# Sistema De Reconocimiento De Palabras Aisladas Dependiente Del Hablante

***Trabajo práctico final de "Procesamiento Digital de Señales"***

Nelson Gerardo Peltzer, Nelson Facundo Lezcano

Carrera de Ingeniería Informática. Facultad de Ingeniería y Ciencias Hídricas(FICH). Universidad Nacional del Litoral. Ciudad Universitaria. Paraje "El Pozo". S3000ADQ. Santa Fe. Argentina.
{nelson.peltzer, facundolezcano}@gmail.com

**Resumen.** Este trabajo tiene como fin presentar un sistema de reconocimiento de voz de palabras aisladas dependientes del hablante dentro del contexto de la entrega de un trabajo práctico final de carácter académico. La idea general se basa en comparar patrones de una señal de voz emitida por un hablante con patrones de voz de un conjunto de palabras o comandos que conforman el diccionario y decidir si la palabra pronunciada es una de las palabras contenidas en este último.
Para la extracción de dichos patrones característicos de las señales de voz se ha recurrido a tres técnicas, estamos hablando del método de Predicción Lineal, Coeficientes Cepstrales, y Coeficientes Cepstrales en escala de Mel.
El núcleo de comparación e identificación sobre el que se apoya el sistema es el algoritmo de alineamiento temporal (Dynamic Time Warping).
Las grabaciones de las muestras de voz para las palabras que corresponden al diccionario de comandos fueron realizadas en ambientes liberados de ruido. Así también el funcionamiento del sistema sólo esta asegurado para su ejecución en ambientes silenciosos.
Durante las pruebas realizadas se obtuvo que para un número reducido de palabras del diccionario, y realizando ajustes de la tolerancia en la estrategia de decisiones, la efectividad conseguida fue del 100%.

**Palabras Clave.** Reconocimiento de voz, Predicción lineal, Coeficientes Cepstrales, Coeficientes Cepstrales en escala de Mel, Dynamic Time Warping.

## 1 Introducción

Los sistemas de reconocimiento automático de voz (RAH) han supuesto un paso adelante en las interacciones de los humanos con las máquinas.

El habla es la forma mas natural de la expresión y comunicación humana, es por ello que la posibilidad de comandar e interactuar con máquinas y dispositivos a través del uso de lenguaje ha permitido en ciertos aspectos solucionar muchos problemas que residen en la complejidad inherentes a la comprensión de las interfaces que permiten el control y manejo de diversos sistemas, así como también a las capacidades tanto físicas como intelectuales con las que se ha de contar. Por citar ejemplos pensemos en las aplicaciones software que usan sistemas RAH para permitir a personas no videntes poder operar un computador, o más aún, personas cuyas discapacidades físicas les impiden el uso de interfaces tangibles, tal como puede ser un mouse o un teclado.

Mirando a nuestro alrededor podemos ver que estos sistemas en cuestión se encuentran aplicados en muchos ámbitos; son ejemplos: electrodomésticos, sistemas de seguridad, robots, sistemas para discapacitados, etc.

Los sistemas RAH se definen y clasifican según el problema al cual es aplicable. Así existen sistemas de tipo independiente o dependiente del hablante y por el otro lado se encuentra la capacidad de reconocimiento del habla continua o solo palabras aisladas. Estas características tienen consecuencias en distintas teorías para el desarrollo e implementación de los mismos.

Para el caso de los sistemas independientes del hablante y el reconocimiento del habla continuo, en la actualidad se usan métodos como el HMM (Hidden Markov Models) y otros basados en ANN (Redes Neuronales). Estos han dado muy buenos resultados y son los más usados al día de hoy.

En el presente trabajo se muestran los fundamentos para el diseño de un sistema RAH dependiente del hablante y con reconocimiento de palabras aisladas. Estos tipos de sistemas necesitan obligatoriamente ser entrenados por la persona quien hará uso de la aplicación basada en dicho sistema.

Las fases de implementación son básicamente tres: el preénfasis, que es la etapa en la cual se encarga del preprocesado de las muestras; la extracción de parámetros, los cuales son un conjunto de valores propios de una señal de voz, y que son representativos de esta; El reconocimiento, que provee la regla para la decisiones en el proceso de identificación.

Para la extracción de parámetros se emplearon tres de los métodos más populares con el fin de establecer una comparación de resultados y eficiencia entre ellos. Estos son la técnica de Preedición Lineal, de Coeficientes Cepstrales, y Coeficientes Cepstrales en escala de Mel.

