IMPLEMENTACIÓN DE UN RECONOCEDOR DE PALABRAS AISLADAS DEPENDIENTE DEL LOCUTOR

César San Martín S.[1] **Roberto Carrillo A.**[1]

Recibido el 14 de julio de 2003, aceptado el 3 de mayo de 2004

RESUMEN

En este trabajo se presenta un sistema de reconocimiento de palabras aisladas dependiente del locutor. Cada palabra se codifica mediante las técnicas de Predicción Lineal y Cepstrum real, mientras que la etapa de clasificación se realiza mediante el alineamiento temporal dinámico, que permite independencia del intervalo de tiempo de cada muestra de voz. Los resultados obtenidos demuestran que el uso de estas técnicas permiten obtener un 85% de clasificación correcta.

Palabras claves: Reconocimiento de voz, análisis Cepstral, extracción de parámetros, patrones de voz.

ABSTRACT

In this work a speaker dependent isolated word recognition system is presented. While each word is encoded using the Linear Prediction and Cepstrum techniques, the classification stage is carried out by means of Dynamic Time-Warping, which allows independence for the interval of time of each voice sample. The results obtained show that using these techniques allows about 85% of correct classification.

Keywords: Speech recognition, Cepstral analysis, extraction of parameters, speech patterns.

INTRODUCCIÓN

La creciente necesidad de mejorar la comunicación hombre – máquina, ha inducido a la técnica del reconocimiento automático del habla (RAH) hacia un inusitado interés, tanto en empresas tecnológicas como en Universidades. Basta nombrar una serie de nuevos productos de control por voz para comprender la creciente necesidad: robot industriales o electrodomésticos, sistemas de ayuda a discapacitados, acceso y navegación por base de datos, operaciones y transacciones comerciales, control de acceso, operaciones telefónicas automáticas, etc.

En términos generales, los sistemas RAH, son la implementación algorítmica de un análisis detallado de las características del habla [3], y que dependerá del lenguaje usado [5] y de otros conceptos que no deben excluirse (frecuencia fundamental, energía del tramo de señal, etc.). Un sistema RAH puede contemplar: Reconocimiento de palabras aisladas, identificación de palabras clave en discurso continuo, reconocimiento de palabras conectadas y reconocimiento de discurso continuo. Cada una de estas con dependencia o independencia del locutor.

Los sistemas de reconocimiento dependiente del hablante, deben ser entrenados para responder a las características particulares de la voz de una persona en particular, es decir, para un solo locutor. Esto restringe el propósito general de todo sistema RHA, pero puede ser perfectamente factible para situaciones como las mencionadas en el párrafo inicial. Además, el primer paso antes de implementar un sistema RAH independiente del locutor, es investigar sobre la dependencia del hablante con el sistema, lo cual, permite cimentar la base teórico práctica para la implementación posterior.

Esta publicación, da a conocer nuestra experiencia en el desarrollo de un sistema reconocedor de palabras aisladas dependiente del locutor. El sistema fue desarrollado para posteriormente ser utilizado en el desarrollo e implementación de un sistema para mejorar el proceso enseñanza aprendizaje del habla en niños con déficit auditivo. La implementación de nuestro reconocedor, consta de tres etapas bien

[1] Universidad de La Frontera, Depto. de Ingeniería Eléctrica, Casilla 54-D, Temuco – Chile, csmarti@ufro.cl, rcarrill@ufro.cl

definidas: Etapa de muestreo y cuantificación de la señal de voz, pre-procesamiento o proceso de adaptación de las muestras y la etapa final de reconocimiento. El interés principal de nuestra investigación lo centramos en una correcta implementación de los algoritmos clásicos de extracción de características y reconocimiento de palabras. Específicamente: Para la extracción de características de voz se utilizó la técnica de codificación por predicción lineal (LP) [1] y análisis cepstral [2], [8]; correspondiendo esta última a un modelo matemático de extracción de patrones por compresión de señal. La etapa de reconocimiento de palabras, la implementamos usando el algoritmo de alineamiento temporal dinámico (DTW) [6], [7]; el cual, es capaz de discriminar entre palabras con duración temporal independientes de la señal.

