*Conciencia Tecnológica No. 44, Julio-Diciembre 2012*

# Reconocedor de Palabras con el uso de Regresión Lineal y Coeficiente Muestral

Investigación

Dr. Carlos Alejandro de Luna Ortega[1,2], Dr. Miguel Mora González[1], Dr. Julio César Martínez-Romo[3], Dr. Francisco Javier Luna Rosas[3]

[1] Universidad de Guadalajara, Centro Universitario de los Lagos.
Av. Enrique Díaz de León 1144, Paseos de la Montaña, C.P. 47460, Lagos de Moreno, Jal. México
alejandro.deluna@upa.edu.mx, mmora@culagos.udg.mx

[2] Universidad Politécnica de Aguascalientes, Departamento de Ingeniería Mecatrónica

[3] Instituto Tecnológico de Aguascalientes, Departamento de Ingeniería Electrónica

**Resumen**

En el presente trabajo se aborda el uso de la técnica del coeficiente de correlación muestral como un método de reconocimiento de palabras aisladas pronunciadas por un hablante, tratando de combatir el problema de respuesta y complejidad en el proceso algorítmico en un sistema embebido, este método se propone para su uso en lugar de algoritmos más complejos en su estructura algorítmica.

La implementación de la técnica, consiste en caracterizar la señal de voz mediante el uso de los Coeficientes de Predicción Lineal (LPC), para la obtención de un vector característico, que será procesado por la técnica del coeficiente de correlación muestral R, para realizar el reconocimiento de la palabra contra la muestra y obtener un resultado. Con este método propuesto, se trabajó con la pronunciación de las palabras adelante, atrás, derecha e izquierda, obteniendo un 8% de error en el reconocimiento de manera global, ofreciendo un porcentaje igual o cercano a algoritmos de estructura compleja como las redes neuronales.

**Palabras clave: c**oeficiente muestral R, reconocimiento de voz, reconocedor de palabras, speech recognition.

**Abstract**

This paper presents the use of the technique of the sample correlation coefficient as a method of recognition of isolated words uttered by one speaker, trying to combat the problem of response and algorithmic complexity in the process in an embedded system, this method are proposed instead of others algorithms with more complex algorithm structure.

The implementation of the technique consists in characterize the speech signal using Linear Prediction Coefficients (LPC) to obtaining a characteristic vector that will be processed by the technique of the sample correlation coefficient R for the speech recognition against the sample and get a result. With this proposed method worked with the pronunciation of words forward, backward, left and right, getting a 8% error in the recognition globally, offering a rate equal or close to structure of complex algorithms such as neural networks.

**Key words: s**ample rate R, word recognition, speech recognition.

## Introducción

Conforme avanza la tecnología existe una tendencia a la automatización de los procesos. El reconocimiento de voz no se ha quedado atrás en dicha tendencia, para su muestra, se encuentra el Reconocimiento Automático de Voz (ASR, por sus siglas en inglés) [1]-[4]. No obstante a los avances en la tecnología el reconocimiento de voz sigue siendo un objetivo difícil de alcanzar, esto debido a que el humano promedio no solo considera la información de entrada al oído para entender una conversación, sino que también considera el contexto en el que se da dicha información [5].

Los principales problemas que evitan lograr altos porcentajes de reconocimiento en los ASR son: la Variación en las Condiciones Fisiológicas de los seres humanos [6]-[9], y el ruido que se incrusta en la señal digitalizada [10]-[12].

El primer problema se presenta debido a la gran disparidad entre los registros vocales, los cuales dependen de variables fisiológicas como lo son: el género, la edad, los acentos típicos de cada región, las distintas estructuras anatómicas (tanto en el oído como en la boca), los estados de ánimo, etc. Por lo que la pronunciación de la misma palabra por la misma persona o por una persona diferente no necesariamente contiene el mismo patrón acústico [13].