Para la fase de reconocimiento, se hace uso del algoritmo de alineamiento temporal (Dynamic Time Warping).

Además en este informe se desea explicar de forma breve pero concisa la manera en que se utilizan los métodos nombrados en cada etapa, así como también dar una somera idea de las bases de dichos métodos.

Finalmente en las conclusiones se señalan los puntos fuertes o flaquezas que tiene el sistema RAH estudiado y se reflexiona sobre el contexto en el que puede ser aplicado útilmente, así como también las alternativas que se nos presentan según el problema a solucionar.

## 2 Métodos y materiales

En la implementación del sistema RAH se contó con la ayuda de un micrófono profesional y el software Audacity para la toma de las muestras de las palabras del diccionario. Para la codificación del sistema se ha empleado el software Matlab © 1994-2011 The MathWorks, Inc.

A continuación se explican brevemente los métodos utilizados en cada etapa. La Fig. 1 muestra el diagrama de bloque de las etapas.

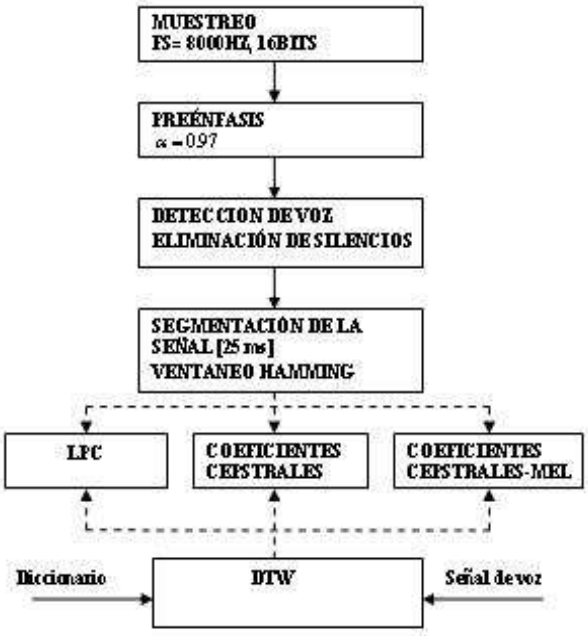


**Fig. 1.** Diagrama de bloque.

### 2.1 Muestreo

Se trata del muestreo y cuantificación de la señal de voz a identificar. La adquisición se hace con una frecuencia de muestreo de 8000 Hz con cuantización de 16 bits, y un solo canal de sonido (mono).

### 2.2 Preénfasis

En esta etapa la señal de voz es procesada por un filtro FIR que tiene como función acentuar las frecuencias altas, que son ante las cuales el oído presenta mayor sensibilidad, más específicamente en la zona de los 3000hz.

La ecuación del filtro es la siguiente:

$$S(n) = V(n) + \alpha S(n-1) \tag{1}$$

Donde para este caso se ha definido $\alpha = 0.97$ .

### 2.3 Detección de voz y eliminación de silencios

Para la eliminación de silencios al principio y al final de la muestra se ha utilizado el algoritmo de Rabiner-Lamel [3]. Este algoritmo se basa en los principios de energía de la señal y la tasa de cruces por cero.

### 2.4 Segmentación y ventaneo

La señal de voz se divide en segmentos de tiempo dentro de los cuales se considera que la señal es cuasi-estacionaria, esto es aproximadamente entre 20 y 30 ms. En nuestro caso se escogió segmentos de 25 ms.

Por otro lado, el incremento o solapamiento entre segmentos elegido fue del 50%, ya que determina una eficiencia razonable para nuestro fin.

A su vez a cada uno de los segmento se les aplica ventaneo de Hamming con el fin de suavizar los valores de los bordes con respecto a los valores centrales del segmento, que son los mas importantes para la extracción de los parámetros característicos de la señal de voz.

### 2.5 Extracción de parámetros

En esta etapa se procede a extraer de cada segmento de la señal de voz, aquellos parámetros que sean característicos de estos fragmentos y que en definitiva, en su conjunto total definen una forma de representar la señal de voz.

En esta experiencia, se ha probado con tres técnicas para obtener tales parámetros, los cuales son el método de Predicción Lineal (LPC), Coeficientes Cepstrales (CC) y Coeficientes Cepstrales en escala de Mel (MFCC).

El orden usado para cada método, es decir, la cantidad de parámetros a extraer por cada segmento, se ha fijado en una cantidad de diez valores.

A continuación se explica brevemente en que consisten cada uno de los métodos.