En [7] se utilizó nuestra metodología, los resultados obtenidos concuerdan plenamente. Lo cual, demuestra que para diferentes acentos dentro de una misma lengua, los algoritmos siguen siendo efectivos. Además, queda demostrado que para un reducido vocabulario, la técnica descrita en este artículo, es la más adecuada y la más económica para reducir el costo computacional. Otros autores se han dedicado a estudiar el reconocimiento de palabras aisladas en presencia de ruido. Para esta situación, en [10] se muestran los resultados basados en un reconocedor desarrollado bajo la técnica MFCC (Mel Frequency Cepstral Coefficient) [1]. Otros trabajos se han orientado al reconocimiento de grandes vocabularios, en ellos se observa que al aplicar el método descrito, la taza de reconocimiento disminuye drásticamente siendo lo más efectivo utilizar técnicas de análisis estocástico. Los resultados de mayor éxito hasta ahora se han logrado utilizando Modelos Ocultos de Markov (HMM). En [9] se logró una taza de 98% de reconocimiento, lo cual prácticamente tendría resuelto el problema.

MODELO DE PRODUCCIÓN DE LA VOZ (MODELO SOURCE-FILTER)

El modelo clásico del tracto vocal [3] se compone de un filtro variable en el tiempo, un generador de ruido aleatorio y de un generador de impulsos (Fig. 1). Los parámetros del filtro varían en función de la acción consciente que se realiza al pronunciar una palabra. El modelo tiene dos entradas, que dependen del tipo de señal. Para señales sonoras (vocales) la excitación es un tren de impulsos de frecuencia controlada, mientras que para las señales no sonoras (consonantes) la excitación es ruido aleatorio. La combinación de estas dos señales modelan el funcionamiento de la glotis. Es importante destacar que la mayor parte de la información del locutor está en las cuerdas vocales (o excitación), mientras que la información de la palabra pronunciada está en la característica del filtro.

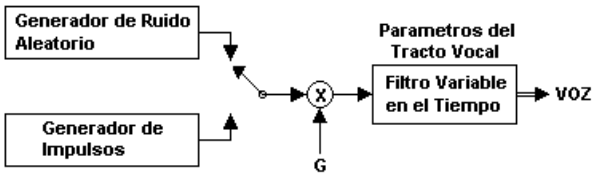

Fig. 1.- Modelo de Producción de Voz

El espectro en frecuencias de la señal vocal se obtiene como el producto de la transformada de Fourier (FT) de la excitación por la respuesta en frecuencia del filtro [3], es decir:

$$S(\omega) = E(\omega) \cdot H(\omega) \tag{1}$$

Donde:

$E(\omega)$: FT de la excitación.

$H(\omega)$: FT del Filtro.

$S(\omega)$: FT de la voz.

Análisis Cepstral

Se define el CEPSTRUM (CS) real de una señal como la transformada inversa de Fourier (IFT) del módulo del espectro en escala logarítmica (en belios) de esa señal [2], es decir:

$$c(t) = F^{-1}\left[\log\left(S(\omega)\right)\right] \tag{2}$$

Desarrollando el CS real para $S(\omega)$ se tiene:

$$c(t) = F^{-1}\left[\log\left|E(\omega)\right|\right] + \left[\log\left|H(\omega)\right|\right] \tag{3}$$
$$c(t) = c_e(t) + c_h(t)$$

De la ecuación (3) se concluye que el CS de una señal es la suma del CS de la excitación y el CS del filtro (precisamente la respuesta impulsiva del filtro). Las *bajas componentes cepstrales* corresponden a variaciones lentas de las componentes espectrales y por tanto contienen información de la *envolvente del espectro*, la cual se relaciona con la respuesta en frecuencia del filtro que modela el tracto vocal. En el caso de reconocimiento de patrones de voz, normalmente lo que nos interesa no son las características de la excitación, sino las *características*

del tracto vocal [4], por lo que se usan las bajas componentes cepstrales para reconocer voz (no locutores). Para el reconocimiento de voz es usual considerar 10 a 12 coeficientes (10 ó 12 primeros) obtenidos sobre una ventana temporal de unos 20 ó 30 milisegundos de duración [1]. Sin embargo, es preferible añadir algunas características adicionales como una medida de la energía de la señal en la ventana de tiempo en cuestión y una estimación del pitch o frecuencia fundamental [3], [5].

Predicción Lineal

La Predicción Lineal (LP), en el reconocimiento de voz, consiste en modelar el tracto vocal como un filtro digital constituido únicamente por polos (respuesta a impulso infinita o IIR) permitiendo así calcular la próxima muestra como una suma ponderada de las muestras pasadas (Fig. 2). Este filtro de predicción [3] se traduce en la función de transferencia de la ecuación (4):

$$H(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i \cdot z^{-i}} \qquad (4)$$

donde G es la ganancia del filtro, que depende de la naturaleza de la señal (sonora o no sonora). Entonces, dada la señal $s(n)$, el problema consistirá en determinar los coeficientes de predicción y la ganancia.

