Reconocimiento automático de palabras aisladas mediante energía a corto plazo, cruces por cero y alineamiento temporal dinámico

(1)Álvaro Anzueto-Ríos, Dr.* (1)Blanca Tovar-Corona, Dr.
(1)Yesenia E. González-Navarro, Dr.

(1) Instituto Politécnico Nacional
UPIITA
Academia de Biónica, Ciudad de México, México
aanzuetor@ipn.mx

Referencia de este artículo [1].

RESUMEN

Un sistema de reconocimiento de palabras aisladas basado en dos descriptores temporales es descrito en este trabajo y la idea forma parte de los cursos de reconocimiento de patrones; la energía a corto plazo y la tasa de cruces por cero (ZCR), y en el alineamiento temporal dinámico (*Dynamic Time Warping*, DTW) como medida de similitud entre patrones. La señal de voz se procesa por tramas no solapadas de 20 ms; las tramas cuya energía supera el 5 % del pico se conservan y las restantes se descartan como silencio. De cada trama activa se obtienen los dos descriptores que forman un vector de características; el vector concatenado se compara con prototipos de tres clases (*agua*, *café*, *jugo*) mediante DTW, esta palabras (sonidos) se han elegido por su simpleza fonética.
Sobre las base de datos, que se compone de 30 grabaciones (10 por clase), se aplico la métrica leave-one-out y se reporta una exactitud global del 80 %; en una de la pruebas, aleatorias y de manera independiente, es aplicada sobre la palabra “café” que reporta un valor de 9,23 como valor mínimo y reconocida con las propia palabra “café”. Una fortaleza a resaltar del método, es su simplicidad matemática y su bajo consuma computacional, que lo convierte en una opción pertinente para ser implementada en sistemas embebidos y para fines didácticos.

Palabras clave: reconocimiento de voz, energía a corto plazo, tasa de cruces por cero, DTW, palabras aisladas, procesamiento de audio.

ABSTRACT

A system for recognizing isolated words based on two temporal descriptors is described in this work. The concept is part of the pattern recognition curriculum, which incorporates short-term energy and zero-crossing rate (ZCR), and dynamic time warping (DTW) as a measure of pattern similarity. The speech signal is processed in non-overlapping 20 ms frames. Frames with energy exceeding 5 % of the peak are retained, while the remainder are discarded as silence. From each active frame, two descriptors are obtained, forming a feature vector. This concatenated vector is compared to prototypes of three classes (water, coffee, juice) using DTW. These words (sounds) were chosen for their phonetic simplicity. The leave-one-out metric was applied to the database, which consists of 30 recordings (10 per class), and an overall accuracy of 80 % is reported. In one of the tests, randomly and independently, the method was applied to the word “coffee”, which yielded a minimum value of $9{,}23$ and was recognized by the word “coffee” itself. A key strength of the method is its mathematical simplicity and low computational resource consumption, making it a suitable option for implementation in embedded systems and for educational purposes.

Keywords: speech recognition, short-time energy, zero-crossing rate, DTW, isolated words, audio processing.

MSC 2020: 68T10, 94A12

Introducción

Dentro del procesamiento digital de señales, la tarea del reconocimiento del habla siempre ha tenido un apartado especial, considerando que es una de las principales vías de comunicación entre las personas [1, 2]. Dentro de este campo, el reconocimiento de palabras aisladas representa el caso más simple, ya que cada locución se delimita por silencios y puede tratarse como una secuencia finita de vectores de características. Aun en ese escenario, la variabilidad acústica del hablante, la duración y la prosodia impiden comparar dos señales mediante una distancia punto a punto: los patrones que corresponden a la misma palabra rara vez tienen la misma longitud.

