Wavelets en el Reconocimiento de voz

Ismael Chávez, *Non Member, IEEE*, Antonio Camarena-Ibarrola, *Non Member, IEEE*

***Resumen*—El esquema tradicional utilizado en sistemas de reconocimiento de voz por varias décadas ha consistido en llevar a cabo un procesamiento basado en la transformada de Fourier de tiempo corto. Hace algunos años se introdujo un esquema basado en ondas de corta duración y localizadas en el tiempo denominadas "wavelets". Hemos implementado un identificador de palabras aisladas basado en wavelets y lo hemos probado con diferentes funciones wavelet con muy buenos resultados**

***Temas claves*— Wavelets, Daubechies, Haar, Reconocimiento de voz.**

I. Introducción

Para llevar a cabo el reconocimiento de voz debemos realizar un análisis espectral dinámico, esto es, determinar las frecuencias que le dan forma a la señal de voz (formantes) pero también saber en qué instante ocurren, es decir, necesitamos información tanto del dominio del tiempo como del dominio de la frecuencia [1]. El enfoque tradicional para logar esto usa de la denominada Transformada de Fourier de Tiempo Corto (STFT por sus siglas en inglés), este método consiste en dividir la señal de voz en segmentos de corta duración denominados marcos, con una longitud típica de 30 ms, con un traslape entre marcos de 20 ms. A cada marco se le aplica una ventana como la de Hann para que la señal correspondiente a un marco se desvanezca en sus extremos y de esta manera disminuir un efecto indeseable conocido como "escurrimiento" [2]. El enfoque de la STFT está sujeto al principio de incertidumbre de Heisenberg que nos dice que hay un compromiso entre la resolución en tiempo y la resolución en frecuencia [3]. Por ejemplo, en la Fig. 1 (centro) se incrementó el tamaño del marco para aumentar la resolución en frecuencia sacrificando la resolución en el tiempo, en la misma figura (derecha) se redujo el tamaño del marco para incrementar la resolución en el tiempo sacrificando la resolución en frecuencia.

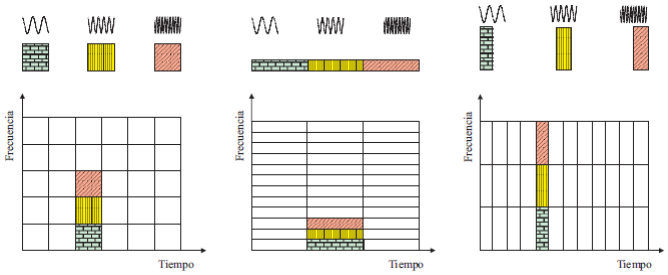

Fig. 1. Compromiso entre la resolución en frecuencia y la resolución en tiempo al usar la STFT

En realidad no necesitamos la misma resolución en tiempo para todas las bandas de frecuencia, en las frecuencias bajas los cambios son lentos y no requerimos mucho detalle relativo a los cambios temporales, por el contrario, para frecuencias altas donde los cambios son rápidos se necesita mayor resolución en tiempo. La idea de tener diferentes resoluciones en el tiempo para diferentes frecuencias se conoce como "multiresolución" y es uno de los aspectos clave de la teoría de wavelets [3]. En la Fig. 2 se ilustra este concepto

