Nova Scientia
Revista de Investigación de la Universidad De La Salle Bajío

Revista Electrónica Nova Scientia

Reconocimiento del habla mediante el uso de la correlación cruzada y una perceptrón multicapa

Speech recognition by using cross correlation and a multilayer perceptron

Carlos A. de Luna-Ortega[1,2], Miguel Mora-González[2], Julio C. Martínez-Romo[3], Francisco J. Luna-Rosas[3] y Jesús Muñoz-Maciel[2]

[1]Ingeniería en Sistemas Estratégicos de Información, Universidad Politécnica de Aguascalientes, Aguascalientes
[2] Universidad de Guadalajara, Centro Universitario de los Lagos, Lagos de Moreno, Jalisco
[3] Departamento de Ingeniería Eléctrica-Electrónica, Instituto Tecnológico de Aguascalientes, Aguascalientes

México

Carlos A. de Luna-Ortega. E-mail: alejandro.deluna@upa.edu.mx

Resumen

En el presente artículo se da a conocer una alternativa algorítimica a los sistemas actuales de reconocimiento automático del habla (ASR), mediante una propuesta en la forma de realizar la caracterización de las palabras basada en una aproximación que usa la extracción de coeficientes de la codificación de predicción lineal (LPC) y la correlación cruzada. La implementación consiste en extraer las características fonéticas mediante los coeficientes LPC, después se forman vectores de patrones de la pronunciación conformados por el promedio de los coeficientes LPC de las muestras de las palabras obteniendo un vector característico de cada pronunciación mediante la autocorrelación de las secuencias de coeficientes LPC; estos vectores se utilizan para entrenar un clasificador de tipo perceptrón multicapa (MLP). Se realizaron pruebas de desempeño previo entrenamiento con los diferentes patrones de las palabras a reconocer. Se utilizó la fonética de los dígitos del cero al nueve como vocabulario objetivo, debido a su amplia aplicación, y para estimar el desempeño de este método se utilizaron dos corpus de pronunciaciones: el corpus UPA, que contempla en su base de datos la pronuncación de la región occidente de México, y el corpus Tlatoa, que hace lo propio para la región centro de México. Las señales en ambos corpus fueron adquiridas en el lenguaje español, y a una frecuencia de muestreo de 8kHz. Los porcentajes de reconocimiento obtenidos fueron del 96.7 y 93.3% para las modalidades de mono-locutor para el corpus UPA y múltiple-locutor para el corpus Tlatoa, respectivamente. Asimismo, se realizó una comparación contra dos métodos clásicos del reconocimiento de voz y del habla, Dynamic Time Warping (DTW) y Hidden Markov Models (HMM).

Palabras clave: reconocimiento automático del habla, correlación cruzada, perceptrón multicapa, codificación de predicción lineal

Recepción: 14-02-2013 *Aceptación: 27-01-2014*

Abstract

It this paper we present an algorithmic alternative to the current Automatic Speech Recognition (ASR) systems by proposing a way to characterize words based on approximations that use an extracted coefficient from Linear Predictive Coding (LPC). The method consists in extracting phonetic characteristics through the use of LPC coefficients, after which pattern vectors are formed from the LPC coefficient averages taken from the word sampling, thus creating a unique vector for each pronunciation through the auto correlation of the LPC coefficient sequences. These vectors are used to train a Multilayer Perceptron (MLP) classifier. After training performance trials were executed. The sounds from the digits zero through nine where used as a target vocabulary, given its general use, and to estimate the performance of this method two corpus where used: the UPA corpus, which in its vocabulary uses a pronunciation familiar to the western part of Mexico, and the Tlatoa corpus, who's vocabulary presents a pronunciation typical of the central region of Mexico. The signals from both corpus where sampled in the Spanish language, and at a sampling frequency of 8kHz. The recognition rate for the mono-speaker from the UPA corpus and the multiple-speaker from the Tlatoa corpus were 96.7% and 93.3% respectively. Additionally, there where comparisons done against two classic methods used for speech recognition, Dynamic Time Warping (DTW) and Hidden Markov Models (HMM).

Keywords: automatic speech recognition, cross-correlation, multilayer perceptron, linear predictive coding

1. Introducción

Uno de los principales problemas del reconocimiento automático del habla (ASR) es la variación de las condiciones fisiológicas entre los humanos (Benzeghiba et. al., 2007, 763), tales como: la gran disparidad de los registros vocales, el género, la edad, la estructura anatómica, entre otras; otro aspecto que puede tener influencia en el problema es el estado de ánimo de la persona. Esto ocasiona que la pronunciación de la palabra por una misma persona no genere el mismo patrón acústico en diversas pronunciaciones (De Luna-Ortega et. al., 2006, 32). La solución a dicho problema tiende a formularse como un problema de clasificación estadística (Trentin, 2001, 91), con la idea de generar patrones que abarquen la mayor parte de las variaciones posibles, obteniendo con ello un aumento en el porcentaje de reconocimiento.

