RECONOCIMIENTO DE SONIDOS VOCALICOS EMPLEANDO REDES NEURONALES ARTIFICIALES

ING. VINICIO CARRERA R.

QUITO – ECUADOR

RESUMEN

Este trabajo presenta una de las alternativas existentes para el desarrollo de sistemas que sean capaces de realizar un reconocimiento eficiente de la voz; esta nueva alternativa se fundamenta en la aplicación de la tecnología de Redes Neuronales Artificiales al campo del Procesamiento Digital de Señales.

ABSTRACT

This work present an alternative for development of systems capables of making efficient recognition of voice; this new alternative is based on the aplication of the tecnology of Artificial Neural Networks to the Digital Signal Processing.

INTRODUCCION

La ventaja de poder tener un computador capaz de entender el lenguaje humano en forma hablada es indiscutible y ha sido la meta de muchas investigaciones en los últimos años; sin embargo, su observancia no se llevó a cabo sino hasta hace muy poco tiempo.

Esta tarea del reconocimiento de la voz, pudo llegar a ser una realidad debido al gran avance que han sufrido las técnicas para el procesamiento digital de señales, y si hoy, a todo ello se añade el advenimiento de la tecnología de Redes Neuronales Artificiales, se llegan a romper muchas de las limitaciones existentes al respecto.

Una Red Neuronal Artificial permite que una vez procesadas digitalmente las señales sonoras del habla, no se tenga que establecer reglas o realizar análisis estadísticos complejos para la determinación del fonema en proceso de reconocimiento; es decir, debido a la capacidad de aprendizaje de las Redes Neuronales, únicamente se requiere una fase de entrenamiento donde se le presentan a la red, ejemplos típicos de determinados fonemas con sus correspondientes valores deseados.

De esta forma, una vez que el sistema está entrenado con un conjunto básico de fonemas, se puede presentar a las entradas de la red cualquier señal de voz, y establecer un reconocimiento efectivo de los fonemas presentes dentro de la misma.

Figura 1

diagrama de bloques del sistema necesario para implementar el reconocimiento de la voz, es el que se muestra en la *figura 1*. En él se observan un bloque de adquisición de datos, un bloque de procesamiento digital de señales y un bloque correspondiente a la Red Neuronal Artificial.

EL SISTEMA DE RECONOCIMIENTO

El objetivo desde el cual se parte en el presente trabajo, es el reconocimiento de los sonidos vocálicos en español (*a, e, i, o, u*) a partir de una Red Neuronal entrenada con parámetros característicos de la Densidad Espectral de Potencia en ventanas de corta duración de la señal de voz. Se ha iniciado con el reconocimiento de las vocales, debido a que sus fonemas presentan una mayor Energía Cuadrática Media, lo cual viene a simplificar el diseño del sistema así como la tolerancia del mismo.

El bloque de adquisición de datos para este caso, consta simplemente de un transductor acústico (micrófono), un acondicionador y amplificador de señales y un conversor análogo–digital. Todo ello está integrado en una tarjeta mediante la cual se logra muestrear la señal de voz a 8 KHz, con 8 bits por muestra; adicionalmente se generan los archivos correspondientes con los valores de las muestras, para usarlos posteriormente en los procesos de entrenamiento y/o recordación de la Red Neuronal.

En la parte referente al procesamiento digital de señales, se requieren procesar los archivos de voz digitalizada para encontrar los parámetros característicos de la Densidad Espectral de Potencia, dentro de este paso ha sido preciso determinar la eficacia de algunos métodos propuestos como estimadores espectrales, siendo los usados:

1. El Periodograma Clásico representado por los Coeficientes de la FFT (Transformada Rápida de Fourier), y;
2. El Modelo Autoregresivo representado ya sea por los Coeficientes Predictores Lineales o por los Coeficientes de Reflexión de la estructura Lattice.

Este bloque, así como el correspondiente a la Red Neuronal, son implementados a través de un programa escrito en lenguaje C, que puede ejecutarse bajo el sistema operativo MS–DOS o UNIX V.

