PAIDEIA XXI

Vol. 3, Nº 4, Lima, diciembre 2013, pp. 120-127

SINTETIZADOR NUMÉRICO DE VOZ PARA PERSONAS CON DISCAPACIDAD DEL HABLA

Pedro Freddy Huamaní Navarrete / Magaly Jannet Ramírez Carvo

Resumen

En este artículo se muestra el desarrollo de un sintetizador numérico de voz utilizando técnicas de procesamiento digital de imágenes y redes neuronales artificiales. Tales técnicas fueron implementadas utilizando el software Matlab instalado en una PC, la cual corresponde a la interface oral de comunicación para el discapacitado. Se realizó el reconocimiento de los números a partir de manuscritos digitalizados, hasta un total de tres dígitos para la parte entera y dos para la parte decimal, obteniendo resultados satisfactorios.

Palabras Clave: Transformaciones Morfológicas, Segmentación de imágenes, Red Neuronal Multicapa, Software Matlab.

Abstract

This article shows the development of voice synthesizer numerical techniques using digital image processing and artificial neural networks. Such techniques were implemented using Matlab software installed on a PC, which corresponds to oral communication interface for the handicapped. Recognition was performed numbers from digitized manuscripts, to a total of three digits for the integer part and two for the integer part, obtaining satisfactory results.

Key words: Morphological transformations, image segmentation, Neural Network Multilayer, Software Matlab.

INTRODUCCIÓN

Actualmente existe un gran número de personas con discapacidad del habla (mudos), que para expresarse emplean un lenguaje de signos o señas, el cual posee un alto grado de dificultad en su aprendizaje. Por tal motivo, el presente artículo, a través de las técnicas de procesamiento de imágenes y entrenamiento de una red neuronal, brinda una alternativa para lograr una comunicación por medio de un sintetizador de voz, solo para números.

La carencia de sintetizadores de voz numéricos en el mercado local y la importancia de aportar a la comunicación de personas con discapacidad del habla, motiva al uso de la tecnología digital a través del uso de diversos algoritmos de procesamiento de imágenes, para la realización de técnicas de segmentación y posterior reconocimiento de patrones, utilizando redes neuronales multi capas con algoritmo de aprendizaje backpropagation.

Cabe resaltar que el sintetizador planteado está basado en la síntesis de voz de números con tres dígitos en la parte entera y dos dígitos en la parte decimal, los cuales corresponden a manuscritos en hojas de trabajo predeterminadas, y con dimensiones particulares. De esta manera, la implementación del sintetizador de voz es realizado mediante una PC, el cual corresponde a la interfaz oral de comunicación entre el discapacitado y las personas con los que desee comunicarse, y haciendo uso de las bondades que ofrece las librerías del software Matlab para elaborar la estructura lógica del sintetizador propuesto.

DESARROLLO DEL TRABAJO

Para el desarrollo de este trabajo se optó por utilizar: un micrófono omnidireccional, con respuesta de frecuencia de 100 Hz a 11 KHz, impedancia de salida 2000 Ohms y una dirección de -60Db; un escáner convencional con resolución de 2400x4800 ppp, color de 48 bits y tecnología OCR. Además, se usó una PC convencional con procesador de 2.4 GHz, RAM de 512 MB y HD de 32 GB, el software Matlab y una base de datos con grabaciones de voz en archivos WAV, codificados con PCM, frecuencia de muestreo de 8KHz, 16 bits por muestra y mono canal.

Este trabajo está compuesto por cinco etapas, que a continuación se procede a describir:

1. Creación de base de datos con grabaciones de voz en archivos WAV.

La base de datos creada está conformada por grabaciones digitales, las cuales fueron realizadas para la construcción del patrón digital de voces en formato *.WAV. Estas grabaciones tuvieron diferentes tiempos de duración y fueron agrupadas en 5 clases:

- 1ra clase. Grabaciones correspondientes a los números del 0 al 9.
- 2da clase. Grabaciones correspondientes a los números del 11 al 19.

- 3ra clase. Grabaciones correspondientes a los números 10, 20, 30, 40, 50, 60, 70, 80 y 90.
- 4ta clase. Grabaciones correspondientes a los números 100, 200, 500, 700 y 900.
- 5ta clase. Grabaciones correspondientes a nexos y conectores: Y, PUNTO.