Una etapa clave en el reconocimiento de patrones es la extracción de características: en el habla, las técnicas más usuales para extraer características incluyen la extracción de coeficientes de predicción lineal (LPC), Cepstrum y los coeficientes cepstrales de frecuencia Mel (MFCC), entre otros (Rabiner, 2007,75); la técnica de LPC es utilizada debido a su capacidad para proporcionar estimaciones precisas de los parámetros de voz (Rabiner, 2007,75). El Cepstrum es otra alternativa de uso para caracterización de la voz debido a que obtiene espectros en ventanas de corto tiempo, además de que es menos susceptible a las distorsiones lineales (Schafer, 2007, 176), y el MFCC (que es una derivación del LPC) y sus derivados son los algoritmos de extracción más popular en uso para los sistemas de reconocimiento de voz (Rabiner, 2007, 166), debido a que están basados en la percepción que tiene el oído humano, es decir, que trabajan por bandas de frecuencia.

A la fecha, se han aplicado diferentes técnicas en el reconocimiento del habla, entre las cuales se encuentran la programación lineal (*Dynamic Time Warping*, DTW), los modelos ocultos de Markov (*Hidden Markov Models*, HMM), las redes neuronales artificiales (*Artificial Neural Networks*, ANN), Redes Bayesianas (*Bayesian Networks*, BN), y otras, obteniendo porcentajes de reconocimiento entre el 80% y el 97%, según la técnica utilizada y las palabras a reconocer (De Wachter et. al., 2007, 1377; Kinjo and Funaki, 2006, 3477; Romo et. al., 2008, 163; De Luna-Ortega et. al., 2006, 32; Irwin, 1988, 1412; Nefian et. al., 2002, 1; Oropeza and Suárez, 2006, 270; Chen et. al., 2011, 919; Zweing, 1999, 253; Livescu, 2003, 1 ; Wollmer, 2010, 867). Actualmente, los modelos complementarios entre extracción de características y técnicas de reconocimiento de voz han favorecido al aumento de la tasa de correcto reconocimiento.

El propósito de este trabajo es mostrar que el uso de la correlación cruzada aumenta la capacidad discriminante de los coeficientes LPC codificando palabras aisladas; para demostrarlo, los dígitos en español del cero al nueve son clasificados utilizando un perceptrón multicapa. Como se verá, esta combinación de LPC/correlación cruzada/perceptrón multicapa permite obtener altas tasas de clasificación correcta. Con el método propuesto no es necesario estandarizar la dimensionalidad de los vectores de las palabras para realizar el reconocimiento, obteniendo desempeños iguales o mayores que los algoritmos clásicamente aplicados como lo son HMM y DTW.

En las siguientes secciones se describe la correlación cruzada como una forma de realce de la caracterización de los coeficientes de LPC de la pronunciación de una palabra, aunado al perceptrón multicapa como reconocedor en un sistema ASR. Se describe el diseño experimental así como la configuración del perceptrón multicapa. Finalmente, en las dos últimas secciones se presentan los resultados y las conclusiones, respectivamente.

2. Método

La correlación tiene como objetivo principal el cálculo de la similitud de dos señales. La correlación cruzada entre dos señales dada por la señal $r_{xy}(l)$, que está definida por (Proakis, 2007, 106):

$$r_{xy}(l) = \sum_{n=-\infty}^{\infty} x(n)y(n-l) = \sum_{n=-\infty}^{\infty} x(n+l)y(n), \qquad l = 0,\pm1,\pm2,... \tag{1}$$

donde *x,y,l* y *n* son las dos señales a correlacionar, un parámetro de offset (tiempo o retardo), y el número de muestras de ambas señales, respectivamente. El subíndice *xy* que es usado en la ecuación (1) indica cuales señales son correlacionadas. En específico, si la señal tiene correlación consigo misma es nombrada autocorrelación que está definida como:

$$r_{xx}(l) = \sum_{n=-\infty}^{\infty} x(n)x(n-l) = \sum_{n=-\infty}^{\infty} x(n+l)x(n). \tag{2}$$