Para el caso de utilizar el Periodograma Clásico como estimador espectral, se lo aplica con 256 muestras de las cuales únicamente se utilizan para entrenamiento y/o reconocimiento, los primeros 127 módulos (se excluye X[0] y desde X[128] a X[255]), de tal forma que la Red Neuronal debe poseer 127 elementos de procesamiento en su estrato de entrada. Se podría utilizar un mayor número de elementos en el estrato de entrada, pero el incremento en la eficiencia del sistema es despreciable respecto al incremento de recursos computacionales

necesarios para un rendimiento aceptable de la Red Neuronal en lo que tiene que ver a tiempo de ejecución.

Cuando se usa el Modelo Autoregresivo, se hace el entrenamiento y/o recordación de la red, mediante los doce primeros valores, ya sean estos los Coeficientes Predictores Lineales o los Coeficientes de Reflexión, lo cual implica que la Red Neuronal ha de tener apenas 12 elementos de procesamiento en su estrato de entrada. Se ha demostrado a través de las técnicas de procesamiento digital de señales que este número de coeficientes es suficiente para representar cualquier señal de voz (proceso estocástico ligeramente variable en el tiempo).

Antes de aplicar a la Red Neuronal los valores de los coeficientes de la FFT o los coeficientes del Modelo Autoregresivo, estos deben ser normalizados a valores dentro de un rango que varía entre 0.00 y 1.00.

A	1.0	0.0	0.0	0.0	0.0
E	0.0	1.0	0.0	0.0	0.0
I	0.0	0.0	1.0	0.0	0.0
O	0.0	0.0	0.0	1.0	0.0
U	0.0	0.0	0.0	0.0	1.0

Tabla 1

Junto a todo esto, como el objetivo del sistema es reconocer los cinco sonidos vocálicos del alfabeto, la Red Neuronal utilizada debe poseer cinco elementos de procesamiento en el estrato de entrada, de tal forma que se pueda representar cada sonido mediante la simbología dada en la *tabla 1*.

Con ese tipo de representación se garantiza la independencia de los fonemas en el proceso de reconocimiento, ya que sólo un elemento del estrato de salida va estar activo y por ende se pueden aplicar estrategias de competición que aseguran la separación de las señales de entrada en clases válidas.

La Red Neuronal

Las subrutinas del programa que corresponden a la simulación de la Red Neuronal Artificial, utilizan el algoritmo Back-Propagation para implementarla; el número de estratos es seleccionable por medio de menús, pero en el presente trabajo se ha utilizado para la mayor parte de las pruebas únicamente tres (3). La constante de aprendizaje (η) y la constante de momento (α) son también ajustables a través de menús, pero debido a pruebas de rendimiento que se han llevado a cabo anteriormente, se ha escogido $\eta = 0.90$ y $\alpha = 0.60$, aunque en algunos casos se han dado resultados más aceptables con $\eta = 1.00$ y $\alpha = 0.50$. Las conexiones entre todos los elementos de procesamiento son inicializadas a un valor aleatorio entre -0.10 y 0.10.

Cuando se emplea la FFT como estimador espectral de la señal de voz, la Red Neuronal consta de 127 elementos de procesamiento en el estrato de entrada, 20 en el estrato intermedio y 5 elementos en el estrato de salida. El número de elementos de procesamiento en los estratos de entrada y de salida queda determinado por el problema; el número de elementos en el estrato intermedio se determina en base a la cantidad de regiones linealmente independientes que se pueden obtener en el dominio de clasificación.

Para el caso de emplear el Modelo Autoregresivo, la red consta de 12 elementos de procesamiento en el estrato de entrada, 2[illegible] el estrato intermedio y 5 en el estrato salida. Para este caso se usó también redes 4 estratos, conteniendo 12, 20, 15 [illegible] elementos de procesamiento respectivamente. número de elementos de procesamiento en [illegible] uno de los estratos queda determinado tan en base a las observaciones anteriores.