Para el segundo problema existen múltiples factores que pueden originar el ruido al digitalizar la voz, como lo son: el ruido ambiente, el ruido eléctrico producido por lámparas y transformadores, la calidad en las tarjetas de captura de sonido, así como los métodos de digitalización, etc. [14]. Estos tipos de ruido son de

factores indeseables que impactan negativamente en el desempeño de un sistema ASR [15].

Generalmente un reconocedor de palabras predefinidas utiliza algoritmos complejos, esto es, estructuras algorítmicas con complejidad computacional alta, que hacen tener consumos elevados en recursos computacionales con tiempos de respuesta altos para este tipo de procesos. Por lo que el reto es encontrar un algoritmo simple y sencillo que permita utilizar menos recursos computacionales para el reconocimiento de voz.

Los métodos más utilizados en el reconocimiento de voz son el alineamiento dinámico del tiempo (DTW, por sus siglas en inglés) [16]-[17], modelos ocultos de Markov (HMM, por sus siglas en inglés) [13], modelos híbridos que utilizan además redes neuronales [6]-[7], [18]; estos algoritmos representan procesos computacionales complejos para las etapas de entrenamiento y reconocimiento, lo que ocasiona que los recursos informáticos estén basados en computadoras potentes dedicadas exclusivamente para su operación.

Con esta premisa se debe buscar algoritmos que demanden pocos recursos informáticos con una complejidad computacional baja, misma que puede darse con el uso de operaciones simples como lo son la suma y el producto.

La arquitectura propuesta se basa en utilizar el modelo de regresión lineal con el uso del coeficiente muestral R (CMR), para reconocer una palabra de otra. Sin necesidad de utilizar algoritmos más complejos, como prueba para poder extenderlo a un vocablo más amplio.

En las siguientes secciones se presentan los métodos utilizados para la captura, el filtrado, la caracterización y el reconocimiento de la voz. En las últimas dos secciones se discuten los resultados obtenidos y se presentan las conclusiones de utilizar el método de regresión lineal como un reconocedor de voz.

## Diseño metodológico

El sistema de reconocimiento de voz que se plantea se puede observar en la figura 1, para ello se abordan los aspectos teóricos de las técnicas utilizadas para dicho sistema.

### A. Captura de Voz

La captura de la voz se realizó mediante el micrófono integrado de la computadora (Laptop Dell XPS), con una serie consecutiva de pronunciaciones de la misma palabra durante un minuto, logrando obtener variaciones en los parámetros de velocidad, frecuencia e intensidad.

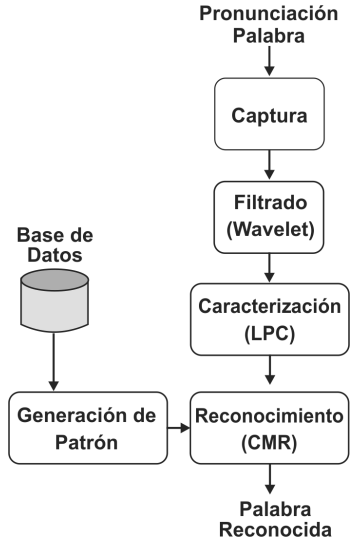


**Figura 1**. Etapas del Reconocedor de voz propuesto

El corpus que se utiliza en dicho experimento consta de cuatro palabras en español: izquierda, derecha, atrás, adelante.

### B. Filtrado

El filtro *wavelet denoising* se basa en suponer que se quiere recuperar una función desconocida *f* de un dato ruidoso descrito por [19]:

$$d_i = f(t_i) + \sigma z_i, \qquad i = 0,\ldots,n-1 \tag{1}$$

donde $t_i$, $z_i$ y $\sigma$ son el rango de muestreo $i/n$, un ruido blanco gaussiano estándar (señal aleatoria con distribución normal) y el nivel de ruido, respectivamente.

