PROCESAMIENTO DE SEÑALES DE VOZ CON REDES NEURONALES

Leonardo Romero Muñoz
(Instituto de Investigaciones Eléctricas - Cuernavaca, Morelos)
Ricardo Fernández del Busto
(Instituto Tecnológico y de Estudios Superiores de Monterrey - Campus Morelos - Cuernavaca, Morelos)
Ofelia Cervantes de Poty
(Universidad de las Américas - Puebla, Puebla)

RESUMEN

Se desarrolló una herramienta para evaluar la aplicación de un tipo de red neuronal, llamada perceptrón multicapa, al reconocimiento de palabras aisladas de voz (entre una palabra y otra se hace una pausa). La herramienta consiste en un conjunto de programas que permiten implementar y analizar tres sistemas de reconocimiento de voz. Los primeros resultados de reconocimiento muestran que el perceptrón multicapa, combinado con técnicas tradicionales, tiene un gran potencial en el desarrollo de sistemas de reconocimiento de palabras aisladas. (Romero 1990).

INTRODUCCION

Desde hace un par de años a la fecha (Mariani, 1989), las redes neuronales

(una forma de computación paralela de un gran número de procesadores elementales interconectadas) se empezaron a aplicar al reconocimiento de voz, y, actualmente, los resultados son prometedores. Normalmente, los sistemas de reconocimiento de palabras aisladas tienen dos componentes principales: extracción de características y clasificación (ver figura I). El objetivo de la extracción de características es transformar la entrada de voz, en un conjunto pequeño de características (patrón de características) que representen la señal original. La clasificación se puede interpretar como una partición del espacio de características en regiones mutuamente excluyentes, tales que cada región esté asociada a una palabra (Sotelo, 1989).

CLASIFICACION

La clasificación se lleva a cabo en dos etapas: aprendizaje y reconocimiento. En la etapa de aprendizaje, se forma un diccionario de patrones de palabras. En la etapa de reconocimiento, simplemente se calculan los "grados de similitud" del patrón de prueba, con cada uno de los patrones del diccionario y se escoge aquél que esté más cercano (mediante la regla de decisión). La figura 2 muestra los conceptos mencionados.En la clasificación basada en una red neuronal el diccionario de patrones (de la figura 2) está implícito en la red y los cálculos de "grados de similitud" se realizan en paralelo. Para ello, la red neuronal tiene tantas salidas como palabras del vocabulario a reconocer.

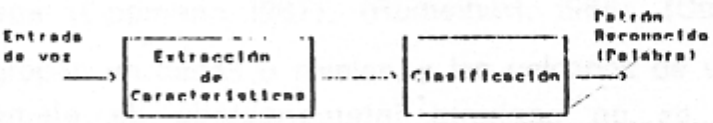

FIGURA 1.- DIAGRAMA DE BLOQUES DE UN SISTEMA DE RPA.

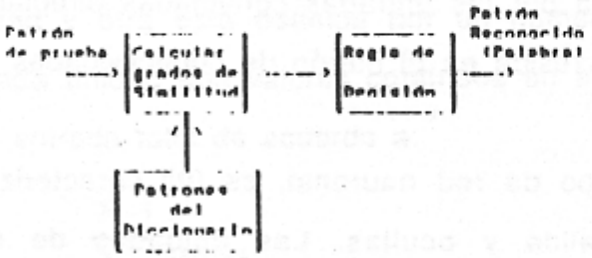

FIGURA 2.- DIAGRAMA DE BLOQUES DEL CLASIFICADOR DE PATRONES.

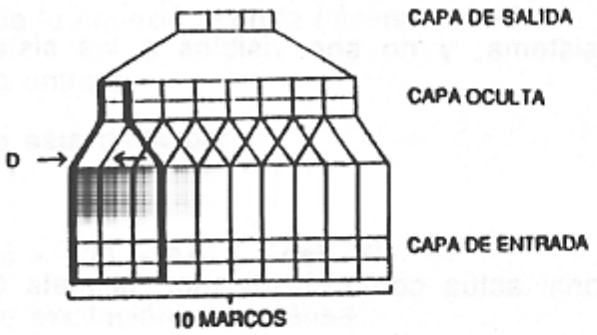

FIGURA 3.- UNA RED NEURONAL DE TIEMPO DIFERIDO.

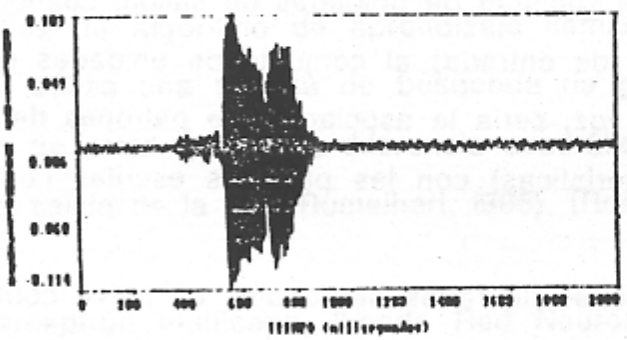

FIGURA 4.- SEÑAL ADQUIRIDA DE LA PALABRA /CERO/ NORMALIZADA AL INTERVALO [-1,1].