Adicionalmente, debido a que el sis debe centrar su reconocimiento en los fon vocálicos, se requiere que éste sea capaz distinguir los sonidos de las vocales de correspondientes a las consonantes, ya que una palabra completa se tendrá una combina de ambos tipos de sonidos. Es por ello que accionar de la Red Neuronal debe [illegible] habilitado o deshabilitado a través de un v determinado por la Energía Cuadrática M presente en la ventana de corta duración d señal de voz analizada; si el valor ex cierto límite (o valor umbral) se procede [illegible] ejecución del mecanismo neuronal, [illegible] contrario se deshabilita su funcionami hasta que las muestras vuelvan a presentar densidad espectral suficiente para consider un sonido vocálico.

Debido a la gran capacidad de cál necesaria para el desarrollo de [illegible] aplicación, tanto en el procesamiento dig de señales como en el algoritmo de la Neuronal, es preciso la utilización de equipo con características no inferiores a sistema 80386, de tal forma que se p observar un rendimiento aceptable (todavía en tiempo real) del sistema.

Aprendizaje

El proceso de entrenamiento de la red lo lleva a cabo mediante la aplicación muestras seleccionadas de los archivos que obtuvieron a través de la tarjeta adquisición de datos. En este caso son fon puros representativos de cada una de vocales; debiendo estar acompañadas de indicador que especifica el sonido vocálico cual corresponden. Las muestras para realización del entrenamiento son presente repetidamente en forma alternada h conseguir que la red reconozca adecuadam cada uno de los ejemplos enseñados.

En este caso, se utilizó para entrenamiento, sonidos vocáli correspondientes a diez personas diferentes pronunciaron la secuencia:

aaaa eeee iiii oooo uuuu

Estos archivos son usados para entrenamiento de la red, tanto para el caso reconocimiento dependiente del locutor (la reconoce un sólo locutor a la vez) como para caso de reconocimiento independiente locutor (la red se entrena con los fonemas la diez personas al mismo tiempo).

De cada vocal pronunciada por cada pers se extrajeron 4 muestras representativas toda la señal en el caso de la FFT y muestras en el caso del Modelo Autoregresi La aplicación de estos conjuntos de valores la red se llevó a cabo fuera de li (off-line), de acuerdo a los requerimientos Modelo Back-Propagation.

Cuando se emplean los fonemas de un s locutor, se requieren menos de 200 pasos entrenamiento al usar la FFT, y aproximadame 2000 pasos con los Coeficientes Predicto Lineales o con los Coeficientes de Reflexi lográndose tener un error menor al 10% en reconocimiento de los archivos entrenamiento.

Para el caso de reconocimiento pendiente del locutor, y al usar la FFT, necesario realizar mil pasos de amiento (5 veces cada muestra), itándose 45 segundos en un computador /33 Mhz bajo el sistema operativo MS-DOS y egundos bajo el sistema UNIX V. Al final de mil pasos de entrenamiento, el sistema fue de reconocer las muestras seleccionadas un error menor al 3% (con 500 pasos de trenamiento el error es menor al 10%).

Cuando se emplean el Modelo Autoregresivo se necesitan 30 mil pasos de entrenamiento (20 veces cada muestra), necesitando 150 segundos en el mismo computador anterior bajo MS-DOS y 110 segundos bajo UNIX. Al final del trenamiento, el sistema es capaz de reconocer s archivos de entrenamiento con un error nor al 10%.

De todas maneras, en el número de pasos de entrenamiento va a influir en cierta forma la randomización inicial que se efectúa sobre las conexiones de la Red Neuronal. Esta es una característica aplicada a cualquier utilización del Modelo Back-Propagation.

Recordación

El proceso de recordación (reconocimiento por parte de la red) consiste en la presentación de una señal de voz cualquiera y la obtención de su correspondiente clasificación de acuerdo a la presencia de los sonidos vocálicos dentro del contexto de la señal.

Para probar el sistema se digitalizaron varias palabras que contienen diferentes vocales, por ejemplo: *electrónica*, *estadio*, *eucalipto*, *neuronal*, *vaselina*. Estas señales fueron tratadas con el respectivo procesamiento digital y analizadas conforme las variaciones de la Densidad Espectral de Potencia presente en cada una de las ventanas de corta duración.