En primer lugar describiremos el método de predicción lineal o LPC [1]. Este parte de la suposición de que el valor de una muestra dada de la señal puede ser obtenido a partir de la combinación lineal de los valores de las muestras anteriores.

$$S_p = \sum_{k=1}^{N} a_k \, s_{n-k} \tag{1}$$

Los coeficientes $a_k$ son los parámetros deseados. Para comenzar su búsqueda se define primero el error de predicción:

$$e[n] = S[n] - S_p[n] \tag{2}$$

O lo que es lo mismo:

$$e[n] = S[n] - \sum_{k=1}^{N} a_k \, s_{n-k} \tag{3}$$

Los coeficientes que minimizan el error surgen de aplicar el método de mínimos cuadrados y resolver el sistema de ecuaciones que se presenta. Para eso se ha usado el algoritmo de Levinson-Durbin [1].

En segundo lugar tenemos el método de Coeficientes Cepstrales [2]. Este consiste en extraer los primeros componentes del Cepstrum, en lo cuales se halla la información de la envolvente del espectro de frecuencias de la señal de voz, es decir, la información de la respuesta en frecuencia del filtro que modela el tracto vocal.

El Cepstrum real de una señal de voz $V(f)$ se define como sigue:

$$c(t) = DFT^{-1}\left(\log V(f)\right) \tag{4}$$

De esta manera la elección de un número determinado de coeficientes cepstrales por cada segmento de la señal, conforman un conjunto de elementos representativos de la señal.

Por último tenemos los Coeficientes Cepstrales en escala de Mel [8], [2]. Lo primero que se debe saber es que la escala de Mel es una escala de frecuencia que aproxima la sensibilidad de percepción auditiva humana. La relación con la frecuencia lineal viene dada por:

$$\theta = 2595 \log_{10}(1 + \frac{f}{700}) \tag{5}$$

Los coeficientes cepstrales expresados en escala de Mel, son una buena forma de caracterizar una señal de voz.

La extracción se realiza filtrando el espectro de la señal de voz a partir de un banco de filtros diseñado en el dominio de Mel como el que se muestra en la Fig. 2.

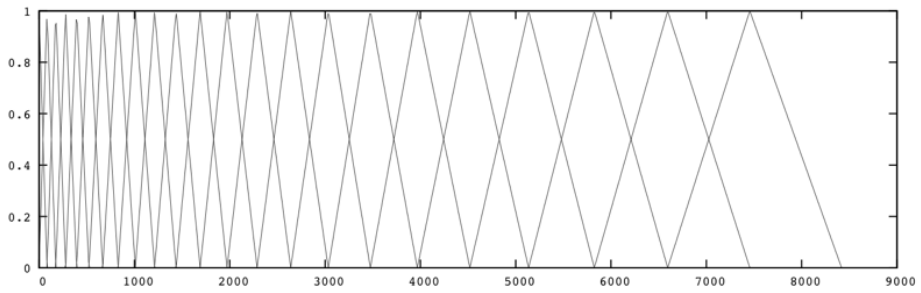


**Fig. 2.** Banco de filtros de Mel.

Las bandas de energía que determinan los filtros dan lugar a coeficientes que al aplicarle la función logarítmica y la transformada coseno discreta, dan como resultados coeficientes cepstrales, pero en escala de Mel.

La ventaja de utilizar estos coeficientes para caracterizar una señal de voz deriva del hecho de que en un número reducido de parámetros se concentra toda la información y además se prioriza la información del tracto vocal y no tanto la de la excitación y/o de las cuerdas vocales, lo que nos ayuda en la tarea del reconocimiento de palabras. Por otro lado contar con un número reducido de parámetros característicos, disminuye la carga de operaciones a realizar en etapas posteriores.

### 2.6 Reconocimiento

Para esta fase se cuenta con el diccionario de palabras. Esto no es otra cosa más que una base de datos que contiene almacenados los parámetros característicos

extraídos de cada realización (o muestras) de los comandos/palabras-aisladas que se espera que el sistema sea capaz de reconocer.

Los parámetros de la señal de voz y los parámetros de las señales de voces registradas en el diccionario son los argumentos de comparación para el núcleo del sistema de reconocimiento de voz en el que se ha basado el presente trabajo: el método de alineación temporal o Dynamic Time Warping (DTW) [4], [9], [10]. Este método tiene como principal característica el poder determinar una "distancia" que determine un grado de similitud entre dos señales(o conjunto de parámetros de la señal para este caso de aplicación particular) cuyas duraciones temporales sean distintas. De este modo las comparaciones entre parámetros de un conjunto de palabras iguales, pero que difieran en duración y velocidad de pronunciación deberían presentar distancias más bien homogéneas.