La energía a corto plazo y la tasa de cruces por cero (*Zero-Crossing Rate*, ZCR) son consideradas como dos técnicas comúnmente aplicadas en el procesado de señales digitales y son descriptores fáciles de calcular que aportan características ricas en detalles que describen las señales. La primera concentra información sobre la sonoridad y permite localizar los puntos de inicio y fin de la locución [5]; la segunda refleja el contenido espectral de la trama y separa sonidos sonoros de sordos [6]. Los dos métodos se aplican en tramas o ventanas fijas, lo cual tiene la ventaja de analizar las señales de manera particular y sintetizar la información especifica de cada trama que, al conjuntarlas, otorgan la información completa de la palabra.

Para cuantificar la semejanza entre dos secuencias de descriptores de los sonidos de distinta longitud, se aplica el alineamiento temporal dinámico (DTW), considerando la propiedad de que está técnica puede ser programada de manera dinámica; fue introducida para el reconocimiento del habla a finales de los años setenta [3, 4] y actualmente sigue siendo una referencia por su interpretabilidad y bajo costo, sobre todo en sistemas con corpus pequeños o en dispositivos con recursos limitados [7, 8].

Considerando este enfoque, en el presente trabajo, se desarrolla un reconocedor de palabras aisladas que cumple con los siguientes puntos: (i) detecta la actividad de voz por umbral de energía, (ii) extrae por ventana los descriptores de energía y ZCR, (iii) normaliza los vectores y los compara con prototipos mediante el calculo de semejanza por DTW, y (iv) determina la clase del prototipo más cercano. La validación se realiza con una base de tres palabras del vocabulario cotidiano de una persona —*agua*, *café* y *jugo*, las muestras fueron grabadas en condiciones semi-controladas dentro de una habitación.

Marco Teórico

La matemática que sustenta el sistema de reconocimiento de voz propuesto, se expresa y analiza en esta sección. Como primer paso, se formaliza la representación discreta de la señal de voz, la cual, es normalizada para tener estandarizado los niveles de sonido, luego se aplica la descomposición en tramas o ventanas de la señal completa.

Posteriormente, se abordan los dos métodos empleados para formar el vector de características; energía a corto plazo y tasa de cruces por cero. Como proceso final, se revisan las partes que conforman el algoritmo de alineamiento temporal dinámico (DTW), responsable de comparar dos secuencias de descriptores de longitud distinta que corresponden a dos palabras.

Señal de voz discreta y ventaneo

Una señal de voz discreta $s[n]$, $n = 0, 1, \ldots, L-1$, se cuantifica por toma de muestras equidistantes a una frecuencia F_s. Para extraer información de manera particular, la señal se divide en tramas o ventanas $w_i[k]$ de $N = \lfloor T\, F_s \rfloor$ muestras, con $T = 20$ ms y con tramas consecutivas, es decir, sin solapamiento:

$$w_i[k] = s[iN + k], \qquad k = 0, 1, \ldots, N-1. \tag{1}$$

Energía a corto plazo

Los datos de la trama i-ésima, son procesados aplicando la suma de los cuadrados para obtener el valor de la energía a corto plazo de la señal, en dicha trama [9]:

$$E_i = \sum_{k=0}^{N-1} w_i[k]^2. \tag{2}$$

El vector $\{E_i\}$ es procesado para, determinar si la trama contiene información que corresponde a sonido o si se trata de una sección de silencio, mediante un umbral relativo al pico:

$$\theta_E = \alpha \cdot \max_i E_i, \qquad \alpha = 0{,}05. \tag{3}$$

Tasa de cruces por cero (ZCR)

Se calcula el número de veces de los cambios de signo de la señal dentro de la ventana [6] de procesado y este valor constituye la tasa de cruce por cero:

$$Z_i = \tfrac{1}{2} \sum_{k=1}^{N-1} \left|\operatorname{sgn}(w_i[k]) - \operatorname{sgn}(w_i[k-1])\right|. \tag{4}$$

La energía aporta una medida de sonoridad y la ZCR una medida indirecta de la frecuencia dominante de la trama; el vector de características que se emplea en este trabajo es por lo tanto $\mathbf{x}_i = [E_i,\ Z_i]^\top$.