Buscando optimizar el error cuadrático medio

$$n^{-1}E\left\|\hat{f}-f\right\|_{\ell_n^2}^2 = \tag{2}$$
$$n^{-1}\sum_{i=0}^{n-1}E(\hat{f}(i/n)-f(i/n))^2,$$

es sujeto a la condición de tener un alta probabilidad de que $\hat{f}$ es al menos tan suave como $f$.

Donoho y Johnstone proponen un umbral para recobrar funciones de una señal de ruido descrito en los siguientes pasos [20]:

1. Se aplica un filtrado piramidal utilizando coeficientes Dauchechies a los datos medidos, para obtener los coeficientes wavelet en varios niveles.
2. Se aplica el umbral para cada nivel mediante el cálculo de un umbral universal determinado por:

$$T = \sqrt{2\sigma^2 \log N} \tag{3}$$

donde $\sigma^2$ es la varianza de los valores de la señal original y N es el tamaño.

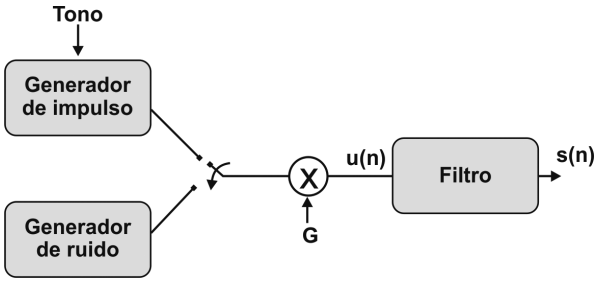


**Figura 2**. Modelo LPC de producción del habla.

3. Se invierte el filtro piramidal, recobrando la señal sin ruido.

### C. Caracterización

Para la extracción de características, la Codificación por Predicción Lineal (LPC, por sus siglas en inglés) es una de las técnicas más utilizadas, partiendo de la idea que se puede predecir la muestra presente a partir de una combinación lineal de las muestras pasadas, generando una descripción espectral basada en segmentos cortos de señal, considerando una señal $s[n]$ a una respuesta de un filtro todo-polo de una excitación $u[n]$.

La Figura 2 muestra el modelo en que se basa el LPC, considerando que la excitación $u[n]$ es el patrón a reconocer. La función de transferencia del filtro se describe como [21]:

$$H(z)=\frac{S(z)}{U(z)}=\frac{G}{1-\sum_{k=1}^{p}a_k z^{-k}}=\frac{G}{A(z)}, \qquad (4)$$

donde $G$, $a_k$ y $p$ son el parámetro de ganancia, los coeficientes del filtro y el orden de dicho filtro, respectivamente. La señal filtrada $s[n]$ se relaciona mediante la excitación $u[n]$, esto es:

$$s[n]=\sum a_k s[n-k]+Gu[n]. \qquad (5)$$

La estimación de los coeficientes de predicción se obtiene minimizando el error de predicción, dado por:

$$e[z]=s[z]\left(1-\sum_{k=1}^{p}a_k z^{-k}\right). \qquad (6)$$

El error de predicción en un corto tiempo se define como:

$$E=\sum_m e[m]^2=\sum_m (s[m]-\sum_{k=1}^{p}a_k s[m-k])^2. \qquad (7)$$

El mínimo error cuadrático total, denotado por $E$, se obtiene al expandir la ecuación (7), lo que es

$$Ep=\sum_n s^2[n]+\sum_{k=1}^{p}a_k\sum_n s[n]s[n-k]. \qquad (8)$$

Utilizando el método de auto-correlación para solucionarla y asumiendo que el error es minimizado sobre la duración infinita la ecuación (8) se puede rescribir como:

$$Ep=R(0)+\sum_{k=1}^{p}a_k R(k), \qquad (9)$$

donde

$$R(i)=\sum_{n=-\infty}^{\infty}s[n]s[n-k]. \qquad (10)$$

Una de las formas más comunes de encontrar los coeficientes de predicción $\{a_k\}$, es mediante métodos computacionales, el algoritmo de Levinson-Durbin es uno de los más utilizados, el cual se describe en la Tabla 1 [22].