Una de las propiedades de la autocorrelación es la simetría geométrica que se obtiene, siendo ésta una función par como se puede ver en la Figura 1a, donde *x(n)* representa los 12 coeficientes de la pronunciación del dígito cuatro. La autocorrelación se representa matemáticamente por la siguiente ecuación:

$$r_{xx}(l) = r_{xx}(-l). \tag{3}$$

Esta ecuación es una propiedad exclusiva de la autocorrelación, pero es posible utilizarla en la correlación cruzada si las secuencias de ambas señales *x(n)* y *y(n)* son muy similares, esto es:

$$r_{xy}(l)\Big|_{y(n)\to x(n)} \approx r_{xx}(l) = r_{xx}(-l), \qquad (4)$$

por lo cual, la aproximación a la propiedad de simetría par que se observa en la ecuación (4) es una herramienta válida para discriminar características en el reconocimiento de patrones, como se puede observar en la Figura 1b, donde existe una correlación cruzada entre *x(n)* y *y(n)*, donde *y(n)* son los 12 coeficientes LPC de una pronunciación del dígito cuatro diferente a *x(n)*. Con ello, si se correlacionan dos secuencias altamente similares, su tendencia será orientada a ser una función par. Esto se puede determinar mediante el establecimiento de un grado de similitud que se procesa mediante técnicas de reconocimiento de patrones, y con ello establecer una clasificación, misma que se puede utilizar para la discriminación de una pronunciación corta. En caso de que las dos secuencias no fueran altamente similares, la función no se aproxima a una función par, como se puede observar en la Figura 1c, donde *z(n)* representa los coeficientes LPC de una pronunciación del dígito cinco en correlación cruzada con *x(n)*.

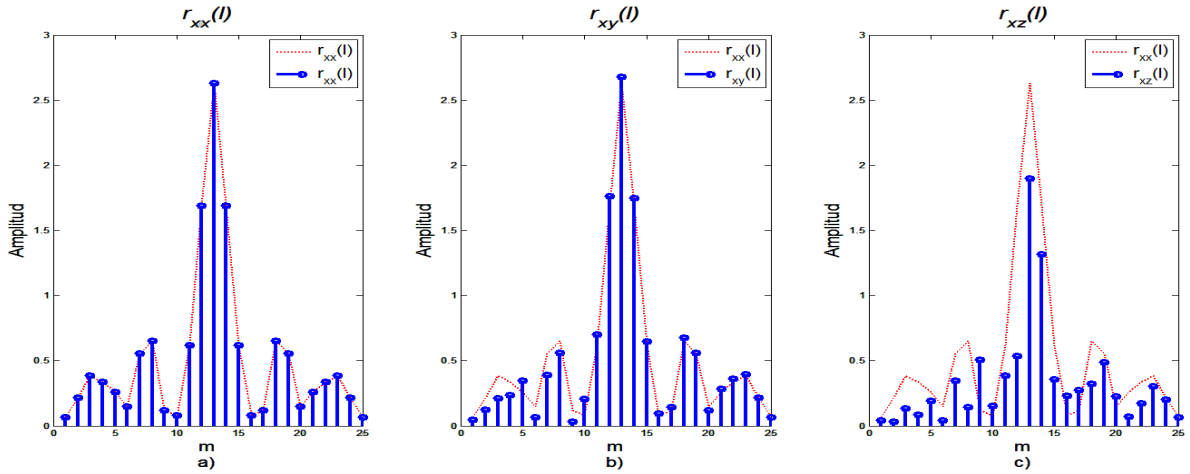

Figura 1. a) Auto-correlación *x(n)*, b) Correlación cruzada *x(n)* y *y(n)* y c) Correlación cruzada *x(n)* y *z(n)*.

3. Diseño Experimental

El diseño experimental utilizado en este sistema ASR se muestra en la Figura 2; en ésta, se muestran cinco etapas en las que se lleva a cabo el proceso: la adquisición y almacenamiento de datos, el pre-procesamiento (normalización y el filtrado), la extracción de características (LPC orden 12 y la Correlación), el entrenamiento y el reconocimiento.

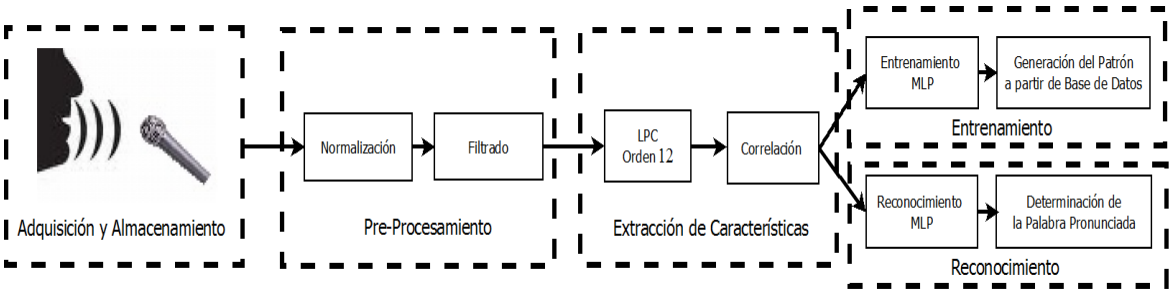

Fig. 2. Sistema ASR propuesto.