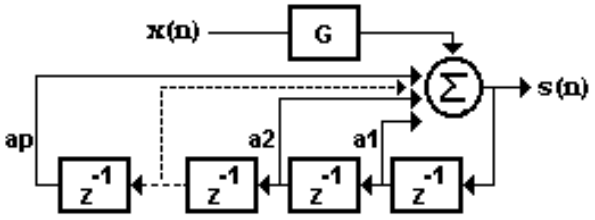

Fig. 2.- Filtro de Predicción

Entonces, serán los coeficientes de predicción los que se usarán como parámetros de reconocimiento de palabras. Estos se determinan minimizando el error que se comete cuando se intenta realizar la aproximación de la señal. Esto es:

$$e(n) = s(n) - s_p(n) \qquad (5)$$

Donde:

$e(n)$: Error de predicción.

$s(n)$: Señal de voz.

$s_p(n)$: Señal de voz predicha.

Además, el error de predicción puede escribirse como [1]:

$$e(n) = s(n) - \sum_{k=1}^{p} a_k \cdot s(n-k) \qquad (6)$$

y puede minimizarse obteniendo para los parámetros la relación matricial (7):

$$\begin{bmatrix} r_1 \\ r_2 \\ \dots \\ r_p \end{bmatrix} = \begin{bmatrix} r_0 & r_1 & \dots & r_{p-1} \\ r_1 & r_0 & \dots & r_{p-2} \\ \dots & \dots & \dots & \dots \\ r_{p-1} & r_{p-1} & \dots & r_0 \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix} \qquad (7)$$

Despejando se obtiene la ecuación (8):

$$A = R_{Toeplitz}^{-1} \cdot R_{autocorrelación} \qquad (8)$$

En las ecuaciones (7) y (8), se observa que el vector [r_1 r_2 ... r_p] corresponde a los p primeros coeficientes de autocorrelación de un segmento de señal. A la matriz que multiplica al vector de coeficientes se le denomina *Matriz de Toeplitz*. Por lo tanto, calculando los primeros p coeficientes de autocorrelación de un segmento de señal podemos generar la matriz de Toeplitz correspondiente. Con esto, sólo queda calcular su inversa y multiplicarla por el vector de autocorrelación para obtener los coeficientes del filtro en la ventana de análisis.

La función de autocorrelación proporciona una medida de la relación de la señal con una copia desfasada de sí misma. Se va a definir a p como el orden de análisis. Valores típicos de p pueden ser entre 10 y 15, lo que significa que por cada ventana de señal se extraerán entre 10 y 15 coeficientes, proceso denominado codificación por LP (LPC) [1].

Para la extracción de patrones de voz realizado en la implementación práctica, se considera el cálculo de los coeficientes LPC-CS por ventana de señal. La idea es, que para cada segmento de voz, se extraigan los primeros 12 coeficientes LPC [4]. Luego, se desarrolla el cálculo cepstral para los coeficientes LPC anteriormente obtenidos. Así, la mezcla de estas dos técnicas parametriza de mejor manera la señal vocal, puesto que se consideran las características de la excitación propiamente tal (LPC) y las del tracto vocal (CS) correspondientes con el modelo en cuestión.

DESCRIPCIÓN DEL SISTEMA RHA

Implementación

Mediante la utilización de un micrófono multimedia se implementó la adquisición de señales de voz, para su posterior procesamiento y análisis de datos en computadora personal (PC). Para la adquisición de señal se utiliza la tarjeta de sonido de una computadora y la interfaz gráfica de usuario se desarrolló en LabVIEW de NATIONAL INSTRUMENTS, versión de evaluación.

Se debe reiterar la importancia de evaluar cada uno de los algoritmos en un análisis OFF-LINE, esto es, desarrollar las etapas de reconocimiento en un software de tratamiento de señales sencillo. Matlab de MATHWORKS, posee las características necesarias para este efecto. El correcto funcionamiento de los algoritmos en Matlab permitirá una evolución más simple y sencilla del sistema a un software de procesamiento de señales más complejo para su operación ON-LINE.

Descripción del Sistema

Los puntos básicos a implementar son los que se muestran en la Fig. 3. El desarrollo paso a paso de cada una de las etapas observadas en el esquema realiza la correcta implementación del Reconocedor.

