Interfase de voz en sistemas robóticos para asistencia a discapacitados

Paola Lucia Tellez Ballesteros
Yu Tang
División de Estudios de Posgrado
Universidad Nacional Autónoma de México
Mexico, D.F.
tellezpl@yahoo.com.mx
tang@servidor.unam.mx

Resumen

Este trabajo presenta los resultados de comparación de distintas formas de analizar palabras aisladas para su reconocimiento por medio de una red neuronal multicapa, que formará parte de la interfase cn un sistema robótico para asistencia a discapacitados. Para tal propósito se utiliza una base de datos compuesta por ocho dígitos, obteniendo resultados específicos para tres distintos tipos de procesamiento de voz.

1 Introducción

Un robot es un manipulador mecánico, cuyo comportamiento puede ser programado. Los robots tienen cinco componentes principales:

1. Un *cerebro* controla las acciones del robot y responde a entradas sensoriales. Usualmente el cerebro es una computadora de algún tipo.
2. El *cuerpo* de un robot es el elemento físico que sostiene todas las piezas del robot.
3. Los *actuadores* que permiten el movimiento del robot. Estos son usualmente motores eléctricos aunque hay muchas otras posibilidades, como pistones hidráulicos.
4. Los *sensores* que proporciona información del ambiente al robot. Los sensores son los que permiten dotar de sentidos al robot.
5. Una $fuente\ de\ potencia$ necesaria para mover el cerebro, los actuadores y los sensores [7].

El desarrollo de sistemas de percepción en robótica surge a partir de los progresos tecnológicos de sensores tales como los de visión, tacto e, incluso, audición. Sin embargo, la percepción involucra no sólo la captación de la información sensorial, sino también su tratamiento e interpretación [9]. En el caso de la audición, se trata de llegar como última meta a una forma de reconocimiento de voz continua, con un gran vocabulario, y que sea independiente del usuario.

El problema básico es que cuando una persona habla, un sonido modifica el sonido que viene inmediatamente después. La diversidad de combinaciones de palabras en voz continua multiplica el número de psoibilidades que una computadora debe examinar para determinar lo que se dijo. Este fenómeno aumenta la cantidad de memoria requerida para realizar un reconocimiento de voz continua. Por esta razón para ayuda a discapacitados, se requiere de una aplicación más limitada y más predecible, una tarea con un número promedio reducido de elecciones de palabras que permitan seguir cada palabra en el vocabulario, de acuerdo a su probabilidad de ocurrencia.

Los sistemas de reconocimiento de voz actuales todavía están lejos del desempeño obtenido por los humanos. Los humanos reconocen voz sin esfuerzo aún en condiciones ambientales adversas. Los sistemas de reconocimiento de voz se ven considerablemente afectados por tareas tan simples como un discurso. En esta artículo, se trata de comparar el desempeño en el reconocimiento de palabras aisladas utilizando dígitos.

2 Redes neuronales multicapa

La red que se utilizó para probar los 4 tipos de procesamiento de voz, se muestra en la figura 1. En este caso se tenían 8 unidades en la capa de salida, que representan los 8 dígitos a reconocer. Se trata de una red neuronal de tres capas, en donde las unidades de la capa entrada y de la capa de salida se encuentran completamente conectadas por medio de la capa oculta (que en esta caso tenía 7 unidades). En donde cada unidad y_j de una capa inferior se encuentra conectada con una unidad z_j de una capa superior a ella por

medio de la ecuación 1

$$y_j = \sum_{i=1}^{n_k} w_{ji} z_i \quad j = 1, 2, \ldots, n_p. \tag{1}$$

Aquí, w_{ji} es el peso de conexión entre la unidad i en la capa inferior, y la unidad j en la capa superior. Esta red fue entrenada por medio del algoritmo de retropropagación.