**Entradas**
p = número de coeficientes a calcular
R = matriz de auto-correlación
**Salidas**
$a_k$ = coeficientes de predicción
**Variables utilizadas**
$E^i$ = Error en la posición i
**Algoritmo**

| | |
|---|---|
| **Paso 1** | $E^0=R(0)$ |
| **Paso 2** | $a_0=1$ |
| **Paso 3** | Para $i = 1, 2, \ldots, p$ |
| **Paso 4** | $k_i=\left(R(i)-\sum_{j=1}^{i-1}a_j^{(i-1)}R(i-j)\right)/E^{(i-1)}$ |
| **Paso 5** | $a_i^{(i)}=k_i$ |
| **Paso 6** | si $i>1$ entonces para $k=1,2,\ldots,i-1$ |
| **Paso 7** | $a_k^{(i)}=a_k^{(i-1)}-k_i a_{i-k}^{(i-1)}$ |
| **Paso 8** | fin |
| **Paso 9** | $E^{(i)}=(1-k_i^2)E^{(i-1)}$ |
| **Paso 10** | Fin |
| **Paso 11** | $a_k=a_k^{(p)} \quad j=1,2,\ldots,p$ |

**Tabla 1**. Algoritmo de Levinson-Durbin

Una importante característica de este algoritmo es que al hacer la recursión se realiza la estimación del error de predicción cuadrático medio. Dicha predicción satisface la función del sistema, determinada en la ecuación (8), la cual corresponde al término A(z) de la ecuación (4), esto es:

$$A^{(i)}(z)=A^{(i-1)}(z)-k_i z^{-i}A^{(i-1)}z^{-1}, \qquad (11)$$

donde la parte fundamental para la caracterización de la señal en coeficientes de predicción, radica en establecer un número adecuado de coeficientes p de acuerdo a la frecuencia de muestreo ($f_s$) y en base a la resonancia en kHz [21], esto es:

$$p=4+\frac{f_s}{1000}, \qquad (12)$$

que da como resultado el óptimo número de coeficientes LPC es el que representa el menor error cuadrático medio posible.

**D. Regresión Lineal y coeficiente muestral R**

El concepto de la regresión lineal se basa en encontrar la mejor relación entre las variables dependientes e independientes, esto al cuantificar la intensidad de dicha relación sobre una línea recta [24]. En el caso de la regresión lineal simple las variables dependiente y e independiente x, son nombradas variable de respuesta y regresor (o predictor), respectivamente [23]. La regresión lineal simple está dada por:

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (13)$$

donde $\beta_0$, $\beta_1$ y $\varepsilon$ son el coeficiente de intercepción de la línea, el coeficiente de la pendiente y el término de error aleatorio con promedio cero y varianza desconocida, respectivamente.

El Coeficiente muestral R es utilizado como juez para observar la adecuación del modelo de regresión lineal, está dado por [23]

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad (14)$$

donde $S_{xx}$ es el cuadrado de la diferencia de los elementos del patrón y la media de dicho patrón, $S_{yy}$ es el cuadrado de la diferencia de los elementos del dato que se compara con el patrón y la media del patrón, y $S_{xy}$ es la multiplicación de los elementos del dato que se compara con el patrón por la diferencia de los elementos del patrón y la media del patrón.

**E. Generación de Patrón**

La generación del patrón consiste en tomar 35 muestras de cada palabra, esto es, cada muestra de cada palabra se le calculan los coeficientes R, con estos 35 coeficientes R de cada palabra se les calcula la media aritmética, generando así un patrón característico, de manera estadística.