Normalización por columna

Dado que el calculo de E_i y Z_i genera rangos numéricos diferentes, es necesario estandarizarlos empleando su propia media y desviación estándar para luego ser presentados a DTW:

$$\tilde{x}_i^{(j)} = \frac{x_i^{(j)} - \mu_j}{\sigma_j + \varepsilon}, \qquad j \in \{E,\ Z\}, \tag{5}$$

con $\varepsilon = 10^{-8}$ para evitar divisiones por cero.

Alineamiento temporal dinámico (DTW)

Dadas dos secuencias de vectores $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$ y $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, el algoritmo DTW construye una matriz de costos acumulados $D \in \mathbb{R}^{(M+1)\times(N+1)}$ mediante la recurrencia:

$$D(i,j) = \|\mathbf{x}_i - \mathbf{y}_j\|_2 + \min\big\{D(i-1,j),\ D(i,j-1),\ D(i-1,j-1)\big\}, \tag{6}$$

con $D(0,0) = 0$ y $D(i,0) = D(0,j) = \infty$ para $i, j > 0$ [3,7]. La distancia DTW se define como $D_{\text{DTW}}(\mathbf{X}, \mathbf{Y}) = D(M,N)$ y el camino óptimo se obtiene por *backtracking* desde (M,N) hasta $(1,1)$.

Metodología Propuesta

Corresponde finalmente describir la secuencia ordenada de pasos que conforman el sistema de reconocimiento de sonidos; la secuencia esta organizada en tres bloques; el primero es la arquitectura general, que encadena pre-procesado, detección de voz, extracción de características y decisión por vecino más cercano. El segundo bloque corresponde a la descripción de la base de datos empleada, indispensable para reproducir el experimento. y finalmente el bloque de las métricas con las que se cuantifica el desempeño global del clasificador.

Arquitectura del Sistema

Para presentar la idea general de todo el proceso se organiza en cuatro etapas, que se aplican tanto, a las grabaciones de entrenamiento como a la señal de prueba, estas se enumeran a continuación:

1. **Pre-procesado:** la grabación de la señal de voz debe ser verificada si se encuentra en modo estéreo o mono aural, el sistema considera unicamente el modo mono aural y su normalización dentro del rango $[-1, 1]$.
2. **Detección de actividad de voz:** se calcula la energía por trama mediante (2) y se conservan únicamente las tramas que superan el umbral (3), lo cual indica que la trama contiene información de audio (no silencio).

Tabla 1: Parámetros del sistema y de la base de datos.

Parámetro	**Valor**
Frecuencia de muestreo, F_s	44 100 Hz
Duración de ventana, T	20 ms (882 muestras)
Solapamiento entre ventanas	0 %
Umbral de silencio, α	0.05
Normalización	z-score por columna
Métrica de comparación	DTW (norma ℓ_2)
Clases	agua, café, jugo
Grabaciones por clase	10

3. **Extracción de características:** por cada trama activa se calculan E_i y Z_i; la matriz resultante se estandariza por columna con (5).
4. **Decisión por vecino más cercano:** se calcula la distancia DTW (6) entre la secuencia de prueba y los prototipos; se asigna la clase del prototipo más cercano.

Base de datos

La base de datos empleada contiene tres clases (*agua*, *café* y *jugo*) con diez grabaciones por clase, capturadas a $F_s = 44\,100$ Hz en formato PCM mono de 16 bits. Cada grabación dura aproximadamente dos segundos y se almacena en una subcarpeta nombrada por la clase. La señal de prueba se mantiene en una carpeta independiente y no participa en la construcción de los prototipos.