El procedimiento de comparación del método consiste básicamente en realizar un plano cuyos ejes son los parámetros de la señal de referencia (palabra del diccionario) y la señal a identificar. Para cada punto del plano se debe calcular la distancia euclídea entre sus dos coordenadas.

La distancia resultante de la comparación se obtiene de encontrar la ruta del origen al final de las intersecciones que minimice la distancia recorrida, y que además se atenga a la restricción de que los valores que componen al camino deben estar ubicados cerca de la diagonal.

Una forma muy común de hallar esta distancia es, a partir del plano-matriz con distancias euclídeas, crear una nueva matriz que acumule distancias recorridas en base a las siguientes expresiones:

$$\begin{aligned} DistMinima &= \min(A(i,j), A(i+1,j), A(i,j+1)) \\ A(i+1,j+1) &= D(i+1,j+1) + DistMinima \end{aligned} \tag{6}$$

Donde $A$ es la matriz de distancias acumuladas de tamaño $nxm$ . El elemento de la matriz de índice $n,m$ contiene la distancia final buscada entre los parámetros de señales comparadas.

Se ha de mencionar que para el sistema que se ha implementado, por conveniencia, los parámetros de las señales se almacenan en matrices.

Para el sistema de decisión además de la distancia se requiere una cota de error (cota C1) que sirva de tolerancia para la identificación.

Para obtener tal cota se calculan todas las distancias entre parámetros de realizaciones de una misma palabra para cada palabra del diccionario y se arma un vector. A partir de este se calcula su media $\mathbf{x}$ y desvío $\boldsymbol{\sigma}$.

Finalmente el proceso para identificar una palabra es procesar la señal de entrada y obtener los parámetros que la identifican a partir de alguno de los métodos descriptos anteriormente obteniendo una matriz de dimensión (numero de ventanas x orden del método), luego comparar esta matriz con cada una de las matrices del diccionario mediante DTW y quedándonos solo con la menor distancia encontrada.

Si esta distancia es menor que la media $\mathbf{x}$ se acepta la palabra de la que fue obtenida esta distancia como palabra identificada. Si es mayor a la media se comprueba si se encuentra en el intervalo $\mathbf{x} + \boldsymbol{\sigma}$, si es así también consideramos la identificación como exitosa. Sino el programa simplemente no decide y esperamos una nueva entrada.

## 3 Resultados

Para probar el programa se tomaron 4 muestras de cada palabra del diccionario. Obteniendo un total de 40 muestras para probar el funcionamiento del mismo.

**Tabla 1.** Efectividad de reconocimiento para la cota C1

| | LPC | CC | MFCC |
|---|---|---|---|
| Total palabras probadas | 40 | 40 | 40 |
| Palabras identificadas | 38 | 39 | 37 |
| Palabras no identificadas | 2 | 1 | 3 |
| Porcentaje de Efectividad | 95% | 97.5% | 92.5% |

Los resultados obtenidos al utilizar la cota (C1) explicada anteriormente son muy positivos como se puede ver en la Tabla 1. Sin embargo por lo explicado en la descripción de los métodos se esperaba que el más preciso de ellos sea el MFCC. Al observar la interacción de estos métodos con la cota elegida se decidió probar con otro valor para acotar la decisión, ya que se pudo observar que la menor distancia encontrada por el método de Mel siempre era la palabra correcta, por lo que evaluamos que la cota anterior podía ser relativamente chica.

Entonces se apuntó a utilizar un método que aproveche las diferencias relativas entre las distancias de palabras iguales y la distancia de palabras distintas que según lo observado es mayor en el método de MFCC.

Para ilustrar esto en la Fig. 4 vemos las diferencias relativas entre la media de distancias entre palabras iguales y la media de la distancia entre palabras distintas para cada método.

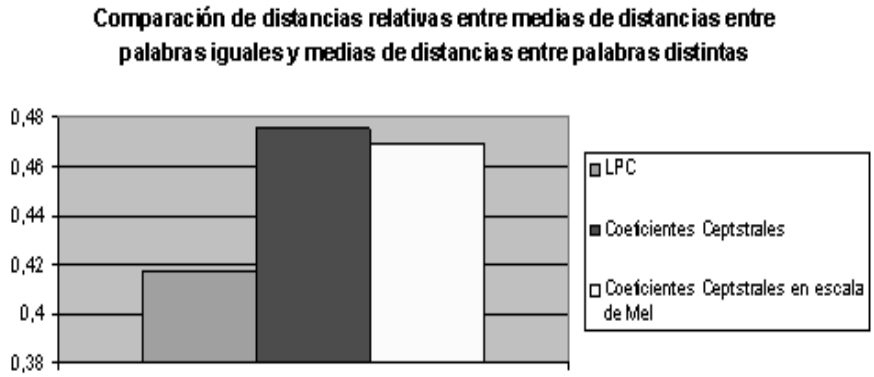


**Fig. 3.** Gráfico comparativo

A partir de lo expuesto se definió entonces una nueva cota (C2) como el promedio de la máxima distancia entre palabras iguales y la mínima distancia entre palabras distintas.