3.1 Corpus Utilizados

Se utilizaron dos bases de datos de pronunciaciones de dígitos con palabras en español con la finalidad de comprobar el algoritmo propuesto. La primera es la base de datos Tlatoa (Tlatoa, 2012), la cual considera pronunciaciones de la región centro de México, de la cual se tomaron para este estudio 100 pronunciaciones de cada uno de los dígitos del cero al nueve. Las pronunciaciones utilizadas de esta base de datos fueron de 12 hablantes (7 hombres y 5 mujeres), los cuales se seleccionaron aleatoriamente para no generar tendencias o desviaciones en este estudio. La segunda es una base de datos, recolectada en la Universidad Politécnica de Aguascalientes en el año 2012; en ésta, la captura de las palabras se realizó con el micrófono integrado de la computadora Laptop Dell XPS, con una serie consecutiva de la misma palabra durante un minuto de un solo hablante (hombre), logrando obtener variaciones en los parámetros de velocidad, frecuencia e intensidad, en una habitación de 3mx3m con ambiente controlado, buscando obtener pronunciaciones locales de la región occidente de México, consistente en 100 pronunciaciones de cada uno de los dígitos del cero al nueve, totalizando 1000 ejemplares. Para ambas bases de datos la frecuencia de muestreo fue de 8kHz.

3.2 Pre-procesamiento

En la etapa de pre-procesamiento, ver Figura 2, todas las señales a procesar fueron filtradas mediante el algoritmo de *wavelet denoising*, con un filtro de 12 niveles de resolución y 4 coeficientes de Daubechies, obteniendo una atenuación de 5dB de relación señal a ruido (SNR). El filtrado se realizó usando Matalb®, versión 2010ª. La selección del número de niveles y coeficientes se realizó mediante experimentación, tomando como base los mejores resultados en promedio, esto es, la menor distorsión de la señal. Para validar este proceso, se calculó el porcentaje de error de reconstrucción o distorsión de la señal, a partir de la ecuación 5 (Benzid,

2006, 1306). Los resultados se presentan en la Tabla 1, en donde se observa que el menor porcentaje de distorsión se obtiene con doce coeficientes.

$$PRD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{N}(x_i - \mu_i)^2}} \times 100 \qquad (5)$$

en donde: PRD, es el porcentaje de diferencia o distorsión de la señal de audio original y la filtrada (Percent Root-mean square Difference); x_i, es la señal original de audio; $\hat{x}_i$, es la señal filtrada; μ_i, es la media de la señal original.

Tabla 1. Porcentaje de distorsión en razón a niveles de resolución y coeficientes

Wavelet Madre	Niveles de resolución	Coeficientes	*PRD (%)*	Atenuación SNR (dB)
Daubechies	10	2	*79.40*	*4.75*
	10	4	*77.64*	*5.33*
	10	6	*76.82*	*5.61*
	12	2	*80.09*	*4.73*
	12	4	*75.84*	*5.33*
	12	6	*77.22*	*5.59*
	14	2	*80.89*	*4.76*
	14	4	*78.29*	*5.33*
	14	6	*77.44*	*5.62*

3.3 Extracción de características

Para la extracción de características de cada palabra, se obtuvieron 12 coeficientes de LPC con el objeto de modelar sus propiedades fonéticas (Makhoul, 1975, 561) y, a su vez, representar cada palabra en un número corto de coeficientes indistintamente de la duración de la misma. Dado que todos los datos adquiridos se muestrearon a 8kHz, se obtuvieron 12 coeficientes LPC por palabra según la expresión (Rabiner, 2007, 90):

$$p = 4 + \frac{f_s}{1000} \qquad (6)$$

en donde p es el número de coeficientes y f_s es la frecuencia de muestreo.

Una vez obtenido el vector con los 12 coeficientes LPC, se calcula la correlación cruzada entre la pronunciación analizada y un prototipo de clase o patrón de cada dígito, para obtener un nuevo vector con 23 coeficientes, que tiene un realce de las características de la pronunciación del dígito. El patrón o prototipo de cada dígitio se obtuvo en el espacio de los coeficients de predicción lineal al promediar los coeficientes de 50 pronunciaciones arbitarias de cada uno.

3.4 Generación de Patrones para Correlación Cruzada

Para la generación de patrones y prueba del desempeño de esta propuesta se definieron dos conjuntos: el de entrenamiento y el de prueba. El conjunto de entrenamiento consistió, en ambos corpus, de 500 pronunciaciones, esto es 50 pronunciaciones de cada dígito, donde se definió cada elemento del conjunto como $LPC_d^{\,n}$, donde d es el dígito pronunciado y n es el índice que identifica cada pronunciación. El conjunto de prueba se formó, para ambos casos, del corpus UPA y corpus Tlatoa, las 50 pronunciaciones restantes de cada dígito.

Con el conjunto de entrenamiento se obtuvieron los patrones para cada dígito, conformados por el promedio de los coeficientes LPC de las 50 pronunciaciones de cada uno de los dígitos de cada corpus. Definiendo dicho promedio como $\overline{LPC}_d$, donde d es el dígito pronunciado.