Métricas de evaluación

Para evaluar el desempeño global del sistema se emplean dos indicadores: la exactitud por validación cruzada *leave-one-out* (LOO) y la matriz de confusión. En LOO cada uno de los 30 patrones se utiliza, por turnos, como consulta mientras los 29 restantes funcionan como prototipos; la exactitud se define como:

$$\text{Exactitud} = \frac{\#\text{aciertos}}{N_{\text{total}}}. \tag{7}$$

Validación Experimental

Con el objetivo de comprobar el funcionamiento del reconocedor, en esta sección se reporta la evaluación experimental realizada sobre la base de datos descrita en la metodología. Primero se detalla la configuración numérica del experimento; posteriormente se ilustra el proceso de detección de voz sobre una grabación representativa, se comparan los descriptores entre las tres clases consideradas y se analiza la separabilidad mediante alineamientos DTW intra-clase e inter-clase; finalmente, se cuantifica el desempeño global con la validación cruzada *leave-one-out*.

Configuración Experimental

Los parámetros del experimento se resumen en la Tabla 1. Se implementan en Python 3 con `numpy` y `scipy`; las funciones de extracción y DTW se programan sin recurrir a paquetes externos de reconocimiento de voz, con el fin de mantener la transparencia algorítmica.

Detección de actividad de voz

La Figura 1 ilustra la detección de actividad sobre la grabación `agua (1).wav`. En el panel (a) se muestra el oscilograma de la señal completa: las tramas activas (en verde) y las silenciosas (en gris) se distinguen mediante sombreado, lo que permite visualizar la duración relativa de las regiones útiles dentro de la grabación. En el panel (b) se grafican simultáneamente, sobre ejes gemelos, la energía a corto plazo

E_i (eje izquierdo) y la tasa de cruces por cero Z_i (eje derecho); la línea discontinua marca el umbral adaptativo $\theta_E = \alpha \ \max E$ y la banda verde resalta el intervalo de tramas que lo superan. Finalmente, en el panel (c) se incluye el espectrograma de la señal, que permite contrastar la detección por umbral con la energía espectral concentrada en la locución. Para esta grabación, la duración efectiva se reduce de 2.00 s a 0.22 s y solamente 11 tramas superan $\theta_E = 6{,}78$.

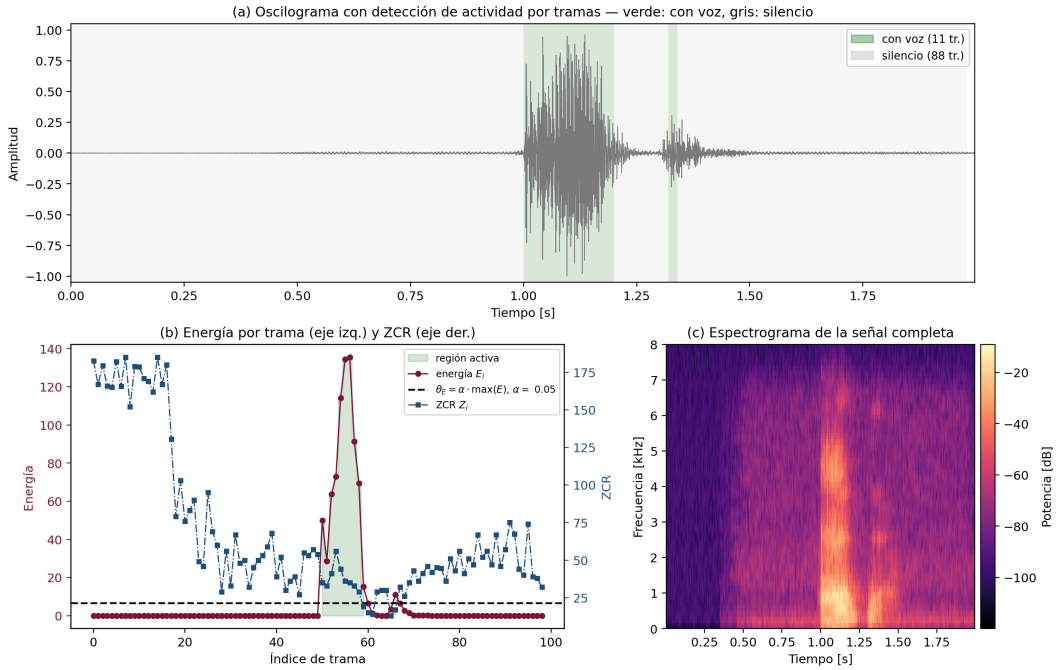

Figura 1: Detección de actividad de voz para `agua (1).wav`: (a) oscilograma con tramas activas (verde) y silencios (gris) coloreados sobre la propia señal; (b) energía a corto plazo E_i (eje izquierdo, guinda) y tasa de cruces por cero Z_i (eje derecho, azul) por trama, con umbral θ_E; (c) espectrograma de la señal completa.