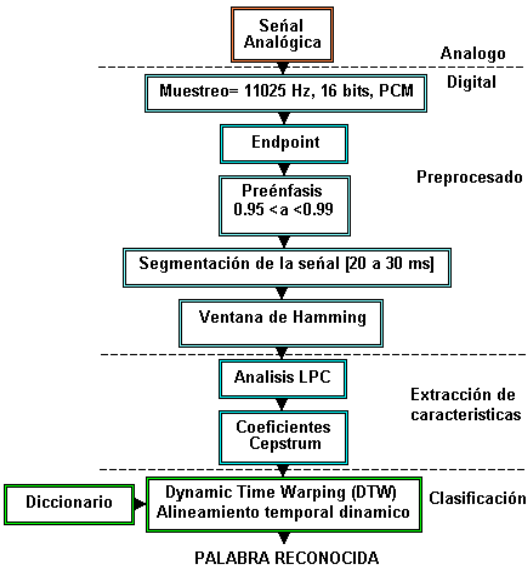

Fig. 3.- RHA basado en DTW

Adquisición, Cuantificación y Muestreo

El primer paso a desarrollar es la *adquisición, cuantificación* y *muestreo* de la señal. Para ello, LabVIEW posee librerías especialmente diseñadas para el trabajo con señales de audio. Con un simple diagrama en bloques es posible configurar esta tarea para operación de adquisición on-line de la señal de voz. Para efectos del trabajo, la adquisición se desarrolló con una frecuencia de muestreo f_s = 11025 Hz, una cuantificación de 16 bits y calidad de sonido *mono estéreo*.

Endpoint Detection

Una vez realizada la adquisición, es necesario implementar un algoritmo detector de comienzo y fin de punto de la señal con el objetivo de eliminar información redundante de entrada al sistema (eliminar silencios al comienzo y al final de la señal de voz). Esto se realiza a través de un cálculo de energía y detección de umbrales de actividad de señal.

Preénfasis

El *preénfasis* permite acentuar las frecuencias altas de la señal de voz, esto debido a que el modelo del tracto vocal utilizado no filtra de buena manera las señales de frecuencias altas (*no sonoras*, por ejemplo: consonantes, "s"), a diferencia de las de frecuencias bajas (*sonoras*, por ejemplo: vocales "a"). Este filtro de preénfasis obedece a la ecuación en diferencia dada en la ecuación (9) y cuya función de transferencia asociada es la ecuación (10), donde $v(n)$ es la señal de voz de entrada y $s(n)$ la señal filtrada.

$$s(n) = v(n) - a \cdot s(n-1) \tag{9}$$

$$H(z) = 1 - a \cdot z^{-1} \tag{10}$$

Segmentación-Enventanado de Hamming

Antes de entrar a la etapa de extracción de características, la señal de voz debe ser segmentada a intervalos de 20 a 30 ms, tiempo durante el cual la señal se considera cuasi estacionaria. Luego, es sometida a una ventana de Hamming [2] con el objeto de suavizar la señal en los bordes de dicha ventana, que generalmente se usa para el análisis de señales de voz, y se define como en la ecuación (11).

$$W_n = \begin{cases} 0.54 - 0.46 \cdot \left(\frac{2\pi n}{N}\right) & 0 \le n \le N \\ 0 & \text{otro caso} \end{cases} \qquad (11)$$

Extracción de Patrones de Voz

Usando la técnica LPC y CS real, es posible parametrizar la señal con un número pequeño de patrones con los cuales se reconstruye adecuadamente. A través de esta técnica podemos representar a la señal de voz mediante parámetros que varían en el tiempo los cuales están relacionados con la función de transferencia del tracto vocal y las características de la fuente sonora. Otra ventaja es que no requiere demasiado tiempo de procesamiento, lo cual es muy importante a la hora de la implementación. El bloque en LabVIEW para la implementación de este punto se muestra en la Fig. 4. En este diagrama, la señal de voz de entrada se identifica por *X*, F_s corresponde a la frecuencia de muestreo, que para efectos prácticos es de 11025 Hz y *n_lpcc* es el orden *p* de análisis. En la salida se tiene una matriz denominada *getceps(X)*, cuyas filas contienen los patrones o características LPC-CEPSTRUM correspondiente a cada segmento de voz.

Fig. 4.- Extractor de parámetros de voz

Clasificación de Patrones de Voz

En esta etapa la computadora debe ser capaz de discriminar entre los diferentes sonidos de entrada realizando el reconocimiento de cada una de éstas con respecto a un diccionario previamente establecido. Este diccionario suele denominarse libro de códigos (codebooks). Por las características que posee la señal de voz, se debe hacer uso de un algoritmo que permita la comparación de dos patrones, independientemente de sus duraciones temporales con que fueron pronunciadas. El algoritmo utilizado para este propósito es el Alineamiento Temporal Dinámico (DTW) [6], [7].