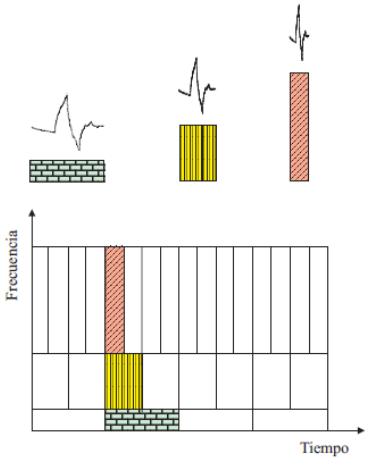

Fig. 2. Multiresolución uno de los aspectos clave en la teoría de wavelets

Los wavelets han tenido gran éxito donde las señales sufren cambios bruscos puesto que estos son difíciles de reproducir con senoides de duración infinita (análisis de Fourier). Por ejemplo, para comprimir imágenes los wavelets son muy utilizados (particularmente para huellas dactilares). La señal de voz también sufre de cambios bruscos, sobre todo en sonidos no vocalizados y fricativos [4]. En general, cuando una señal no es estacionaria la transformada de Fourier no la puede caracterizar adecuadamente. La transformada Wavelet en cambio parece funcionar mejor mientras más caótica

I. Chávez labora en la Universidad Michoacana de San Nicolás de Hidalgo. Facultad de Ingeniería Eléctrica (e-mail: ichavez@umich.mx).

A. Camarena-Ibarrola labora en la Universidad Michoacana de San Nicolás de Hidalgo. Facultad de Ingeniería Eléctrica. División de estudios de Postgrado (email: camarena@umich.mx)

ISBN: 978-607-95476-3-9

parezca ser la señal bajo análisis [5].

A diferencia de la STFT, la Transformada Discreta Wavelet (DWT por sus siglas en inglés) no requiere de dividir la señal en marcos de tiempo ni de aplicación de ventanas como la de Hann, Hamming, Parzen o cualquiera de las que se usan cuando se opta por la STFT [2]. La DWT fue desde su origen diseñada para proporcionar información del dominio del tiempo y de la frecuencia a la vez. Las versiones del wavelet a menor escala tienen menor duración y por tanto se encargan de reproducir los componentes de alta frecuencia de la señal bajo. De manera similar, las versiones del wavelet de mayor escala son de duración mayor y reproducen los componentes de frecuencia baja. En la teoría wavelet en lugar de la frecuencia se utiliza la variable escala.

A. Wavelets

Los wavelets son ondas de corta duración y localizadas en el tiempo [3], a diferencia del análisis de Fourier donde una señal se descompone en una serie de senoides de diferentes frecuencias, todas de duración infinita, el análisis basado en wavelets descompone la señal en una serie de versiones escaladas y desplazadas del "wavelet madre". En la parte superior de la Fig. 2 se aprecian tres diferentes versiones escaladas del wavelet de Daubechies de segundo orden.

Una función $f(t)$ (una señal) se puede expresar como una combinación de funciones base (wavelets) $w_{j,k}(t)$ mediante (1)

$$f(t) = \sum_{j,k} b_{j,k} w_{j,k}(t) \qquad (1)$$

donde es $b_{j,k}$ es el coeficiente que pondera a la función base $w_{j,k}(t)$ que no es otra cosa que el wavelet en la escala j y desplazamiento k, de acuerdo a (2)

$$w_{j,k}(t) = w(2^j t - k) \qquad (2)$$

La codificación en sub-bandas de doble canal constituye un mecanismo eficiente para la implantación de la Transformada Wavelet Discreta usando una pareja de filtros, un filtro pasa-bajas y un filtro pasa-altas. La salida del filtro pasa-bajas es una versión suavizada de la señal de entrada a la que llamamos "aproximación", sus componentes son por esta razón denominados coeficientes de aproximación (cA). La salida del filtro pasa-altas se interpreta como el "detalle" de la señal original y a sus componentes se les denomina coeficientes de detalle (cD) [3]. De acuerdo a la regla de Nyquist [2], tanto la salida del filtro pasa-bajas como la del filtro pasa-altas tienen redundancia, debemos entonces hacer un sub-muestreo que consiste en eliminar a todos los componentes impares de la señal, de otra manera incrementaríamos al doble las necesidades de almacenamiento, ver Fig. 3.