**F. Reconocimiento**

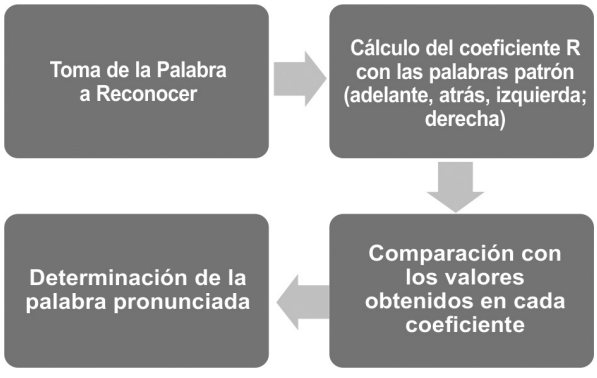


**Figura 3**. Diagrama de bloques del proceso de reconocimiento

Para el reconocimiento de las palabras pronunciadas, se utiliza el diagrama a bloques que se presenta en la figura 3, con el cual se obtienen los coeficientes muestrales R en razón a los patrones obtenidos de cada palabra del corpus.

## Resultados y Discusión

A partir del modelo simplista que ofrece la metodología del coeficiente muestral R en el reconocimiento de palabras de un locutor aislado, se obtuvieron los siguientes resultados.

El porcentaje de reconocimiento por palabra se presenta en la Tabla 2, donde se puede observar que se alcanzó un 92.5% de reconocimiento.

| Palabra | % de Reconocimiento | |
|---|---|---|
| | R | DTW |
| Adelante | 90 | 93 |
| **Atrás** | **90** | **100** |
| Izquierda | 100 | 90 |
| **Derecha** | **90** | **87** |
| **% Promedio** | **92.5** | **92.5** |

**Tabla 2**. Porcentaje de reconocimiento del método R vs. DTW.

La prueba del método se realizó con un corpus constituido por 400 pronunciaciones, 100 de cada palabra, en la Tabla 3 se puede observar una matriz de confusión donde se clarifican que palabra se dio un falso positivo en el método.

| | Adelante | Atrás | Izq. | Der. | % |
|---|---|---|---|---|---|
| **Adelante** | 90 | - | 4 | 6 | 90% |
| **Atrás** | 3 | 90 | 3 | 4 | 90% |
| **Izquierda** | - | - | 100 | - | 100% |
| **Derecha** | 5 | 5 | - | 90 | 90% |

**Tabla 3**. Matriz de confusión del método R.

Los datos obtenidos indican que el método presenta un porcentaje de reconocimiento similar al DTW, con lo cual se da un interés en ver su comportamiento al aumentar el corpus.

El modelo de R presenta una respuesta mayor al 90% de las cuatro palabras con la característica de tener un patrón calculado con promedios y no de manera azarosa como el DTW.

La arquitectura del modelo R se realiza mediante las operaciones básicas (sumas y productos), lo cual ofrece menor complejidad computacional.

## Conclusiones

El método de R ofrece una simplicidad en el reconocimiento de cuatro palabras, dando rapidez en el

tiempo de proceso y un alto porcentaje en el desempeño del reconocimiento, incluso igual que el del coeficiente $R^2$ mostrado en otros estudios.

El método de R puede ser en un algoritmo de mayorías junto con el método DTW, ya que ambos se complementan muy bien en sus porcentajes de reconocimiento.

El uso de esta técnica representa un ahorro en tiempo de cómputo, ya que solamente se realiza mediante sumas y no es necesario entrenamiento alguno, solamente la comparación de los promedios de las palabras de la base de datos contra la palabra a reconocer.

Esta es una técnica que no se ha empleado en el reconocimiento de voz y que puede dar buenos resultados.

**Referencias**

[1] Paulson, L. (2006). Speech Recognition Moves from Software to Hardware. *Computer IEEE* , 39 (11), p. 15-18.

[2] Varga, I., Aalburg, S., Andrassy, B., Astrov, S., Bauer, J., & Beaugeant, C. (2002). ASR in Mobile Phones - An Industrial Approach. *IEEE Transaction on Speech and Audio Processing* , 10 (8), 562-569.