De esta manera, con la combinación apropiada de estos archivos digitales de voz, fue posible reproducir toda cantidad numérica desde 0 hasta 999. De igual forma, para la parte decimal, fue posible realizar la combinación apropiada de archivos digitales, para reproducir digitalmente desde 0.00 hasta 0.99.

2. Digitalización de manuscritos.

La digitalización de los manuscritos se basó en la operación de escaneado de una hoja Bond A4, donde el manuscrito se representó de forma legible y libre de ruido. Para ello se utilizó una resolución de 230x280 pixeles con 1 bit por pixel (blanco y negro), y en un archivo digital de formato BMP. Estas características permitieron un eficiente procesamiento y posterior reconocimiento de cada patrón numérico. Seguidamente, en la Figura 1, se muestra el resultado de la digitalización de un manuscrito que representa a un número con tres dígitos en la parte entera, y dos en la parte decimal.

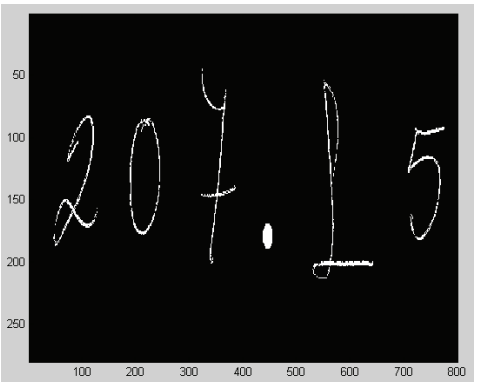

Figura 1: Imagen perteneciente a un manuscrito digitalizado.

3. Procesamiento de los manuscritos digitalizados.

El procesamiento de la imagen digital está basado en seis partes, tal como lo señala el diagrama de bloques de la Figura 2. En este diagrama de bloques se puede observar claramente el uso de operaciones morfológicas, el uso de la segmentación, así como de la decimación, tanto de filas como de columnas, y de una operación de determinación y generación de valores padrones.

Para la realización de las operaciones morfológicas, se utilizó un elemento estructural del tipo "square" con un ancho de pixeles igual a 3. Su elección se produjo después de realizar varias pruebas y de verificación del mejor comportamiento frente a la característica de la imagen digital procesada.

Posteriormente, la imagen digital fue segmentada por cada dígito, y luego decimada con el fin de reducir el número de filas y columnas, pero conservando la forma original. Antes

de aplicarse sobre la red neuronal, se procedió a la determinación de los valores padrones, a partir de tres sumas verticales y tres horizontales sobre cada número de la imagen binaria resultante.

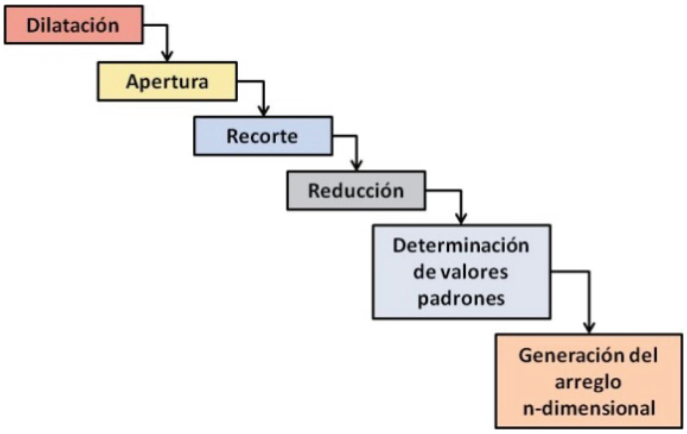

Figura 2: Diagrama de bloques del procesamiento de las imágenes digitales.

4. Entrenamiento de patrones de imágenes.

El entrenamiento de la red neuronal fue realizado en base a vectores conformados por los seis valores padrones normalizados. Para este procedimiento, se trabajó con el número de capas y neuronas anteriormente indicado, así como también con funciones de transferencia del tipo no lineal para las primeras tres capas, y del tipo lineal para la última capa de neuronas. Asimismo, con ayuda del toolbox de Neural Networks del Matlab, se logró realizar el entrenamiento respectivo.

Se planteó el uso de 250000 iteraciones con el interés de lograr un mínimo error. Asimismo, una tasa de aprendizaje de 0.008 y un error mínimo cuadrático de 0.0002. Para una mejor clasificación y reconocimiento de cada dígito a través de la red neuronal, se propuso determinar valores que permitan establecer los rangos de reconocimiento para cada número. Tal como lo muestra la Tabla 1.