REDES NEURONALES

Una red neuronal es un conjunto interconectado de elementos que llamaremos unidades o neuronas, las cuales procesan y almacenan información en forma local; es decir, cada neurona procesa sólo la información que le llega por las entradas conectadas directamente a ella. El conocimiento de la red reside en el patrón de conectividades de la red.

Dentro de cualquier tipo de red neuronal, es útil caracterizar tres tipos de unidades: entrada, salida y ocultas. Las unidades de entrada reciben entradas de fuentes externas a la red (v.gr. las características del patrón de una palabra). Las unidades de salida envían señales fuera de la red (v.gr. los grados de similitud). Las unidades ocultas son aquellas que tienen entradas y salidas dentro del sistema, y no son visibles a los sistemas exteriores. (Rumelhart, 1986).

Cuando una red neuronal actúa como clasificador, la meta es establecer las conexiones necesarias (entre unidades) para que se produzca un patrón de salida deseado en el conjunto de unidades de salida, cuando se alimenta un determinado patrón (de entrada) al conjunto de unidades de entrada de la red. En el caso de voz, sería la asociación de patrones de palabras de voz (patrones de características) con las palabras escritas correspondientes.

Una de las aplicaciones de redes neuronales es servir como clasificador de patrones (Lippmann, 1987) y para el caso de reconocimiento de voz un tipo de red llamada perceptrón multicapa, es particularmente atractivo. En el

perceptrón multicapa (Lippmann 1987), (Rumelhart, 1986), (Caudill, 1987), las unidades se agrupan en capas o niveles, y las unidades de un nivel sólo tienen conexiones (unidireccionales) con el nivel inmediato superior. Las unidades inferiores son las de entrada, las superiores son las de salida y las intermedias son las ocultas. Todas las unidades son del mismo tipo y la conexión entre un unidad y otra está definida por un número real llamado peso. Las salidas de cada unidad son valores contínuos en el intervalo [0,1]. Cada unidad calcula su entrada total de acuerdo a:

$$net = \sum_{i=0}^{N-1} Wi\ Xi$$

Donde N es el número de entradas

Wi es el peso de la conexión con la i-ésima entrada

Xi es la i-ésima entrada

y la salida de la unidad está dada por:

$$s = 1/(1 + \exp (- net - \theta)$$

Donde θ es un valor de excitación espontánea.

Este tipo de red utiliza un algoritmo de aprendizaje llamado regla delta generalizada el cual utiliza una técnica de búsqueda de gradiente, para minimizar una función de error igual a la diferencia cuadrática media entre la salida deseada y la salida de la red (Rumelhart, 1986), (Romero, 1990).

Un estructura del perceptrón multicapa, llmada Red Neuronal de Tiempo Diferido (RNTD) se ha desarrollado para aplicaciones en voz (Mariani, 1989). En la figura 3 se muestra una red do este tipo. Las unidades de una capa o

nivel se representan por una matriz de cuadros, donde cada cuadro representa una unidad. En la figura se aprecia una "ventana" (marcada con líneas de doble grueso) de 3 columnas en la capa de entrada, que está conectada a la primera columna de unidades de la capa oculta. Cada unidad de la capa de entrada, que está dentro de la ventana, tiene conexiones a las dos unidades que se indican en la capa oculta. Existe otra ventana, D columnas desplazada a la derecha, que establece el siguiente conjunto de conexiones. Sucede lo mismo para el resto de las columnas de la capa de entrada.

La ventaja de las RNTD sobre las redes totalmente conectadas (una unidad tiene conexiones con todas las unidades del nivel siguiente) es una drástica disminución del número de conexiones y en consecuencia una disminución en el tiempo de aprendizaje de este tipo de redes.

UNA HERRAMIENTA PARA LA INVESTIGACION EN RECONOCIMIENTO DE VOZ (HIREV)

Se ha desarrollado una herramienta para implementar sistemas de reconocimiento de voz. La herramienta consiste en un conjunto de programas que pueden acoplarse unos con otros para formar sistemas de reconocimiento; además de utilerías para graficación de resultados intermedios (Romero, 1990).

Dentro del contexto expuesto anteriormente, la etapa de extracción de

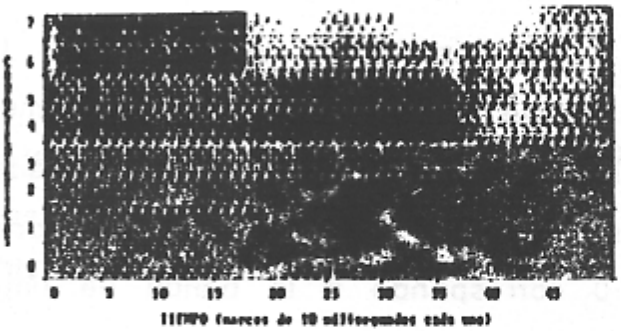

FIGURA 5.- PATRON DE CARACTERISTICAS DE LA PALABRA /CERO/.