Comparación de descriptores entre clases

La Figura 2 resume el comportamiento de los dos descriptores temporales para las tres clases. Los paneles (a) y (b) trazan, respectivamente, la energía y la ZCR por trama para una grabación representativa de cada clase, mientras que el panel (c) proyecta el conjunto completo de la base en el plano $(\overline{E}, \overline{Z})$ de promedios por grabación. Las tres palabras se distinguen tanto en la duración total como en la forma temporal del descriptor, y la dispersión 2D revela que cada clase ocupa una región diferenciada del plano de características.

Alineamiento DTW intra e inter-clase

Para mostrar la capacidad discriminativa de DTW se comparan dos grabaciones de la misma clase (*agua* 1 vs *agua* 7) y dos de clases distintas (*agua* 1 vs *café* 1). La Figura 3 reúne, en los paneles superiores, las matrices acumuladas DTW con el camino óptimo superpuesto, y en los paneles inferiores las secuencias de energía alineadas mediante las correspondencias indicadas por dicho camino. La distancia intra-clase es $D = 6{,}49$, mientras que la inter-clase asciende a $D = 19{,}13$. El factor cercano a tres entre ambas distancias justifica el uso de DTW como medida de similitud entre patrones.

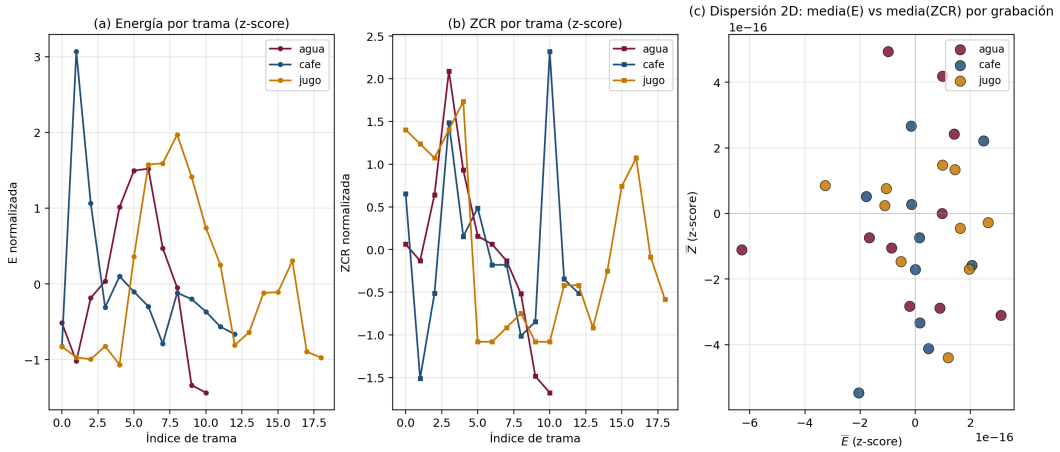

Figura 2: Descriptores por clase: (a) energía por trama y (b) tasa de cruces por cero por trama (z-score, una grabación por clase); (c) dispersión $(\overline{E}, \overline{Z})$ de las 30 grabaciones de la base.