5. Generación de la voz digital.

Una vez entrenada la red neuronal, se procedió a la generación de la voz digital correspondiente. Dicha generación se obtuvo a partir de la concatenación de vectores obtenidos después de la lectura de los archivos de audios digitalizados y correspondientes a cada número reconocido. Esto quiere decir, que después de reconocer cada dígito del número completo, se procedió a la apertura y lectura del archivo de audio digital de la base de datos previamente creada. Esto es realizado con cada número para posteriormente unir los vectores y reproducirlo todo en conjunto a una misma frecuencia de muestreo.

Número	Rango
0	98 - 102
1	8 – 12
2	18 - 22
3	28 – 32
4	38 – 42
5	48 – 52
6	58 – 62
7	68 – 72
8	78 – 82
9	88 – 92

Tabla 1: Rangos permisibles para cada número.

PROCESO DE SIMULACIÓN

Se inicia con el hecho de que el manuscrito se encuentre digitalizado (Figura 1), para luego realizar las operaciones morfológicas de dilatación y apertura, tal como lo muestra la Figura 3. Luego, se procede al recorte de la imagen con la finalidad de obtener cada número de manera independiente. Este recorte o segmentación de la imagen genera a su vez seis imágenes de dimensiones diferentes, incluyendo el punto decimal. La obtención del recorte se logró considerando sumas perpendiculares y teniendo en cuenta los pixels de color blanco. De esta manera se pudo aislar cada uno de los dígitos, con que se encuentra conformado el número completo.

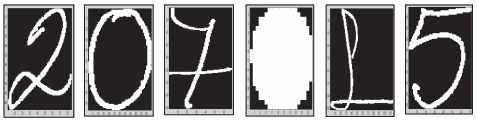

Figura 3. Operación de segmentación sobre la imagen digitalizada.

Posteriormente, se reduce el tamaño de cada imagen procediendo a una operación de decimación por filas y por columnas. Luego, se determinan los valores padrones conformados por seis valores de la representación matricial de la imagen reducida. Este conjunto de valores padrones formará el vector de entrada a la red neuronal. A continuación, en la Figura 4 se muestra un ejemplo del procedimiento de obtención de los seis valores de representación matricial, a partir de sumas de elementos de la matriz en tres filas y tres columnas diferentes.

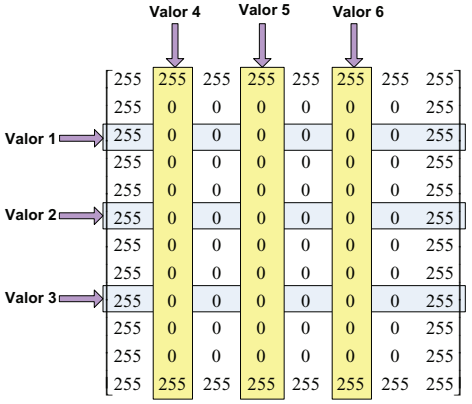

Figura 4: Representación matricial de la Imagen reducida del número 'cero', con sus respectivas posiciones en las filas y columnas.

Obtenida las sumas, se finaliza con la normalización del vector de seis elementos con la finalidad de facilitar el entrenamiento de la red neuronal. En la figura 5 se muestra la matriz con los seis valores padrones por cada dígito y el punto decimal, y correspondiente al número "207.15".

0.3716	0.3235	0.2265	0.3015	0.1873	0.0894
0.1487	0.2157	0.1132	0.4523	0.0937	0.1789
0.4460	0.4313	0.2265	0.4523	0.0937	0.1789
0.2973	0.5392	0.4529	0.3015	0.9366	0.7155
0.5946	0.4313	0.7926	0.4523	0.1873	0.3578
0.4460	0.4313	0.2265	0.4523	0.1873	0.5367

Figura 5: Matriz con valores padrones del número “207.15”

Para el entrenamiento y comprobación del desempeño de la red neuronal, se utilizaron manuscritos legibles sobre hojas de trabajo de tamaño A4 y tipo Bond. Estos manuscritos corresponden a un grupo de seis personas, entre las edades de 20 a 30 años, y a las cuales se les solicitó una muestra de 10 dígitos. Por lo tanto, el universo de trabajo llegó a ser igual a 60 muestras numéricas.