Fig. 5.- Bloque que aplica el DTW

El diagrama de bloque implementado en LabVIEW para realizar esta tarea se observa en la Fig. 5. En ella se tienen las entradas *unknown* (desconocido) y *template* (plantilla). La primera de ellas corresponde a la matriz de patrones calculados para la palabra de entrada al sistema (señal a reconocer), y la segunda a las matrices de características de cada una de las palabras incluidas en el diccionario de códigos. La salida *score* contiene el resultado de la comparación de la palabra a reconocer y las pertenecientes al libro de códigos (palabra reconocida).

CONCLUSIONES

El sistema de reconocimiento fue sometido a prueba para un conjunto de 10 palabras. Se escogió a modo de ejemplo un vocabulario compuesto por los dígitos (números del 1 al 10). Los experimentos se llevaron a cabo pronunciando 20 veces cada una de las palabras y anotando los aciertos y desaciertos, obteniéndose un 85% de clasificación correcta (Tabla 1). Cabe destacar que las pruebas se realizaron por un solo locutor y en condiciones de ausencia de ruido de fondo.

Los resultados obtenidos para el sistema de reconocimiento no han sido del todo satisfactorios (85%). La taza de reconocimiento podría aumentarse mejorando las condiciones para la adquisición de las palabras de referencia, las cuales deberían ser lo más óptimas posible.

Tabla 1.- Matriz de confusión para el sistema RHA

n° / Pal	1	2	3	4	5	6	7	8	9	10
Uno	19	1	–	–	–	–	–	–	–	–
Dos	–	15	–	–	–	2	–	–	–	3
Tres	–	2	18	–	–	–	–	–	–	–
Cuatro	–	–	–	19	–	–	–	1	–	–
Cinco	–	–	–	–	18	–	–	–	–	–
Seis	2	–	–	–	–	18	–	–	–	–
Siete	–	–	–	–	–	3	17	–	–	–
Ocho	–	–	–	–	–	–	–	17	–	–
Nueve	1	–	–	–	–	–	–	–	19	–
Diez	–	5	3	–	–	2	–	–	–	10

La técnica DTW da resultados hasta cierto punto aceptables, pero se presenta un problema dado que requiere más procesamiento que las técnicas empleadas en los sistemas independientes del locutor (Markov y redes neuronales) [5], [9]. A pesar de lo anterior, el trabajo nos entrega un primer paso hacia el reconocimiento de palabras aisladas en español. Así, las futuras investigaciones deben orientarse a sistemas de multi - locutor los cuales deben aplicar los métodos ya mencionados. Para el éxito de esas investigaciones, será fundamental la creación de una base de datos para el lenguaje natural de la región donde se implemente, sólo así, el entrenamiento de los modelos acústicos (basados en parámetros estadísticos), será más fiable.

REFERENCIAS

[1] J. Proakis, Ch.D.G. Manolakis; "Tratamiento Digital de Señales", Prentice - Hall, 1998.

[2] A. Oppenheim, R. Schafer; "Discrete-Time Signal Processing", Prentice-Hall, USA., 1989.

[3] L. Rabiner, B.H. Juang; "Fundamentals of Speech Recognition", Prentice-Hall, USA., 1993.

[4] A. Procházka, J. Uhlír and P. Sovka; "Signal Analysis and Prediction I", Procházka et al, Prague, Czech Republic, 1998.

[5] Llamas Bello, Cardeñoso Payo; "Reconocimiento Automático del Habla. Técnicas y aplicaciones", Publicaciones de la Universidad de Valladolid. España, 1997.

[6] Stuart N. Wrigley; "Speech Recognition By Dynamic Time Warping (DTW)". Disponible en: http://www.dcs.shef.ac.uk/, 1999.

[7] Andrés Flores Espinoza; "Reconocimiento de Palabras aisladas". Disponible en: http://www.alek.pucp.%20edu.pe/~dflores/INDEX.html

[8] M.S. Zilovich, R.P. Ramachandram; "Speaker Identification Based in the use of Robust Ceptral Features Obtained fron Pole-Zero Transfer Functions", IEEE Transactions on Speech and audio Processing, 6, 3, 1998. pp. 260-267.

[9] K. Montri, S. Zahorian; "Signal Modeling for High-Performance Robust Isolated Word Recognition". IEEE Transactions on Speech and audio Processing. Vol-9, No 6, September 2001.

[10] A. Adid, J.P. Barjaktarevic, O. Ozun, M. Smith and P. Steurer; "Automatic Speech Recognition for Isolated Words". Departament of electrical Engineering. Universite of California, Los Angeles. U.S.A. Disponible en: http://www.ee.ucla.edu/~psteurer/projects/ee214b projectreport.pdf