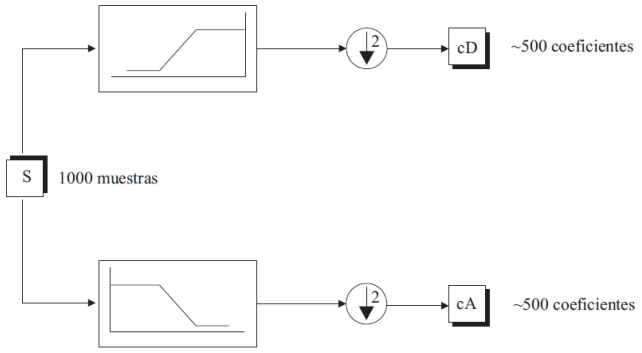

Fig. 3. Descomposición de una señal S en aproximaciones y detalles

.

La aplicación de esta pareja de filtros a la señal original produce los coeficientes de aproximación y de detalle del primer nivel. El filtro pasa-bajas y el pasa-altas se aplican a los coeficientes de aproximación del primer nivel para producir los coeficientes de aproximación y de detalle del segundo nivel respectivamente, el proceso se repite con los coeficientes de aproximación del segundo nivel y así sucesivamente como se muestra en la Fig. 4

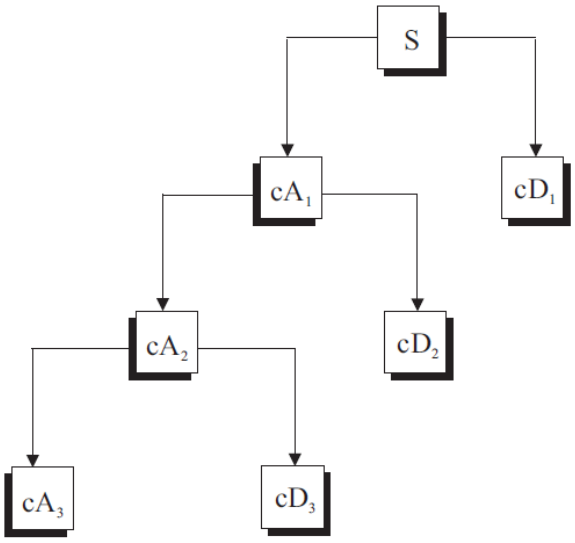

Fig. 4. Descomposición multinivel de una señal S

En el caso del wavelet de Haar, los coeficientes del filtro pasa-bajas son $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ y los del filtro pasa-altas son $\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$ [3].

El algoritmo piramidal o de Mallat [3] determina los coeficientes de aproximación y de detalle del nivel j-1 a partir de los coeficientes de aproximación del nivel j en base a la recurrencia dada en (3) y (4).

$$a_{j-1,k} = \frac{1}{\sqrt{2}} a_{j,2k} + \frac{1}{\sqrt{2}} a_{j,2k+1} \qquad (3)$$

ISBN: 978-607-95476-3-9

$$b_{j-1,k} = \frac{1}{\sqrt{2}} a_{j,2k} - \frac{1}{\sqrt{2}} a_{j,2k+1} \qquad (4)$$

En la Fig. 5, se muestra el algoritmo de Mallat para el wavelet de Haar, el algoritmo recibe la señal de la que va a extraer los coeficientes de aproximación y los de detalle, también recibe el número de niveles requerido [6].

```
TRANSFORMADAWAVELETHAAR(señal, nivel)
1  coeficientesACalcular ← longitudDeSeñal/2
2  l0 ← l1 ← 1/√2
3  para i = 0 hasta i < nivel
4    para j = 0 hasta j < coeficientesACalcular
5      aproximacion[i][j] ← (señal[j*2]*l0 + señal[j*2+1]*l1)
6      detalle[i][j] ← (señal[j*2]*−l0 + señal[j*2+1]*l1)
7    señal ← aproximacion
8    coeficientesACalcular ← coeficientesACalcular/2
```

Fig. 5. Algoritmo de Mallat usando el wavelet de Haar

Los coeficientes del filtro pasa-bajas del wavelets de Daubechies de segundo orden son $\left(\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}}\right)$ [7]. Los coeficientes de filtros relacionados a otros wavelets pueden obtenerse de manera simple con Matlab, ver [8].