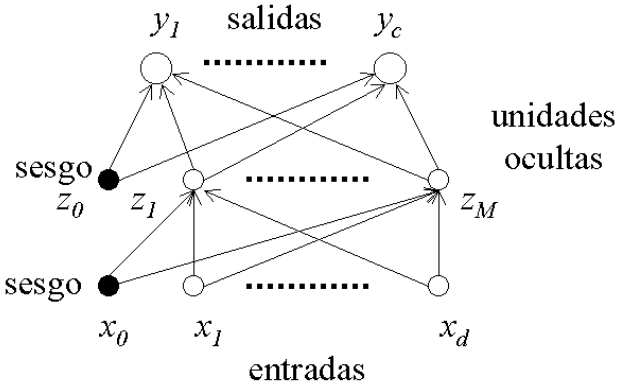

Figura 1: Red neuronal de tres capas

2.1 Algoritmo de retropropagación

Para el entrenamiento de una red neuronal del tipo que se muestra en la figura 1, se utiliza una función de error que es la diferencia entre las salidas de la red y_j, y los valores deseados correspondientes d_j.

$$E = \frac{1}{2} \sum_j (y_j - d_j)^2. \tag{2}$$

En una red común de este tipo, los valores de salida y_j son solo los niveles de activación de la salida de la red y_j. Tomando en cuenta este hecho en la definición de E, y entonces diferenciando sobre y_j, se obtiene la derivada parcial del error con respecto a la activación de las unidades de salida. La retropropagación de estos valores permiten el punto de comienzo del algoritmo de aprendizaje [10].

$$\frac{\delta E}{\delta y_j} = y_j - d_j \tag{3}$$

En la red actualizada, cada valor de salida de la red es la suma de los cuadrados de las activaciones de distintas actualizaciones temporales de una unidad de salida.

$$o_j = \sum_t y_{jt}^2 \tag{4}$$

Insertando la ecuación 4 en la ecuación 2, y diferenciando el error con respecto a la activación de la actualización de la unidad j, en el tiempo τ, se obtiene

$$\frac{\delta E}{\delta y_{j\tau}} = 2 y_{j\tau} \big(\big(\sum_t y_{j\tau}^2 \big) - d_j \big) \tag{5}$$

Este algoritmo se puede detener de acuerdo al esquema 2.2 ó de acuerdo al esquema 2.3 que se presentaran más adelante. Para aplicar ambos esquemas se divide la base de datos en dos conjuntos: el conjunto de entrenamiento que constituye el 80 % de la base de datos, y el conjunto de prueba que constituye el 20 % restante.

2.2 Pronto acabado

El entrenamiento de una red multicapa se realiza de una manera iterativa reducinedo la función de error definida con respecto al conjunto de entrenamiento. Durante una sesión de entrenamiento normal, este error disminuye como una función del número de iteraciones en el algoritmo. Sin embargo, el error con respecto al conjunto de prueba al principio disminuye, pero después comienza a aumentar, como muestra la figura 2 casi al final de la época 500, indicando que los datos se están sobremuestreando.

Por medio del esquema de *pronto acabado*, el entrenamiento se detiene en la época τ en la que el conjunto de prueba presenta un error mínimo, ya que es en este punto donde se espera tener la mejor gene-ralización posible. Para no cometer errores al observar este punto, debe observarse la tendencia de los datos, para ello, pueden utilizarse métodos de suavizamiento como el spline, que puede calcularse cada cierto número de épocas para ver sí el error en los datos tiende a aumentar. Sí esto sucede, se guardan los pesos de la red, ya que es ahí donde se espera tener el mejor desempeño.

En este artículo se evaluó la tendencia en el error del conjunto de prueba cada 100 épocas. Se utilizó el método de spline para evaluar la tendencia de los datos [1], [2], [3], [4], [5] . Este comportamiento se explica en términos del número efectivo de grados de libertad en la red. Este número comienza siendo pequeño, y entonces crece durante el proceso de entrenamiento, correspondiendo al incremento efectivo de la complejidad del modelo. Detener el entrenamiento antes de que se alcance un mínimo en el error de entrenamiento representa una forma de limitar la complejidad de la red [6].

