una de las medidas.

De esta manera, una función continua en el tiempo quedará representada por una serie discreta de valores numéricos. Al proceso combinado de transducción, muestreo y cuantificación se le llama adquisición.

.a. ***.b.*** ***.c.*** ***.d.*** ***.e.*** ***.f.***

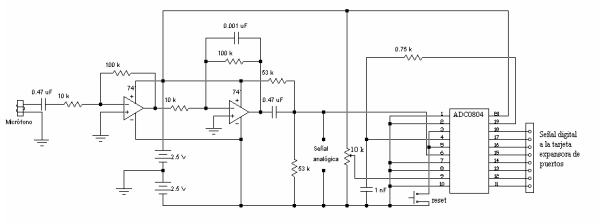

Figura 1.- Circuito prototipo
a) Entrada analógica, b)comparación, c) filtrado, d) entrada analógica, e) convertidor ADC0804, f) salida digital.

MUESTREO

Una señal muestreada a intervalos de tiempo T, $S_M(t)$, puede definirse como el producto de la señal continua s(t) y una función <<peine>> de δ Dirac (función impulso)

$$S_M(t) = S(t) \sum_{.k = -\infty}^{\infty} \delta\,(t - kT)$$

$$S_M(t) = \sum_{.k = -\infty}^{\infty} S(kT)\,\delta\,(t - kT)$$

De donde es inmediato demostrar que la expresión del espectro $S_M(j\omega)$ de la señal muestreada en función de la señal sin muestrear $S(j\omega)$, adopta la forma:

$$S_M(j\omega) = 1/T \sum_{.k = -\infty}^{\infty} S[j(\omega + k\omega_o)];$$

$$\omega_o = 2/T$$

Expresión que representa la superposición del espectro de **s(t)** con las sucesivas versiones del mismo desplazadas en el eje de las frecuencias con periodicidad **1/T.**

Resulta evidente que si el ancho de banda de la señal a muestrear es excesivo con relación a la frecuencia de muestreo, se producirá un solapamiento irreversible de los espectros sucesivos, haciendo imposible la reconstrucción de la señal original. Este solapamiento (aliasing) ocurre siempre que la máxima frecuencia *(Fb)* del espectro no nulo de la señal a muestrear sea superior a la mitad de la frecuencia de muestreo *(Fm)* (frecuencia Nyquist).

Fm > 2Fb

Antes de muestrear una señal será pues necesario limitar la frecuencia máxima de ésta a la mitad de la de muestreo, lo que se puede conseguir mediante un filtro analógico de paso bajo previo al convertidor A/D, cuya frecuencia de corte sea la de Nyquist como máximo.

La anchura de banda de la señal resultante deberá preservar la información relevante necesaria

para una adecuada descripción de los objetos acústicos a tratar.

CUANTIFICACIÓN

En cada impulso de muestreo, el convertidor A/D compara la señal muestreado con cosa dado un conjunto de entradas, podemos usar unidades sumadoras con nivel de disparo (Threshold) como simples compuertas AND, OR, y NOT poniendo apropiadamente el nivel de disparo y los pesos de conexión entre ellas una serie de nivel de cuantificación predefinidos. El número de niveles (N) determina la precisión del análisis y, por tanto, el número de bits (b) necesarios para la presentación digital de cada muestra:

$$b = \log_2 N$$

Teniendo en cuenta la relación de Nyquist se puede determinar el flujo de información, en bits por segundo, resultante del proceso combinado de muestreo y cuantificación:

$$\Phi > 2Fb \cdot \log_2 N$$

Para señales vocales adquiridas directamente en el dominio del tiempo, dicho flujo suele oscilar entre 50 Kbits/s y 300 Kbits/s.

En el prototipo que se muestra en la ***figura 1*** el período de muestreo T es fijado por un retardo a través del programa y la cuantificación que inicialmente puede consistir en 255 valores (8 bits) por necesidades del algoritmo backpropagation son reducidos a un valor numérico entre 0.1 y 0.9 multiplicando el valor máximo (255) por un factor adecuado.

ANÁLISIS EN EL DOMINIO DEL TIEMPO Y LA FRECUENCIA

En el análisis en el dominio del tiempo tenemos el método de *energía y amplitud media* y el método de *densidad de cruce por cero.* En el análisis frecuencial la *transformada de fourier* y la *predicción lineal* ***(LPC)***. Para este proyecto por simplicidad se emplea el método de densidad de cruce por cero.