La red neuronal utilizada (Figura 6) fue entrenada con el tipo de aprendizaje supervisado y una función de transferencia del tipo lineal en la última capa de neuronas. Con esto se eligió un vector de salida deseada, correspondiente a los números del cero al nueve según el vector de entrada utilizado. Ver la siguiente representación para el vector de entrada de entrenamiento de la red, y su correspondiente vector de salida deseada.

```
>> X = [n0 n1 n2 n3 n4 n5 n6 n7
n8 n9 ]; % matriz de 6x60
>> Yd = [10*ones(1,6) 1*ones(1,6)
2*ones(1,6) ...
3*ones(1,6) 4*ones(1,6) 5*ones(1,6)
6*ones(1,6) ...
7*ones(1,6) 8*ones(1,6) 9*ones(1,6)]
/ 10;
```

Donde:

n0, n1, ..., n9, representan a las matrices con valores padrones tal como se señaló en la figura 5.

Yd representa a una matriz de 1x60 correspondiente a los números 1, 0.1, 0.2, 0.3, ..., 0.9. Donde, el valor “1” hace referencia al dígito de entrada “0”, el valor “0.1” hace referencia al dígito de entrada “1”, y así sucesivamente hasta el valor “0.9” que hace referencia al digito de entrada “9”.

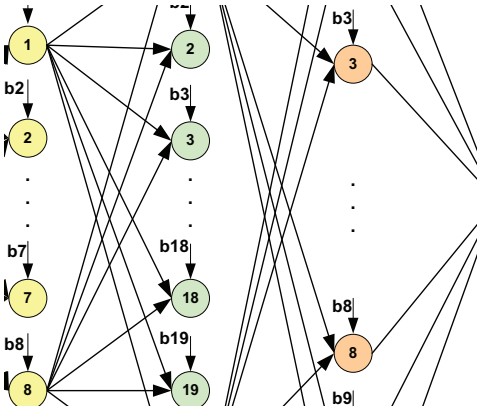

Figura 6: Esquema de la Red Neuronal Multicapa utilizada.

El proceso de entrenamiento de la red neuronal utilizó seis vectores padrones por cada número. Y la etapa de inicialización y entrenamiento se realizaron haciendo uso del toolbox de Neural Networks.

```
limites = [ zeros(6,1) ones(6,1) ];
net = newff( limites , [ 8 20 10 1 ] ,
{ ‘tansig’ ‘tansig’ ‘tansig’ ‘purelin’ } ,
‘traingd’ );
net.trainParam.epochs = 250000;
net.trainParam.lr = 0.008;
net.trainParam.mse = 0.0002;
net = train( net , X , Yd );
```

A continuación, se muestran los archivos de voz utilizados para sintetizar el ejemplo tomado en la Figura 1.

```
>> X1= wavread (‘2.wav’);
```

```
%archivo de sonido DOS
>> X2= wavread ('100s.wav');
%archivo de sonido CIENTOS
>> X3 = wavread('7.wav');
%archivo de sonido SIETE
>> X4 = wavread('punto.wav');
%archivo de sonido PUNTO
>> X5 = wavread('15.wav');
%archivo de sonido QUINCE
>> X = [ X1 X2 X3 X4 X5 ];
%concatenando los vectores
>> sound( X , 8000 )
%reproducción con Fs = 8000 Hz
```

RESULTADOS

La respuesta de la red neuronal artificial, con el grupo de personas que participaron en su entrenamiento, logró un resultado óptimo del 97% a base de simulaciones y pruebas de los diferentes números, desde el 0.00 hasta el 999.99. Por otro lado, con el grupo de personas que no participaron en el entrenamiento de la red neuronal se obtuvo un resultado alrededor del 58%. Cifra que puede mejorarse si incrementamos el conjunto de personas para la etapa de entrenamiento de la red neuronal. A continuación, algunas pruebas de reconocimiento.

1ra Prueba: Número 362.94

Corresponde a un número a reproducir con una cantidad de cifras límite, es decir 3 dígitos enteros y 2 dígitos decimales; por lo tanto, las dimensiones de la matriz principal fue de 6 filas correspondiente a los valores padrones obtenidos de cada cifra y 6 columnas definidas por el número de dígitos que posee el número total a reproducir y del punto decimal.