2.3 Aprendizaje japonés

Si después de cada actualización de los pesos en una época de entrenamiento normal, se evalúa el porcentaje de acierto que se tiene con los datos de

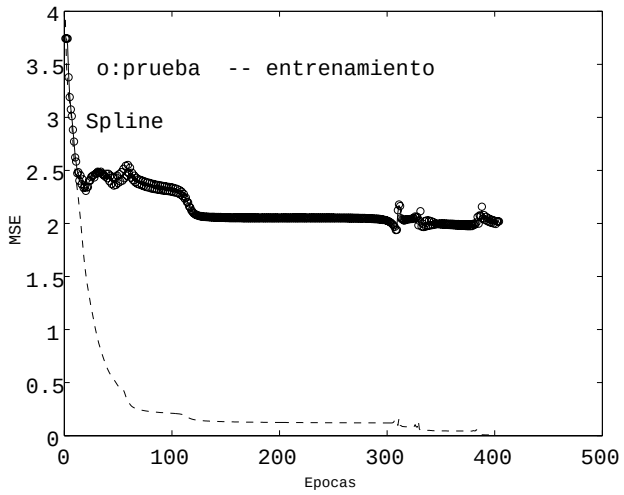

Figura 2: Red neuronal de dos capas

prueba y el porcentaje de acierto que se tiene con los datos de entrenamiento se puede obtener otra medida de aprendizaje de la red. Con respecto al porcentaje obtenido con los datos de prueba se observa que este resultado al principio aumenta mucho, pero después tiene a disminuir y estabilizarse a un cierto nivel como muestra la figura 3.

Por medio de este esquema, (utilizado en Japón), el aprendizaje se detiene cuando se alcanza el máximo en el grado de acierto obtenido con el conjunto de prueba. Para observar este punto también se utilizó el spline, evaluandolo cada 100 épocas.

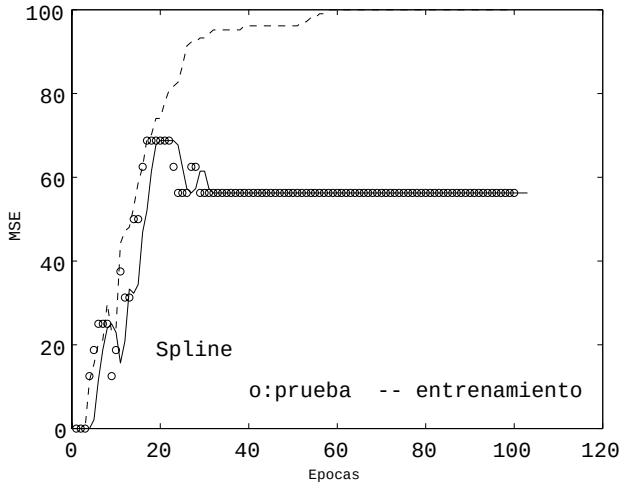

Figura 3: Red neuronal de dos capas

3 Simulación

Para esta investigación, se utilizó una base de datos de 8 dígitos, cada uno de los cuales fue repetido 15 veces. Esta base de datos se muestreo a una frecuencia de 11025 Hz. Para analizar la base de datos, se utilizaron distintos tipos de análisis de voz. Algunas veces se utilizaron escalas auditivas después del análisis de voz.

3.1 Análisis de dígitos

El análisis realizado incluyó disntintos tipos de procesamiento. Uno de ellos fue la utilización de coeficientes LPC, otro fue la transformada rápida de Fourier FFT para obtener la frecuencia de la señal de voz, ó el logaritmo de la FFT para obtener un valor normalizado de la frecuencia.