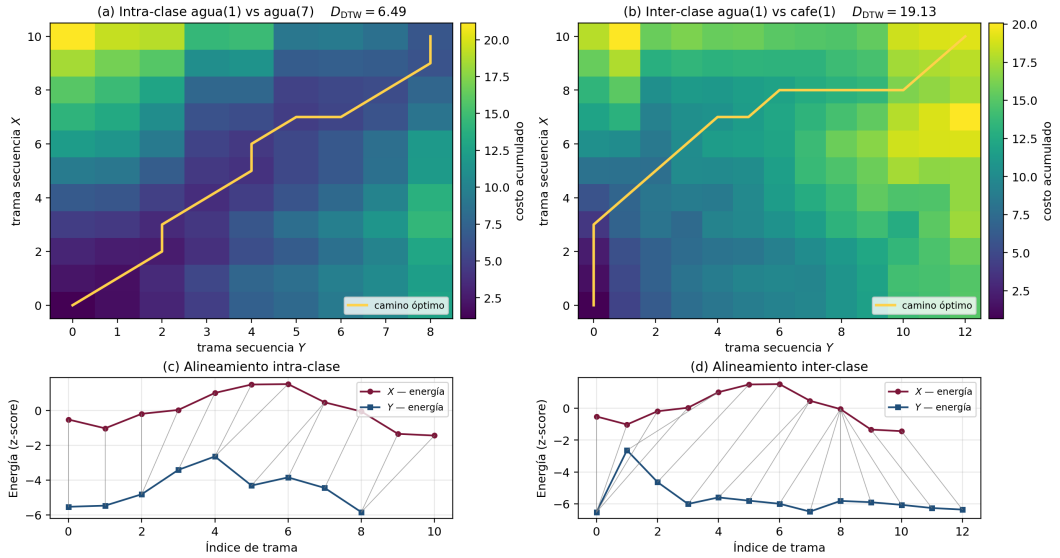

Figura 3: Comparación DTW intra/inter-clase. Paneles superiores: matriz de costos acumulados con camino óptimo, (a) intra-clase ($D = 6{,}49$) y (b) inter-clase ($D = 19{,}13$). Paneles inferiores: secuencias de energía alineadas según el camino óptimo.

Reconocimiento de la señal de prueba

La señal independiente `prueba (1).wav` se examina aplicando la secuencia de etapas, descritas previamente y se compara con los 30 prototipos restantes. La Tabla 2 presenta el resume de los resultado obtenidos.

La Figura 4 resume el resultado del reconocimiento. En el panel (a) se representan las 30 distancias DTW entre la señal de prueba y cada prototipo, ordenadas de menor a mayor y coloreadas por clase; las etiquetas marcan los tres prototipos más cercanos. En el panel (b) se reporta la distancia promedio por clase con su desviación estándar. El mínimo absoluto pertenece a un prototipo de la clase *café*, lo que determina la decisión final del sistema.

Tabla 2: Resultado del reconocimiento sobre la señal de prueba `prueba (1).wav`.

Métrica	**Valor**
Número de tramas activas	13
Distancia DTW mínima	9.23
D_{DTW} promedio a agua	17.58
D_{DTW} promedio a café	14.90
D_{DTW} promedio a jugo	12.40
Palabra reconocida	**café**