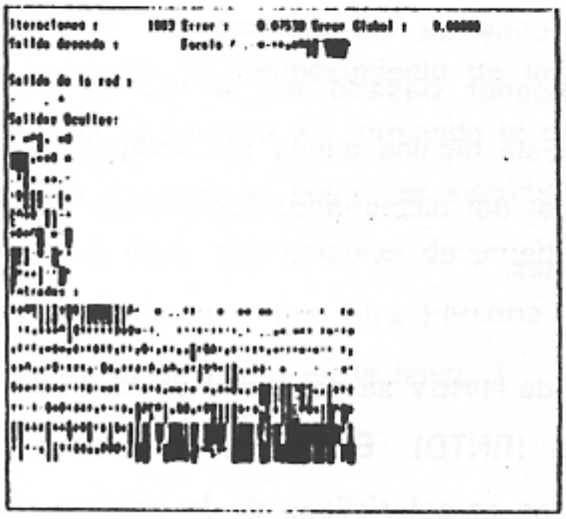

FIGURA 6.- DESPLIEGUE DEL SIMULADOR DE REDES NEURONALES ANTE LA PALABRA /CERO/.

SISTEMA	PERSONA	PALABRAS RECONOCIDAS	PALABRAS NO RECONOCIDAS	PALABRAS EQUIVOCADAS
TRADICIONAL	A	6	2	2
	B	5	3	2
NEURONAL	A	7	1	2
	B	7	1	2
HIBRIDO	A	10	0	0
	B	8	1	1

FIGURA 7- RESULTADOS OBTENIDOS EN EL RECONOCIMIENTO DE LOS DIGITOS.

características está basada en un banco de filtros digitales tipo Butterworth de octavo orden. En la figura 4 se muestra la señal de la palabra cero antes de la extracción (salida digitalizada del micrófono) y en la figura 5 se muestra el patrón de características asociado. La característica 0 corresponde a la banda de bajas frecuencias y la característica 7 a la banda de altas frecuencias.

HIREV permite construir un sistema de reconocimiento tradicional que utiliza un clasificador basado en la técnica de "Doblado Dinámico de Tiempo" (DDT). Esta técnica alínea (en tiempo) los patrones de comparación (de prueba con el del diccionario) como paso previo al cálculo de similitud de los dos patrones.

Como una parte de HIREV se desarrolló un simulador de Redes Neuronales de Tiempo Diferido (RNTD). En la figura 6 se muestra el despliegue del simulador ante las características de la palabra cero.

En resumen, HIREV permite construir tres tipos de sistemas de reconocimiento:

a) Sistema Tradicional.- Extracción de características y clasificador basado en el "Doblado Dinámico en Tiempo".

b) Sistema con clasificador neuronal.- Extracción de característica y clasificación usando el simulador de RNTD.

c) Sistema híbrido.- Extracción de características,

alineamiento temporal de patrones utilizando DDT y clasificación basada en el simulador de RNTD.

EXPERIMENTO

Los tres sistemas anteriores se aplicaron al reconocimiento de los dígitos. Se adquiere (una vez) cada dígito (por la persona A), formando el diccionario de patrones de referencia (explícito o implícito según el sistema), y con ellos se entrenan los 3 sistemas construídos. Las palabras de prueba son los dígitos pronunciados por la misma persona y por otra persona diferente (persona B). Los resultados obtenidos se muestran en la figura 7.

CONCLUSIONES

De los resultados del experimento podemos observar que el perceptrón multicapa (RNTD) realiza eficazmente su tarea de clasificación. El sistema híbrido presenta sólo 2 errores para las palabras pronunciadas por una persona diferente a la que realizó el entrenamiento. En conclusión, el uso de un clasificador basado en una RNTD realiza eficazmente su función y combinado con técnicas tradicionales (alineamiento temporal de patrones por medio de DDT), permite construir sistemas híbridos con un mejor comportamiento que los sistemas que emplean sólo un enfoque.

RECONOCIMIENTO

El trabajo realizado por el M. en C. Leonardo Romero fue patrocinado por CONACYT e Instituto de Investigaciones Eléctricas.

REFERENCIAS

(Caudill, 1987)

M. Caudill, "Neural Networks PRIMER" Partes I, II, III y IV. Aparecieron en la revista "AI EXPERT" en Diciembre de 1987, Febrero de 1988, Junio de 1988 y Agosto de 1988.

(Lippmann, 1987)

R. P. Lippmann, "An Introduction to Computing with Neural Nets". IEEE ASSP Magazine, Abril de 1987.

(Mariani, 1989)

J. Mariani, "Recent Advances in Speech Processing". IEEE, ICASSP, 1989.

(Romero, 1990)

L. Romero M., "Reconocimiento de Voz Mediante Redes Neuronales: Aplicación a Palabras Aisladas". Tesis de Maestría, ITESM, Campus Morelos, Mayo de 1990.

(Rumelhart, 1986)

D. E. Rumelhart y J. L. McClelland (Eds.), "Paralell Distributed Processing" MIT Press, 1986.

(Sotelo, 1989)

R.S. Sotelo H., "Sistema de Adquisición y Extracción de Información para el Reconocimiento de Formas". Tesis de Maestría, ITESM - Campus Morelos, 1989.