El análisis con coeficientes LPC considera que una señal de voz S_n se puede representar por una combinación lineal de valores pasados de la misma señal, modelando el tracto vocal del ser humano como:

$$\widehat{S}_n = -\sum_{k=1}^{p} a_k \cdot Sn - k \qquad (6)$$

Donde p es el orden de predicción lineal, a_k es el k-ésimo coeficiente de predicción lineal. Este modelo corresponde a un filtro auto-regresivo. En este caso, el orden utilizado fue de 16. El análisis realizado con estos coeficientes se puede observar en el cuadro 1. La escala auditiva utilizada se explica en la siguiente sección.

Tabla 1: Análisis de voz LPC

Tasa de muestreo	11.025 kHz
Ventana	Blackman
Ancho de la ventana	10 ms
Desplazamiento de ventana	5 ms
Análisis	LPC de 16o. orden
Escala auditiva	MEL

En el caso de la FFT, se utilizó la escala auditiva Mel, que representa la escala de frecuencia del oído humano. Esta escala se modeló utilizando 5 ventanas triangulares, que pueden observarse en la figura 4.

Así, el tercer análisis de voz utilizó la FFT con la escala auditiva MEL obteniéndose como resultado a los coeficientes SPE. Por último, el cuarto análisis utilizó la FFT normalizada con el logaritmo junto con la escala MEL obteniéndose los coeficientes CEP, como muestran los cuadros 2 y 3.

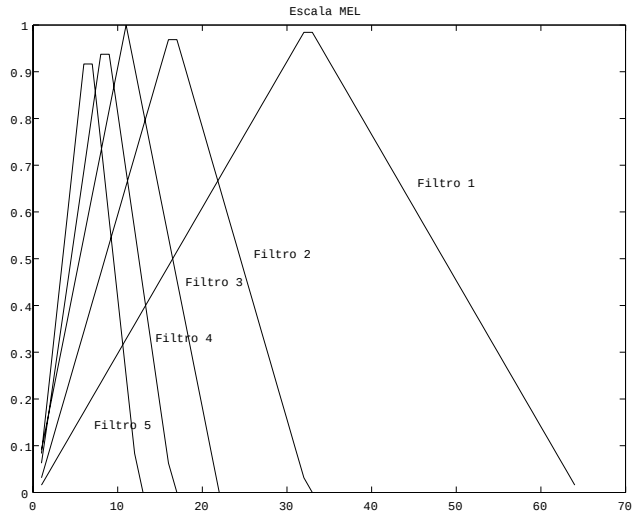

Figura 4: Ventanas Mel

Tabla 2: Análisis de voz SPE

Tasa de muestreo	11.025 kHz
Ventana	Blackman
Ancho de la ventana	10 ms
Desplazamiento de ventana	5 ms
Análisis	FFT
Escala auditiva	MEL

Tabla 3: Análisis de voz CEP

Tasa de muestreo	11.025 kHz
Ventana	Blackman
Ancho de la ventana	10 ms
Desplazamiento de ventana	5 ms
Análisis	Logaritmo FFT
Escala auditiva	MEL

Tabla 4: Reconocimiento de 8 dígitos utilizando pronto acabado con coeficientes LPC

Tipo de coeficiente	Número de épocas	Reconocimiento total
LPC1	100	79.32 %
LPC2	100	82.45 %
LPC3	500	90.62 %
LPC4	100	78.12 %
LPC5	100	84.37 %

4 Reconocimiento de voz con dígitos

Todos los primeros coeficientes obtenidos con el análisis de cada palabra se agruparon en un solo archivo, lo mismo se hizo con los siguientes coeficientes. Los archivos resultantes para cada uno de los análisis mencionados sirvieron para entrenar una red neuronal multicapa. Los cuadros 4, 5, 6, 7, 8, y 9 muestran los resultados obtenidos con los distintos tipos de procesamiento de voz.

En cada uno de los cuadros, los porcentajes de reconocimiento total forman un promedio obtenido con el reconocimiento de los datos de prueba, y con el reconocimiento de los datos de entrenamiento. La tabla también muestra la época donde se detuvo el entrenamiento de acuerdo a cada esquema.