El método de identificación sigue siendo igual sólo que cambia la elección de la cota.

**Tabla 2.** Efectividad de reconocimiento para la nueva cota C2

| | LPC | CC | MFCC |
|---|---|---|---|
| Total palabras probadas | 40 | 40 | 40 |
| Palabras identificadas | 38 | 39 | 40 |
| Palabras no identificadas | 2 | 1 | 0 |
| Porcentaje de Efectividad | 95% | 97.5% | 100% |

## 4 Conclusiones

Se ha podido constatar que el sistema RAH tratado tiene un funcionamiento eficaz para un conjunto pequeño de palabras. En el caso de 10 palabras, y a partir de tolerancias ajustadas para la estrategia de identificación y un costo computacional medio, pudimos obtener una efectividad del 100% para el método MFCC, presentando un buen tiempo de respuesta. Este último método resultó ser el mejor, seguido por CC y LPC.

Estudios realizados han demostrado que usando métodos de orden 30 aproximadamente se ha podido obtener una efectividad del 85% [5], [6], [7]. En nuestro caso se ha podido optimizar el reconocimiento focalizándose en el sistema de decisión y dejando fijo el orden.

Una vez más, este rendimiento se logra para un número reducido de palabras. Por lo que al aumentar significativamente el número de palabras los resultados se vuelven impredecibles, el costo computacional es alto e inútil, y los tiempos de respuesta son pobres.

En definitiva, para aplicaciones que consistan en comandos de un vocabulario reducido, y cuya ejecución se lleva a cabo en ambientes libres de ruidos excesivos, la sencillez de este tipo de sistemas los coloca como la opción mas conveniente a utilizar.

## 5 Agradecimientos

A la ingeniera Susana Vanlesberg por sus consejos en la búsqueda de un sistema de decisión en la fase de reconocimiento.

Al ingeniero Leandro Vignolo por su invaluable ayuda y constante apoyo.

## 6 Referencias

1. Diego H. Milone, Hugo L. Rufiner y otros, “Introducción a las Señales y Sistemas Discretos”, Editorial Eduner, pp. 169-192, 2009.
2. R. Deller, J. G. Proakis, J. H. Hansen, “Discrete-Time Processing of Speech Signals”, Prentice Hall, 1993. 4.1, 4.2.1, 4.2.2, 6.1, 6.2.
3. Lori F. Lamel, Lawrence R. Rabiner, Aaron E. Rosenberg Y Jay G. Wilpon, “An improved endpoint detector for isolated word recognition”, IEEE transactions on acoustics, speech, and signal processing, vol. assp-29, no. 4, august 1981.
4. Titus F. Furtună, “Dynamic Programming Algorithms in Speech Recognition” en Revista Informática Económica, Academy of Economic Studies, Bucharest, 2008.
5. P. Sanz Leon, E. Vera de Payer, “Reconocimiento de comandos de voz aplicado a sistema robótico médico”.
6. Roberto A. Carrillo, César S .San Martín, “Implementación De Un Reconocedor De Palabras Aisladas Dependiente Del Locutor”, revista facultad de ingeniería, u.t.a. (chile), vol. 12 n°1 2004, pp.9-14.
7. H. Borrero, Y. Baquero, Z. Alezones “Reconocimiento de Palabras Aisladas Utilizando LPC Y DTW, para control de navegación de un mini-robot”, IEEE.
8. Sigurdur Sigurdsson, Kaare Brandt Petersen y Tue Lehn-Schiøler, “Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music”, Informatics and Mathematical Modelling,Technical University of Denmark.
9. Lácides A. Ripoll Solano “Verificación de hablante basado en Dynamic Time Warping”.
10. Waleed H. Abdulla, David Chow, and Gary Sin, “Cross-words Reference Template for DTW-based Speech Recognition Systems”, Electrical and Electronic Engineering Department, University of Auckland, Auckland, New Zealand.