3.5 Fase de entrenamiento del MLP

Esta fase de identifica en la Figura 2. El perceptrón multicapa (MLP) está compuesto por una capa de entrada de 23 perceptrones con una función de activación logarítmica sigmoidal, que reciben los valores de la correlación cruzada, una capa oculta de seis perceptrones con una función de activación tangente sigmoidal y una capa de salida con un perceptrón con una función de activación lineal.

El entrenamiento del MLP se llevó a cabo mediante los datos de referencia presentados en la Tabla 2 usando el conjunto de entrenamiento definido en la sección 3.4, ese entreno al perceptron multicapas por un periodo de 10,000 épocas. Nótese que el perceptrón se entrena con las correlaciones del propio dígito vs. cada uno de los patrones o prototipos de clase en el espacio

LPC. Las implicaciones de esta particular selección de patrones de entrenamiento se comentarán en la fase de reconocimiento.

Tabla 2. Datos de referencia para entrenamiento de la MLP.

Dígito Pronunciado	Entrada del MLP (correlación cruzada entre)	Salida deseada del MLP
cero	$LPC_0^{\,n}, \overline{LPC}_0$	**0**
	$LPC_0^{\,n}, \overline{LPC}_1$	**0**
	⋮	⋮
	$LPC_0^{\,n}, \overline{LPC}_9$	**0**
uno	$LPC_1^{\,n}, \overline{LPC}_0$	**1**
	$LPC_1^{\,n}, \overline{LPC}_1$	**1**
	⋮	⋮
	$LPC_1^{\,n}, \overline{LPC}_9$	**1**
⋮		
nueve	$LPC_9^{\,n}, \overline{LPC}_0$	**9**
	$LPC_9^{\,n}, \overline{LPC}_1$	**9**
	⋮	⋮
	$LPC_9^{\,n}, \overline{LPC}_9$	**9**

3.6 Fase de Reconocimiento con el MLP.

La fase de reconocimiento es en la que se verifica el desempeño del sistema. Esta fase se explica basándose en el diagrama de la Figura 2 y los conjuntos de entrenamiento de la Tabla 2. Para estimar el desempeño del sistema se toman los conjuntos de prueba descritos en la sección 3.4. Los coeficientes LPC del dígito a ser reconocido (LPC_d) se someten a correlación cruzada con cada uno de los prototipos de dígito (LPC_d), por lo $\overline{\text{que}}$ al MLP se le ejecuta secuencialmente con cada una de las 10 secuencias de correlación, y la mayoría de valores de salida es directamente el dígito reconocido. Esto es consecuencia directa del método de entrenamiento, en el cual se ha enseñado al MLP que no importa contra cuál prototipo o patrón de dígito se correlacionen los coeficientes LPC de un dígito, éste siempre deberá de tratar de generar a la

salida el valor numérico del dígito de la palabra. Esto convierte al MLP en un generador secuecial de votos para un votante de mayoría.

4. Resultados

Se realizaron pruebas de reconocimiento monolocutor y multilocutor, obteniendo resultados de entre 90 y 100% para las 1000 pronunciaciones probadas en monolocutor, de las cuales 500 fueron de la base de datos de entrenamiento y 500 de la base de datos de prueba, así como entre 88 y 97% para las 1000 pronunciaciones probadas en la modalidad de múltiples locutores, de donde se tomaron 500 de la base de datos de entrenamiento y 500 de la base de prueba. Se observó que las palabras *tres*, *cuatro*, *cinco* y *siete* fueron las que mayor tasa de reconocimiento correcto presentaron, así como las palabras *uno* y *cero* obtuvieron una menor tasa de reconocimiento, esto para ambos corpus utilizados. En las Tablas 3 y 4 se muestran las comparaciones monolocutor para el corpus UPA y de múltiples-locutores para el corpus Tlatoa.

Se realizó la comparación del método propuesto contra los métodos de DTW y HMM, que son métodos muy populares en reconocimiento del habla, según la literatura (Rabiner, 89, 257; Takiguchi, et. al., 2001, 127; Abdulla, 2003, 1576; Itakura, 1975, 52). Tómese en cuenta que para el DTW no se extrajeron características, sino que se compararon las pronunciaciones de los digitos contra plantillas o patrones de los dígitos, en forma de vectores característicos, definidas en base a la menor distancia intraclase del patrón contra todas las pronunciaciones. Para el desarrollo del HMM se utilizó el Hidden Markov Model Toolkit. Los tres métodos se probaron usando 1000 pronunciaciones en monolocutor, utilizando la base de datos UPA. Se obtuvo una tasa de reconocimiento con el método aquí propuesto, esto es, para el total de las palabras probadas se lograron los siguientes resultados: 96.7, 94.3 y 90.9% de reconocimiento para el método propuesto, el DTW y el HMM, respectivamente. Los resultados por cada dígito se observan en la Figura 3 que representa las tasas de reconocimiento correcto por método.