II. Sistema implementado

Se implementó un reconocedor de palabras aisladas que consiste de los siguientes módulos:

A. Captura de la señal de audio

La señal de audio proveniente del micrófono es digitalizada con una frecuencia de muestreo de 8KHz, con una precisión de 8 bits por muestra sin signo y por un solo canal (monoaural). Se utilizó el API JavaSound [9] para llevar a cabo la lectura de la señal digitalizada.

B. Segmentación

Desde el momento en que empieza a capturarse la señal de audio hasta que el usuario del sistema realmente articula una palabra transcurre un tiempo en el que se grabó básicamente ruido ambiental, de igual manera, desde que finaliza la elocución de la palabra hasta que se termina la captura de la señal de audio se graba una señal sin contenido de voz, es pues necesario encontrar el inicio y el final de la elocución. Para ello, la señal fue procesada en marcos de tiempo de 10 ms sin traslape (80 muestras en nuestro caso). El oído humano no puede detectar cambios que ocurran en menos de 10 ms [1]. Por cada marco de tiempo se determinó el régimen de cruces por cero y la energía de la señal. Si la energía y el régimen de cruces por cero rebasan un valor umbral consistentemente por varios marcos podemos estar seguros que la elocución de la palabra comenzó en el primero de los marcos donde estas características rebasaron el umbral. El final de la elocución se determina de igual manera pero partiendo del final de la grabación hacia atrás.

C. Extracción de características

Se implantó el algoritmo piramidal de Mallat [3] para calcular la Transformada Discreta Wavelet, utilizando la función base de Haar y las funciones base de Daubechies [7] de segundo a décimo orden. La validación del correcto funcionamiento de los algoritmos correspondientes a estos módulos del sistema de reconocimiento se probó exhaustivamente, y se cotejaron los resultados obtenidos de su uso con los que producen los códigos disponibles en Matlab [8] a partir del programa propio escrito en Java, encontrándose coincidencia en todos los casos considerados. Se cuenta entonces con un módulo de caracterización de señales confiable que permite construir los vectores de Coeficientes Wavelet de aproximación en el sistema autónomo que se propone, de manera que no se depende de funciones implantadas en aplicaciones de terceros.

D. Comparación

Dos elocuciones de la misma palabra son normalmente de duración diferente (considere estas duraciones en milisegundos para convencerse). Después de la extracción de características las palabras son sustituidas por secuencias de coeficientes de aproximación. Debemos comparar dos secuencias de longitud distinta para establecer que tan diferentes son y concluir si en efecto proceden de dos elocuciones de la misma palabra o de dos palabras distintas. El doblado dinámico en tiempo [10]-[11] (DTW por sus siglas en inglés) soluciona el problema de descubrir la función de doblado óptima que establece cuales coeficientes de una secuencia se deben comparar con cuales coeficientes de la otra secuencia, esto se hace para establecer que tan parecidas serían las dos secuencias bajo comparación si estas fueran alineadas de manera óptima. La distancia entre la secuencia x de longitud N y la secuencia y de longitud M se determina mediante DTW que puede ser implementado de manera eficiente mediante programación dinámica en base a la siguiente recurrencia:

$$D[0][0] = 0$$

$$D[i][0] = D[i-1][0] + |x[i] - y[0]| \quad \forall i \le N$$

$$D[0][j] = D[0][j-1] + |x[0] - y[j]| \quad \forall j \le M$$

$$D[i][j] = \begin{cases} D[i-1][j-1] + |2x[i] - y[0]| \\ D[i-1][j] + |x[i] - y[0]| \\ D[i][j-1] + |x[i] - y[0]| \end{cases} \quad \forall i \le N, j \le N$$