Tabla 5: Reconocimiento de 8 dígitos utilizando aprendizaje japonés con coeficientes LPC

Tipo de coeficiente	Número de épocas	Reconocimiento total
LPC1	500	79.32 %
LPC2	500	82.45 %
LPC3	100	85.57 %
LPC4	100	78.12 %
LPC5	491	84.37 %

Tabla 6: Reconocimiento de 8 dígitos utilizando pronto acabado con coeficientes SPE

Tipo de coeficiente	Número de épocas	Reconocimiento total
SPE1	100	69.23 %
SPE2	100	80.04 %
SPE3	200	78.84 %
SPE4	500	81.49 %
SPE5	400	75.00 %

Tabla 7: Reconocimiento de 8 dígitos utilizando aprendizaje japonés con coeficientes SPE

Tipo de coeficiente	Número de épocas	Reconocimiento total
SPE1	100	69.23 %
SPE2	100	80.04 %
SPE3	100	73.79 %
SPE4	100	82.69 %
SPE5	100	75.72 %

Tabla 8: Reconocimiento de 8 dígitos utilizando pronto acabado con coeficientes CEP

Tipo de coeficiente	Número de épocas	Reconocimiento total
CEP1	100	87.50 %
CEP2	100	87.50 %
CEP3	497	90.62 %
CEP4	385	93.75 %
CEP5	300	93.75 %

Tabla 9: Reconocimiento de 8 dígitos utilizando aprendizaje japonés con coeficientes CEP

Tipo de coeficiente	Número de épocas	Reconocimiento total
CEP1	100	87.50 %
CEP2	100	85.50 %
CEP3	100	90.62 %
CEP4	385	93.75 %
CEP5	454	93.75 %

5 Conclusiones

Por medio de este trabajo se puede observar que los mejores resultados de reconocimiento se pueden obtener con un análisis normalizado a través del logaritmo de la FFT, y con la utilización de la escala auditiva MEL, como muestran los cuadros 8, y 9. En estos cuadros, se observa que aún con un distinto numero de épocas en cada caso, se alcanza el mismo porcentaje de reconocimiento usando aprendizaje japonés ó usando pronto acabado.. De esta forma se demuestra que al modelar el reconocimiento de acuerdo al funcionamiento del oído humano se pueden obtener mejores resultados de reconocimiento de voz utilizando dígitos como un ejemplo.

Referencias

[1] Simonoff, J. S. 1996. Smoothing Methods in Statistics. New York. Springer Verlag.

[2] Wand, M.P. and M.C. Jones. 1995. Kernel Smoothing. London. Chapman & Hall.

[3] Bowman, A. W. and A. Azalini. 1997. Applied smoothing techniques for data analysis: The kernel approach with S-PLUS. Ilustrations. Oxford.

[4] Green P.J. and B. W. Silverman. 1994. Nonparametric REgression and Generalizaed Linear Models. A Roughness Penalty Approach. Chapman & Hall.

[5] Scott, David W. 1992. Multivariate Density Estimation. Theory, Practice and Visualization. New York. John Wiley & Sons.

[6] Bishop, Christopher M. Neural Networks for Pattern Recognition. Oxford University Press, 2000.

[7] Knudsen, Jonathan B. The unofficial guide to LEGO MINDSTORMS Robots. O rreilly, 1999.

[8] Nakano Miyatake, Mariko. Sistema de reconocimiento de voz con vocabulario limitado para el control de dispositivos periféricos. IEEE ROC 2001. 12a. reunión de otoño de comunicaciones, computación, electrónica y exposición industrial.

[9] Ollerto Baturne, Aníbal. Robótica, manipuladores y robots móviles. Alfaomega marcombo, 2001.

[10] Waibel Alex H., Lang Kevin. A time delay neuroal network architecture for isolated word recognition. *Neural Networks*, vol. 3. 1990.