Tabla 3. Matriz de confusión para Corpus UPA monolocutor

Palabra	Uno	Dos	Tres	Cuatro	Cinco	Seis	Siete	Ocho	Nueve	Cero	Porcentaje
Uno	90	5	-	-	3	-	-	2	-	-	90%
Dos	-	94	2	2	-	2	-	-	-	-	94%
Tres	-	-	100	-	-	-	-	-	-	-	100%

Cuatro	-	-	-	100	-	-	-	-	-	-	100%
Cinco	-	-	-	-	100	-	-	-	-	-	100%
Seis	-	-	-	-	-	94	3	1	-	2	94%
Siete	-	-	-	-	-	-	100	-	-	-	100%
Ocho	-	-	-	-	-	-	3	96	3	-	96%
Nueve	-	-	-	-	-	-	-	-	100	-	100%
Cero	-	-	-	2	1	2	1	1	-	93	93%

Tabla 4. Matriz de confusión para Corpus Tlatoa Múltiple locutor

Palabra	Uno	Dos	Tres	Cuatro	Cinco	Seis	Siete	Ocho	Nueve	Cero	Porcentaje
Uno	95	2	-	-	1	-	1	1	-	-	95%
Dos	-	88	3	2	3	2	-	-	1	1	88%
Tres	1	2	97	-	-	-	-	-	-	-	97%
Cuatro	-	-	-	95	-	1	2	-	-	2	95%
Cinco	-	-	-	-	93	-	2	3	2	-	93%
Seis	-	2	1	2	-	94	-	1	-	-	94%
Siete	1	1	1	1	-	-	95	-	1	-	95%
Ocho	-	2	1	2	2	-	-	91	-	2	91%
Nueve	1	2	-	2	-	-	-	-	95	-	95%
Cero	-	2	-	2	2	2	2	-	-	90	90%

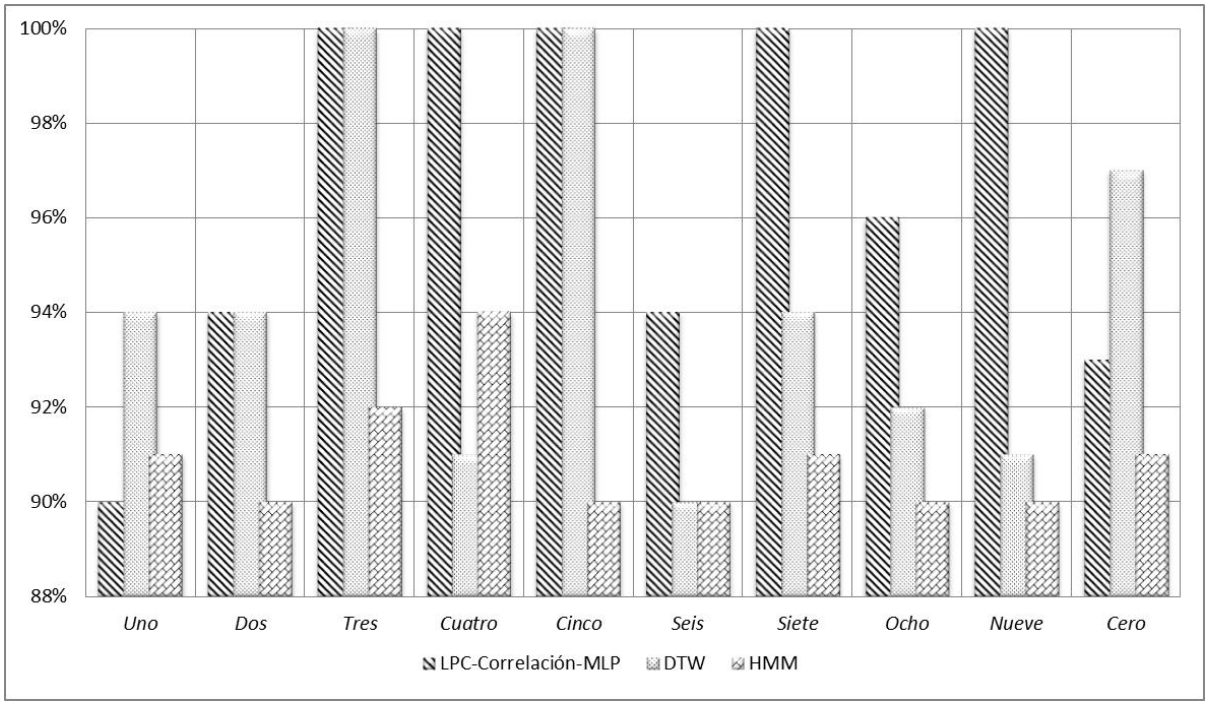

Fig. 3. Comparación mono-locutor entre el algoritmo propuesto, DTW y HMM, utilizando el Corpus UPA.