D es un arreglo auxiliar de tamaño $N \times M$,la distancia normalizada entre las dos secuencias se determina mediante (5)

ISBN: 978-607-95476-3-9

$$d(x,y) = \frac{D[N-1][M-1]}{N+M} \qquad (5)$$

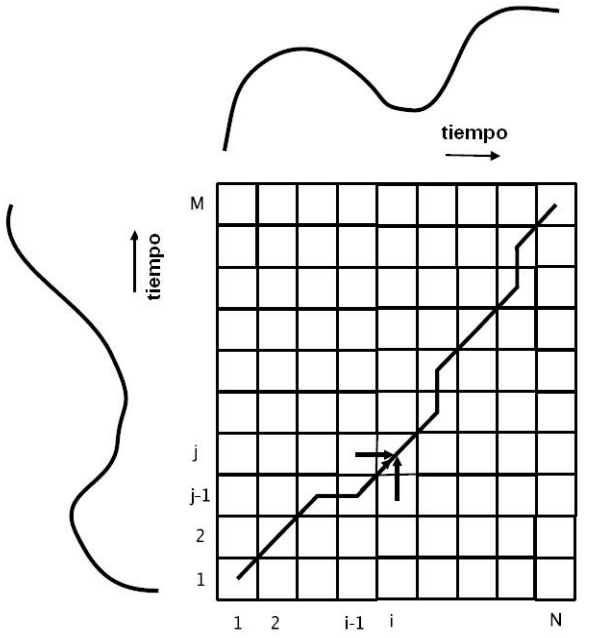

Fig. 3. Descomposición de una señal S en aproximaciones y detalles

E. Clasificación

Para realizar el reconocimiento de una palabra, esta es buscada en el diccionario de manera secuencial y se decide mediante el criterio de los K-vecinos [12] el cual es un esquema de votación. Para implementar este método se debe contar con varias elocuciones de cada una de las palabras incluidas en el diccionario.

III. Experimentos

Se realizaron pruebas al sistema identificador de palabras aisladas cambiando la función wavelet. Se probó con el wavelet de Haar, y los wavelets de Daubechies de órdenes 2 al 10. El diccionario consistió de los diez dígitos, por cada entrada del diccionario se tenían varias elocuciones para poder aplicar el esquema de reconocimiento de los K-vecinos. Para evaluar al sistema se realizó un análisis de sensibilidad descrito a continuación.

A. Análisis de sensibilidad.

Cuando la distancia entre dos elocuciones de la misma palabra es menor que un cierto valor umbral decimos que estamos en presencia de un positivo verdadero, si en cambio, la distancia entre las dos elocuciones de la misma palabra es mayor que dicho umbral, entonces calificamos al resultado de la comparación como un negativo falso. Por otra parte, si estamos comparando palabras distintas y la distancia entre ellas es mayor que el valor umbral designamos al resultado como un negativo verdadero. Finalmente si al comparar dos palabras distintas resulta que la distancia entre ellas es menor que el valor umbral, sentenciamos como positivo falso al resultado. La Tabla I resume estas definiciones.

TABLA I
DEFINICIONES PARA EL ANÁLISIS DE SENSIBILIDAD

Señales bajo comparación	Distancia menor que el umbral	Distancia mayor que el umbral
Dos elocuciones de la misma palabra	Positivo verdadero (PV)	Negativo falso (NF)
Dos palabras distintas	Positivo falso (PF)	Negativo verdadero (NV)

El régimen de predicción verdadero (TPR por sus siglas en inglés) es la fracción de palabras que el sistema identificó correctamente (positivos verdaderos) entre las que debería haber identificado, al TPR se le conoce también como sensibilidad y se determina mediante (6)

$$TPR = \frac{PV}{PV + NF} \qquad (6)$$

El TPR también es equivalente a 1-FRR, donde FRR es el régimen de rechazos falsos (por sus siglas en inglés).