El tiempo empleado en el proceso de recordación es bastante bajo, principalmente en el sistema que utiliza el Modelo Autoregresivo, ya que el tiempo usado en el cálculo de los doce coeficientes es mucho menor que el tiempo empleado en el cálculo de la FFT con 256 muestras. Igualmente una Red Neuronal de 42 elementos de procesamiento y 465 conexiones se procesa mucho más rápido que una de 152 elementos y 2665 conexiones.

RESULTADOS

Para determinar la eficiéncia del sistema, primeramente se probó el comportamiento del procedimiento a una entrada que es una replica de las muestras usadas en el entrenamiento de la red. En este caso, el proceso que utiliza la FFT como estimador espectral, obtuvo resultados mucho más claros, con una tasa de aciertos igual al 100% en el caso de un locutor y 98% en el caso de la combinación de varios locutores (*Tabla 2*).

ENTRADA	aaaaeeeeiiiiiooooouuuu
F.F.T.	aaaaeeeeiiiiiooooouuuu
C.P.L.	aaa.eeeiiiiiooooooouo
C.R.	a.oa.eee.iiooooouuou

Tablá 2

De igual manera se propone a la red, el reconocimiento de sonidos vocálicos dentro del contexto de una palabra completa, notándose cierta degradación que en las experiencias realizadas mostré ser mucho menor al usar la FFT como método para el procesamiento digital de señales. La degradación con el Método Autoregresivo es sumamente alta especialmente en el caso del reconocimiento con independencia del locutor; cuando se usa un reconocimiento dependiente del locutor, el rendimiento alcanzado es todavía aceptable. (Tabla 3).

ENTRADA	eelleeccctrrooooooonniicaaa
F.F.T.	ee..ee.....aaoooo..ii..ao
C.P.L.	ee.....aaaaaa...i..oa
C.R.	e... e.....aaaoaa...i...a

Tabla 3

Esta degradación mayor del sistema al trabajar con los Coeficientes Predictores Lineales o con los Coeficientes de Reflexión como métodos de procesamiento digital de señales, se debe básicamente al menor número de valores mediante los cuales la red debe clasificar el sonido vocálico; sin embargo, queda todavía la expectativa que si se aumenta el número de bits por muestra durante el proceso de adquisición de datos, el reconocimiento va a tener un mejor desempeño, conforme lo muestran las experiencias análogas a este tipo de tratamiento de señales de voz.

Cabe mencionar, que todas las muestras de voz utilizadas corresponden a personas del sexo masculino, lo cual conlleva la restricción de no utilizar en los procesos de reconocimiento muestras de voz pertenecientes a personas del sexo femenino, ya que ello degrada significativamente los resultados obtenidos por el sistema.

Sin embargo, es motivante el hecho de que el tiempo empleado en el reconocimiento de una señal al usar el Modelo Autoregresivo es muy inferior si se lo compara con el tiempo empleado por el Método que emplea el Periodograma Clásico. De igual manera el tiempo de entrenamiento por cada paso que se realice es menor, aunque se puede vislumbrar como un inconveniente el hecho de que aumente el número de pasos de entrenamiento necesarios.

Discusión

Este trabajo refleja las ventajas de utilizar la tecnología de Redes Neuronales Artificiales para facilitar el Reconocimiento de Voz, debido a que reduce significativamente la complejidad de los algoritmos utilizados, así como el tiempo empleado en el desarrollo del proceso (a nivel de reconocimiento). Naturalmente existen todavía algunas variaciones al esquema propuesto que deben investigarse para mejorar, y en algunos casos ampliar, la eficiencia y capacidad del reconocimiento efectuado.

Como se mencionó, actualmente se ha limitado el trabajo al reconocimiento de los cinco sonidos vocálicos, además de no cumplir todavía ciertas características de procesamiento en tiempo real.

Una dificultad manifiesta es también el tiempo de entrenamiento relativamente elevado para este modelo de red, pero con el advenimiento de equipos cada vez más potentes, así como sistemas operativos que aprovechen al máximo el hardware disponible, este tiempo se podrá reducir.