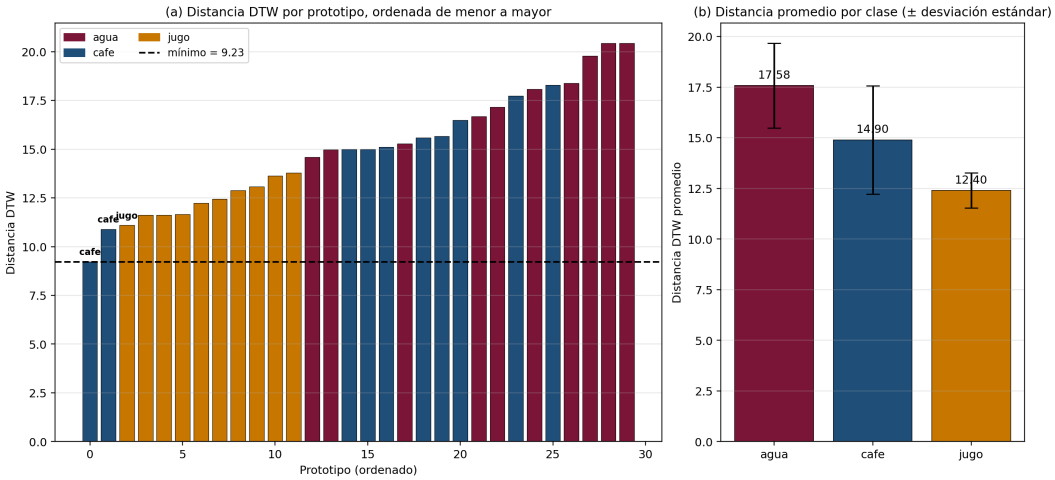

Figura 4: Reconocimiento de la señal de prueba: (a) distancias DTW a cada prototipo ordenadas de menor a mayor (los tres prototipos más cercanos se etiquetan con la clase a la que pertenecen); (b) distancia DTW promedio por clase con barras de desviación estándar.

Evaluación global (leave-one-out)

La validación cruzada, LOO (por sus siglas en inglés: leave-one-out) es aplicada a las muestras de que corresponden a la base de datos, para nuestro caso, 30 grabaciones. La Tabla 3 y la Figura 5 exponen los resultado: la exactitud global alcanza el 80 % (24 aciertos sobre los 30 patrones considerados), la matriz de confusión demuestra el equilibrio que presentan las tres clases. La precisión por clase es de 0.80 para *agua*, 0.88 para *café* y 0.75 para *jugo*; la exhaustividad logra alcanzar los valores de 0.80, 0.70 y 0.90, respectivamente.

Discusión

Los resultados muestran que la concatenación de energía a corto plazo y tasa de cruces por cero; aunque modesta desde el punto de vista espectral, contiene información relevante para discriminar palabras breves cuya envolvente temporal difiere de manera apreciable. La diferencia entre la distancia DTW intra-clase ($D = 6{,}49$) y la inter-clase ($D = 19{,}13$) confirma la separabilidad de las tres clases en el espacio bidimensional propuesto. En la evaluación LOO la clase *jugo* resulta la más distinguible, mientras que la mayoría de los errores se concentra en el par *café*/*jugo*, ambos con una estructura fonética CV-CV similar.

El sistema mantiene una complejidad computacional baja: considerando que DTW es $\mathcal{O}(MN)$ y que las secuencias de características rondan las $M \approx 14$ ventanas, el cálculo de las 30 distancias ronda en los valores de milisegundos, considernado para su implementación: una computadora con un procesador intel i5 y con 16 gigas en RAM. Esta característica resulta especialmente relevante para la enseñanza del procesamiento de voz y para implementaciones en sistemas de hardware de bajos recursos; en contraste, las alternativas basadas en coeficientes *mel-cepstrum* y redes profundas requieren bases de entrenamiento más grandes y recursos de cómputo considerablemente mayores.

Tabla 3: Matriz de confusión de la validación leave-one-out (30 patrones).

	Pred. agua	Pred. café	Pred. jugo	Total
Real agua	8	1	1	10
Real café	1	7	2	10
Real jugo	1	0	9	10
Exactitud	80.0 % (24/30)			