Asimismo, se realizó una validación con el Corpus Tlatoa del algoritmo propuesto mediante el método "*Hold out*" 50-50, obteniendo los resultados que se muestran en la Figura 4; de los cuales se puede establecer, que en diez pruebas de la validación se obtuvo un promedio de tasa de reconocimiento del 94%, presentando una tasa con el porcentaje más alto del 97% tanto para el conjunto de entrenamiento como para el conjunto de prueba.

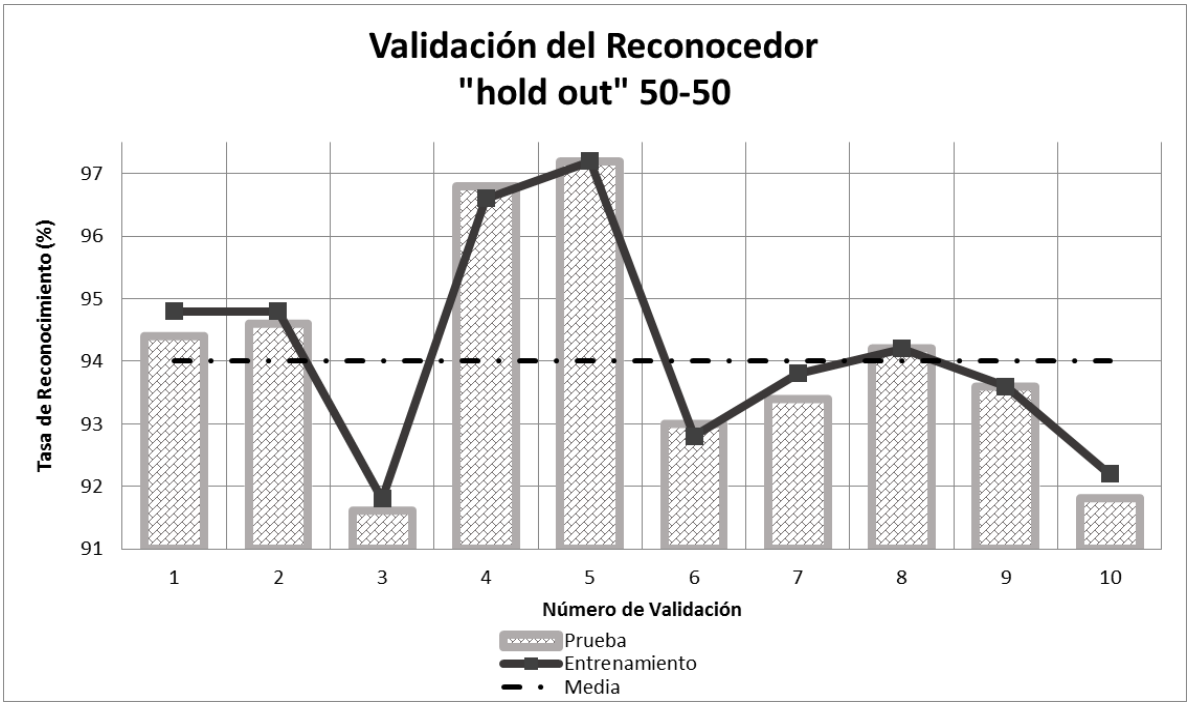

Fig. 4. Tasa de Reconocimiento de 10 pruebas de validación con el método "Hold out" con 50% de muestras de prueba y 50% de muestras de reconocimiento.

5. Conclusiones

Se ha presentado el método propuesto como una variante del reconocimiento de palabras aisladas, en el cual la correlación cruzada se ha utilizado para potenciar la extracción de características, lo cual promueve una mayor tasa de clasificación correcta en el algoritmo de clasificación. La correlación cruzada puede considerarse como una autocorrelación cuando existe una similitud entre el patrón y la palabra pronunciada.

Se escogieron los dígitos del cero al nueve como pronunciaciones a reconocer, debido a que son palabras cortas con gran uso en una vasta cantidad de aplicaciones. El lenguaje español mexicano, de dos regiones de pronunciación diferente, se seleccionó para realizar el desarrollo de la aplicación, ya que existen desarrollos para otros idiomas y al tratar de aplicarlos al español mexicano su desempeño podría degradarse por las diferencias fonéticas entre idiomas diferentes.

El experimento se realizó con los dos corpus descritos en la metodología, obteniendo tasas de reconocimiento correcto del 96.7% para el corpus UPA y 93.3% para el corpus Tlatoa.