DENSIDAD DE CRUCE POR CERO

La densidad de cruces por cero ha sido objeto de numerosos estudios teórico y práctico. Su utilidad en reconocimiento del habla radica en que proporciona una estimación aproximada del contenido frecuencial de una señal, basada en la idea de que una sinusoide pura cruza el eje de abscisas 2 veces por período. Es un parámetro de muy baja complejidad de cálculo, y se le ha utilizado para detectar segmentos fricativos (señal de pequeña energía y elevada densidad de cruces por cero).

Después de obtenidos los parámetros, estos se alimentan a la red neuronal de aprendizaje (Técnica de Inteligencia Artificial) que distinguirá entre uno y otro fonema.
Redes Neuronales (Backpropagation).

¿Qué es lo que una red multicapas puede computar? La respuesta es: cualquier.

El mayor problema es el aprendizaje. La representación del conocimiento en las redes Neuronales es un poquito opaco: Las redes deben aprender su propia representación debido que programarlas a mano es imposible.

Las redes Neuronales pueden aprender cualquier cosa que ellas puedan computar.

Primero trataremos con una subclase de las redes Neuronales llamada redes en capas completamente conectadas de propagación hacia delante; una de la cual se muestra a continuación (fig. 2)

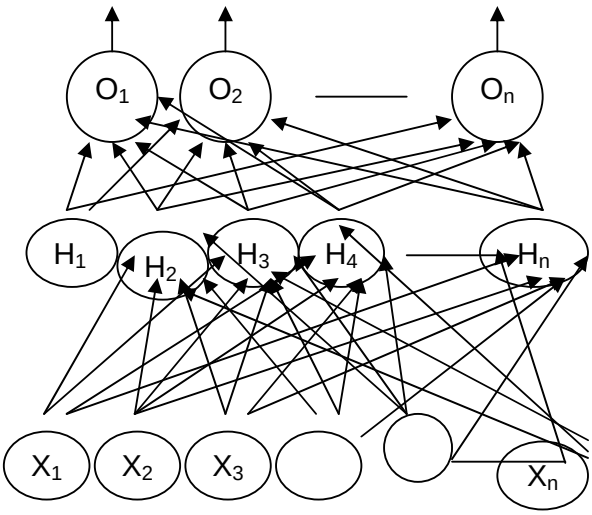

Fig. 2.- Ejemplo de Red Neuronal en Capas

En la figura 2, Xi, Hi y Oi representan las unidades de niveles de activación de las unidades de entrada ocultas y de salida. Los pesos en conexión entre las unidades de entrada y las ocultas están denotadas aquí por las relaciones que existen entre los niveles de las O´s y los niveles de las H´s, al mismo tiempo los pesos entre las unidades ocultas con las capas de salida están denotadas por las relaciones que existen entre los niveles de las H´s y los niveles de las X´s.

Estas redes tienen tres capas, aunque esto es posible y algunas veces es útil tener más. Cada unidad en una capa está conectada en la dirección hacia delante con cada unidad en la capa próxima. La activación fluye desde la unidad de entrada hacia la capa oculta, entonces pasa a la capa de salida. Como es usual, el conocimiento de la red se codifica en los pesos de conexión entre las unidades. En contraste con el método paralelo de relajación usado por las redes de Hopfield. Las redes de propagación hacia atrás lleva a cabo una simple serie de cálculos. Debido a que la activación fluye solo en una dirección, no hay necesidad de un proceso iterativo de relajación.

Los niveles de las unidades de la capa de salida determinan la salida de la red.

La esperanza al atacar problemas como el reconocimiento de la escritura a mano es que las redes Neuronales no solamente aprenderán a clasificar las entradas con que fueron entrenadas sino que generalizarán y serán capaces de clasificar entradas que aún no han sido vistas.

CONCLUSIONES

Las técnicas usadas en el reconocimiento de patrones de voz como FFT (Fast Fourier Transform) son mucho mejor que las de cruce por cero. esta técnica usada junto con la de redes Neuronales logran resultados óptimos. En un futuro próximo se espera tener reconocedores del lenguaje humano que no sean sólo para un locutor definido.

BIBLIOGRAFÍA.

Landee, Robert W., Davis, Donovan C. & Albrecht, Albert P. Electronics designers´handbook, second edition, McGraw Hill.

Close, Charles M. & Frederick, Dean K. Modeling and Analysis of Dynamic Systems. Rensselaer Polytechnic Institute. Houghton Mifflin Company.

Bowker, Albert H. & Lieberman, Gerald J., Engineering Statistics, 2nd edition, Prentice Hall Inc.