2da Prueba: Número 92.53

Corresponde a un número de 2 dígitos enteros y 2 dígitos decimales, por lo tanto las dimensiones de la matriz principal fue de 6 filas referidas a los valores padrones y 5 columnas correspondiente a los dígitos del número a reproducir y el punto decimal.

3ra Prueba: Número 05.27

Corresponde a un número con 1 dígito entero, ya que el cero a la izquierda no tiene valor, y 2 dígitos decimales; por lo tanto, la dimensión de la matriz principal fue de 6 filas referidas a los valores padrones y 4 columnas correspondientes a los dígitos del número a reproducir y el punto decimal.

4ta Prueba: Número 0.03

Para este ejemplo, el número a reproducir consta de un digito entero y 2 dígitos decimales; por lo tanto las dimensiones de la matriz principal fue de 6 filas referidas a los valores padrones y 4 columnas correspondiente al único digito de la parte entera, el punto decimal y los 2 dígitos decimales.

CONCLUSIONES

El sintetizador de voz planteado en este trabajo de investigación, no responde a todas las muestras numéricas que conforman los números a ser sintetizados; es decir, reproducidos, debido a que la red neuronal artificial utilizada no logra reconocer todos los números, pues su entrenamiento ha sido realizado con un grupo conformado por pocas personas (universo pequeño). Por tal razón, mientras mayor

sea el número de vectores de entrada (representación matricial de las muestras numéricas) a la red neuronal para su entrenamiento o aprendizaje, mayor será el rango de reconocimiento de números; es decir, la red neuronal reconocerá números de un grupo mayor de personas, sean niños y/o adultos.

El sintetizador de voz fue diseñado para reproducir números conformados hasta por tres dígitos enteros y dos decimales, en tal sentido, si el número a sintetizar no posee el número de cifras límite, entonces el algoritmo diseñado para la reproducción automáticamente asignará a las variables correspondientes, sea en la parte de los dígitos enteros o decimales, el número cero. Con esto completaría el número de cifras límite.

La calidad del sintetizador de voz dependerá de la frecuencia de muestreo seleccionada, para la digitalización de las diferentes voces. En tal sentido es preciso indicar que a mayor frecuencia de muestreo, mayor es el número de muestras y el número de bits. En consecuencia, la digitalización de la señal analógica resulta más precisa y la calidad de la voz mejora, lo cual da como resultado un incremento del número de KBytes del archivo de voz digital.

REFERENCIAS BIBLIOGRÁFICAS

Alan V. OPPENHEIM, Alan S. WILLSKY, S Hamid NAWAB. Señales y Sistemas, 2da Edición, Prentice Hall Hispanoamericana, México, 1998.

Rafael C. GONZÁLEZ, Richard E. WOODS. Tratamiento Digital de Imágenes, Addison-Weslley Iberoamericana S.A, U.S.A, 1996.

Rafael C. GONZÁLEZ, Richard E. WOODS and Steven L. Eddins. Digital Image processing using MATLAB, Editorial Dorling Kindersley, 2004.

Simon HAYKIN. Neural Networks a comprehensive foundation, Editorial Prentice Hall, 1994.

Martín del Brío BONIFACIO, SANZ MOLINA Alfredo. Redes Neuronales y sistemas borrosos, 3ra Edición, Editorial Ra-Ma, 2006.

Luis A. HERNANDO. Introducción a la teoría y estructura del lenguaje, 2da Edición, Editorial Verbum, S.L., Madrid-España, 1995.

Matías ZAÑARTU SALAS. Aplicaciones del análisis acústico en los estudios de la voz humana [Artículo], Escuela de Fonoaudiología-Universidad Mayor, Santiago-Chile, 2003. [en línea]. Disponible en: URL:http://www.sld.cu/galerias/pdf/sitios/rehabilitacionlogo/aplicaciones_del_analisis_acustico_de_lavoz_humana.pdf.

Eduardo LLEIDA SOLANO. Conversor Texto-Voz.[Artículo]. [en línea]. Disponible en: URL:http://dihana.cps.unizar.es/investigacion/voz/ctv.html

Genoveva VELÁSQUEZ RAMÍREZ. Sistema de Reconocimiento de Voz en Matlab. Universidad de San Carlos de Guatemala-Facultad de Ingeniería, Tesis de Pre-Grado, Guatemala, 2008. Disponible en: URL:http://biblioteca.usac.edu.gt/tesis/08/08_0223_EO.pdf