[3] Alewine, N., Ruback, H., & Deligne, S. (2004). Pervasive Speech Recognition. *Pervasive computing* , 3 (4), 78-81.

[4] Smaragdis, P., & Shashanka, M. (2007). A frameword for secure speech recognition. *IEEE Transaction on Audio, Speech and Language Processing* , 15 (4), 1404-1413.

[5] Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing. A guide to Theory, Algorithm, and System Development*. Prentice Hall PTR (USA).

[6] Tebelskis, J. (1995). *Speech Recogniton using Neural Networks*. PhD Thesis . Pennsylvania, U.S.A: School of Computer Science Carniege Mellon University.

[7] Merlo, G., Fernández, V., Caram, D., & Priegue, R. (1997). Reconocimiento de voz mediante una red neuronal de Kohonen. *Proceedings of CACIC* 97 , p. 1-7.

[8] Campbell, J. (1997). Speaker Recognition: A tutorial. *Proceedings of the IEEE* , 85 (9), p. 1437-1462.

[9] Benseghiba, M., De Mori, R., & Deroo, O. (2007). Automatic Speech Recognition and Speech Variability: A Review. *Speech Communication*, 49 (1), p. 763-786.

[10] Proakis, J., & Manolakis, D. (2007). *Digital Signal Processing*. Michigan: Prentice Hall. (USA)

[11] Tsoukalas, D., Mourjopoulos, J., & Kokkinakis, G. (1997). Speech Enhancement Based on Audible Noise Suppression. *IEEE Transaction on Speech an Audio Processing* , 5 (6), p. 497-614.

[12] Oppenheim, A., & Schafer, R. (1999). *Discrete Time Signal Processing*. New York: Prentice Hall. (USA)

[13] De Luna-Ortega, C., Mora-González, M., & Martínez-Romo, J. (2006). Reconocimiento de voz con redes neuronales, DTW y Modelos Ocultos de Markov. *Conciencia Tecnológica* , 32 (1), p. 13-17.

[14] Benesty, J., Chen, J., & Arden, Y. (2009). Noise Reductions Algorithms in a generalized transform domain. *IEEE Transactions on Audio, Speech and Language Processing* , 49 (1), p. 1109-1123.

[15] Obaidat, M., Lee, C., Sadoun, B., & Nelson, D. (1999). Estimation of pitch period of speech signal using a new dyadic wavelet algorithm. *Information Sciences* , 119 (1), p. 21-39.

[16] Pham, C., Plötz, T., & Olivier, P. (2010). A Dynamic Time Warping Approach to Real-Time Activity Recognition for Foord Preparation en Ambient Intelligence de Lecture Notes in Computer Science, Springer (Berlin), p. 21-30

[17] Müller, M. (2007). *Information retrieval for music and motion*, Springer. (USA)

[18] Hung-Yan Gu; Chang-Yi Wu (2009). Model spectrum-progression with DTW and ANN for speech synthesis. ECTI-CON 2009. *6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2009. , vol.02, no., pp.1010-1013.

[19] Donoho, D. L. (1995). De-Noising by Soft-Thresholding. *IEEE Transactions on Information Theory*, 41 (3), p. 613-627.

[20] Donoho, D., & Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* , 81, p. 425-455.

[21] Rabiner, L. R., & Schafer, R. W. (2007). *Introduction to Digital Signal Processing*. Hanover, USA: Now Publishers Inc. (USA).

[22] Zaknich, A. (2005). *Principles of adaptive filters and self-learning systems*. Londrés: Springer.

[23] Montgomery, D., & Runger, G. (2003). *Applied Statistics and Probability for Engineers*. New York: John Wiley & Sons. (USA).

[24] Walpole, R., Myers, R., & Myers, S. (2006). *Probability & Statistics for Engineers & Scientists*. New York: Prentice Hall. (USA)

**Recibido**: 7 de noviembre de 2011
**Aceptado**: 12 de junio de 2012