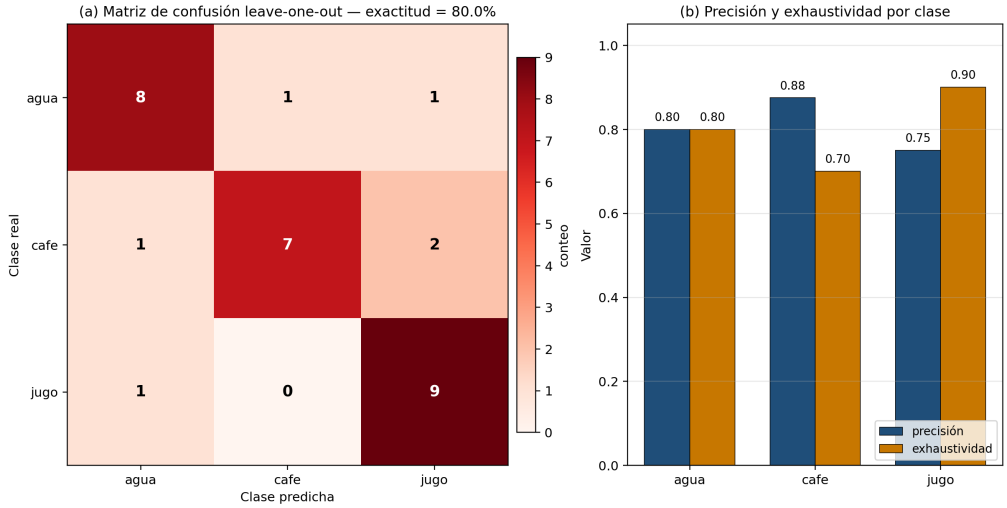

Figura 5: Validación leave-one-out: (a) matriz de confusión 3×3, exactitud global 80 %; (b) precisión y exhaustividad por clase.

Entre las limitaciones que podemos mencionar del sistema propuesto se encuentra la dependencia del umbral α frente al ruido que puede estar presente en el ambiente. Además, del uso de un único prototipo por audio (sin promediado por clase) y la ausencia de restricciones de banda en DTW (*Sakoe-Chiba band*), que logran reducir, tanto el costo como la sensibilidad a desalineamientos de las secuencias de datos.

Conclusión

Un sistema de reconocimiento de palabras aisladas se ha examinado en este trabajo, el cual desarrolla tres bloques clásicos del procesamiento de voz: detección de actividad por energía, que ha logrado reconocer tramas con sonido y con silencios. Después, se calcula tanto la energía de la señal en estado temporal como el número de veces que cruza por el nivel cero (línea base de la señal, ZCR) y la métrica de semejanza entre secuencias mediante DTW, de donde se han extraído las siguientes contribuciones:

1. Se ha conjuntado un detector de actividad sonora y silencios por umbral adaptativo con una representación bidimensional por trama o ventana de información, lo que redujo, de manera significativa, la duración efectiva y análisis de la señal —de 2.00 s a 0.22 s en el ejemplo citado; sin perder la estructura discriminativa de la palabra.
2. Se ha verificado, sobre una base de 30 grabaciones, que representan a tres clases, que el calculo del alineamiento y semejanza de secuencias de datos por DTW logra separar y clasificar, de manera apropiada, las clases de cada muestra: la distancia intra-clase es del orden de un tercio de la inter-clase, y la valoración leave-one-out alcanza el valor de 80 % en la métrica exactitud.
3. El sistema se mantiene simple desde el punto de vista matemático y con bajo consumo computacional, lo que demuestra y garantiza, tanto, su uso didáctico como su implementación en plataformas con recursos limitados (hardware embebido). Como perspectivas de este trabajo

se contempla ampliar el vocabulario, incorporar coeficientes *mel-cepstrales* y acotar el camino DTW con la banda de Sakoe–Chiba para reducir aún más el costo.

Referencias

[1] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[2] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ: Prentice Hall, 2001.

[3] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[4] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.

[5] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[6] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proceedings of the IEEE*, vol. 74, no. 11, pp. 1477–1493, 1986.

[7] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*, Berlin: Springer, 2007, pp. 69–84.

[8] S. Salvador and P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[9] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York: IEEE Press, 2000.

Referencias

[1] Anzueto-Ríos, A., Tovar-Corona, B., González-Navarro, Y. E. ***(julio - agosto, 2026)*** *Reconocimiento automático de palabras aisladas mediante energía a corto plazo, cruces por cero y alineamiento temporal dinámico. Boletín UPIITA. año 20, (113) 2026.*
https://www.boletin.upiita.ipn.mx/index.php/ciencia/1105-cyt-numero-113