El algoritmo propuesto se desempeña razonablemente bien, además que se demuestra que para esta clase de aplicación es superior al HMM y se compara favorablemente contra el DTW,

con lo cual se generan indicios a que, quizá en el problema de palabras aisladas el HMM pueda no ser la mejor opción.

Agradecimientos

Agradecemos al Doctorado en Ciencia y Tecnología del Centro Universitario de los Lagos de la Universidad de Guadalajara por el apoyo y soporte de esta investigación, a la Universidad Politécnica de Aguascalientes por las facilidades para su realización. El primer autor agradece a PROMEP por el apoyo otorgado mediante la beca de estudios de posgrado de calidad, la cual influyó en el desarrollo de la tesis de doctorado, de la cual una parte se presenta en este artículo. Los autores Martínez Romo y Luna Rosas, forman parte del cuerpo académico de Sistemas Inteligentes, y agradecen al Instituto Tecnológico de Aguascalientes por su apoyo en la realización de este proyecto. De igual forma agradecemos a César Andros López Luévano por su valiosa participación dentro de este artículo.

Referencias

Abdulla, Waleed H., David Chow y Gary Sin. (2003). Cross-words reference template for DTW-based speech recognition systems. TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region IEEE. (4): 1576 - 1579 Vol.4

Benzeghiba, M., De Mori, R., Deroo, O. (2007). Automatic speech recognition and speech variability: a review. Speech Communication. (49): 763-786.

Benzid, R., Marir, F., & Bouguechal, N. E. (2006). Quality-controlled compression method using wavelet transform for electrocardiogram signals. International Journal of Biomedical Sciences, 1(1), 1306-1216.

Corpus Tlatloa ®. http://info.pue.udlap.mx/~sistemas/tlatoa (20 de Julio de 2012).

Chen, B., Wei-Hau, C., Shih-Hsiang, L., Wen-Yi, C. (2011). Robust speech recognition using spatial-temporal feature distribution characteristics. Pattern Recognition Letters. (32): 919-926.

De Luna-Ortega, C.A., Mora-González, M., Martinez-Romo, J.C. (2006). Reconocimiento de voz con redes neuronales, DTW y modelos ocultos de Markov. Conciencia Tecnológica. (32): 13-17.

De Watcher, M., Matton, M., Demuynch, K., Wambacq, P., Cools, R. (2007). Template-based continuous speech recognition. IEEE Trans. on Audio, Speech, And Language Processing, 15(4): 1377-1390.

Irwin, M.J., (1988). A digit pipelined dynamic time warp processor. IEEE trans. On acoustics speech and signal processing, 36(9): 1412-1422.

Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-23: 52-72.

Kinjo, T., Funaki, K. (2006). On HMM speech recognition based on complex speech analysis. Proc. IECON 2006 – 32nd Annual Conference on IEEE Industrial Electronics, 3477-3480.

Livescu, Karen, James Glass, and Jeff Bilmes (2003). Hidden feature models for speech recognition using dynamic Bayesian networks. 8th European Conference on Speech Communication and Technology (Eurospeech), 1-4.

Makhoul, J. (1975). Linear Prediction: a tutorial review. Proc. Of the IEEE, 63(4): 561-580.

Nefian, A., Liang, L. Pi, X., Lui, X., Murphy, K. (2002). Dynamic Bayesian Networks for Audio-visual speech recognition. EURASIP Journal on Applied Signal Processing (11): 1-15.

Oropeza, R., Suarez, G. (2006). Algoritmos y métodos para el reconocimiento de voz en español mediante silabas. Computación y sistemas, 9(3): 270-286.

Proakis, J.G., Manolakis, D. (2007). Digital Signal Processing. Prentice Hall. U.S.A.

Rabiner, Lawrence R. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". Proceedings of the IEEE, 257-286.

Rabiner, L. R., & Schafer, R. W. (2007). Introduction to digital speech processing. Foundations and trends in signal processing. U.S.A.

Romo, J.C., Rosas, F.J., Mora-González, M. (2008). Combining Genetic Algorithms and FLDR for Real-Time Voice Command Recognition. Proceedings of the 2008 Seventh Mexican International Conference on Artificial Intelligence, México, 163-169.

Schafer R. W. (2007). "Homomorphic systems and cepstrum analysis of speech," Springer Handbook of Speech Processing and Communication, Springer, U.S.A.

Takiguchi, T.; Nakamura, S.; Shikano, K. (2001). HMM-separation-based recognition for a distant moving speaker. Speech and Audio Processing, IEEE Transactions on. 9(2):127-140

Trentin, E., Gori, M. (2001). A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing. (37): 91-126.

Wollmer, Martin, et al. (2010). Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. Selected Topics in Signal Processing, IEEE Journal of, 867-881.

Zweig, Geoffrey, and Stuart Russell (1999), "Probabilistic modeling with Bayesian networks for automatic speech recognition." Australian Journal of Intelligent Information Processing, 253-260.