Finalmente, hay que destacar que la selección de las muestras a usar en el

entrenamiento de la red es un serio inconveniente en el desenvolvimiento normal del sistema, ya que las muestras deben reflejar los rasgos más generales del fonema en cuestión. Los resultados presentados conciernen a la toma de muestras que corresponden a secciones aleatorias de los archivos obtenidos; no se descarta la posibilidad de que se pueda obtener un mejor rendimiento si se hace un análisis previo de la generalidad o particularidad de las muestras obtenidas a través del bloque de adquisición y procesamiento digital.

Trabajo Futuro

En base a las anotaciones anteriores, se espera poder compactar el esquema propuesto en un sistema mucho más dedicado (con inclusión de hardware). Este mejoramiento podrá incluir: adquisición de datos a través de un canal D.M.A. (Acceso Directo a Memoria), estandarización de los tratamientos para el procesamiento digital de señales de tal forma que se permita ampliar el campo de los fonemas a reconocer, y realización de parte del procesamiento digital y la Red Neuronal mediante la utilización de hardware.

Es también necesario tratar de aprovechar las características del Modelo Autoregresivo, para lo cual se deben realizar pruebas de eficiencia con señales digitalizadas a 12 bits por muestra, ya que la expectativa de superación en los resultados obtenidos está latente de acuerdo a las características procuradas al trabajar con los Coeficientes Predictores Lineales y con los Coeficientes de Reflexión.

Como una última alternativa, se plantea la utilización de otros Modelos de Redes Neuronales, de tal forma que se permita el aprendizaje en línea (on-line), o un aprendizaje no-supervisado. Los modelos con mayores expectativas a este respecto son el modelo *Counter-Propagation* y el modelo A.R.T. (*Adaptive Resonance Theory*). También surge como un tema de investigación la aplicación del modelo S.P.R. (*Spatio-temporal Pattern Recognition*) el cual permitiría el reconocimiento de una secuencia espacio-temporal de sonidos, que en definitiva sería la clasificación de palabras completas en lugar de fonemas aislados

CONCLUSIONES

El proceso de reconocimiento de voz ha evolucionado tanto que hoy en día, es ya un hecho; sin embargo, la utilización de la tecnología de Redes Neuronales Artificiales abre nuevas expectativas al reducir la complejidad de los algoritmos y la forma en que se lleva a cabo. Se presenta de esta forma una nueva alternativa de investigación que sin duda conlleva la consecución de resultados sorprendentes en este campo tan importante e interesante denominado Reconocimiento de Voz.

El desempeño del sistema está sobre muchos otros esquemas existentes para el reconocimiento del habla humana, siendo algo representativo de ello, el caso en que se usan como estimadores espectrales de la señal de voz los coeficientes de la FFT.

La desventaja que se evidencia, en el tiempo de entrenamiento requerido fuera de línea, pero como se cita al comparar el Modelo Back-Propagation con otros modelos de Redes Neuronales existentes: "algún precio se tiene que pagar por un mejor desempeño".

Las perspectivas abiertas con esta tecnología son inmensas tanto al usar el Periodograma Clásico como a través del M Autoregresivo, estando caracterizada la F un reconocimiento más eficiente y Coeficientes Predictores lineales así co Coeficientes de Reflexión por la velocid el procesamiento.

RECONOCIMIENTOS

Un agradecimiento especial al Ing. León MsC., ya que sin su colaboració hubiera sido posible la realización de investigación.

REFERENCIAS

1. CARRERA VINICIO, Las Redes Neuron Artificiales Aplicadas al Reconocimi de Patrones, Tesis de Grado-ESPE, Ecua 1992.

2. D.A.R.P.A., DARPA Neural Network St AFCEA International Press, U.S.A., 19

3. LEON RUBEN, Clasificación de Son Vocálicos Utilizando Técnicas Predicción Lineal, ESPE, Ecuador, 19

4. VEMURI V., Artificial Neural Netwo Theoretical Concepts, IEEE, U.S.A., 1

5. OPPENHEIM & SCHAFER, Digital Si Processing, Prentice/Hall Internatio U.S.A., 1975.