Por otra parte, el régimen de predicciones falsas (FPR por sus siglas en inglés) es una medida de la frecuencia con la que el sistema se equivoca confundiendo una palabra con otra. El FPR es conocido también como régimen de falsas alarmas y es equivalente a 1-especificidad. El FPR se determina mediante (7)

$$FPR = \frac{PF}{PF + NV} \qquad (7)$$

Si usamos al TPR como eje vertical y al FPR como eje horizontal formamos al plano ROC, estas siglas provienen de Receiver-Operator Characteristics debido a que originalmente se diseñó para evaluar radares aunque ahora se usa para evaluar cualquier tipo de clasificadores. Un solo punto en el plano ROC es una indicación de que tan bien funciona un reconocedor para cierto valor umbral. Variando dicho valor umbral se genera una curva ROC [13].

B. Resultados

En la Fig. 6 se muestran las curvas ROC obtenidas para los sistemas reconocedores de palabras aisladas basados en el

ISBN: 978-607-95476-3-9

wavelet de Haar y en los wavelets de Daubechies del primero al décimo orden. Observamos de las curvas producidas por nuestros ensayos que el wavelet de Haar y el de Daubechies de tercer orden obtuvieron el mejor desempeño, esto se concluye al analizar la Fig. 6 teniendo en mente que el clasificador ideal es aquel que se asemeja más a un escalón que a un FPR de cero ya tiene un TPR de uno y se sostiene TPR en uno al ir el FPR de cero a uno.

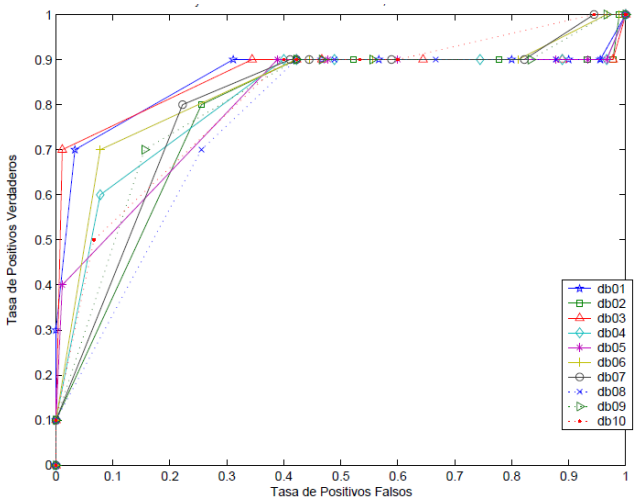

Fig. 7. Curvas ROC de los reconocedores de voz basados en diversos wavelets en nivel 5

IV. Conclusiones y Trabajos Futuros

Se ha implementado un sistema de identificación de palabras aisladas basado en wavelets con desempeño similar al de sistemas basados en coeficientes LPC (Linear Predictive Coding) o en MFCC (Mel-Frequency Cepstral Coeficients). Nuestro sistema no realiza análisis de tiempo corto como en el caso de los sistemas tradicionales. En lugar de comparar palabras usando escalagramas completos, nuestro sistema utiliza solamente un nivel para realizar el reconocimiento, en este sentido el nivel 4 resultó ser el más adecuado para nuestros propósitos. Para efectos de aumentar aún más la tasa de reconocimiento de nuestro sistema llevaremos a cabo en breve una mejora simple consistente en involucrar varios niveles de los coeficientes de detalle en la comparación de las elocuciones en lugar de un solo nivel de coeficentes de aproximación que fué lo que se realizó en este trabajo.

V. Referencias

[1] L. Rabiner, R. Schafer, *Digital Processing of Speech Signals*, Ed Prentice Hall. 1978.

[2] J. Proakis, D. Manolakis. *Digital Signal Processing 4th Edition.* Ed. Prentice Hall. 2006

[3] G. Strang, T. Nguyen. *Wavelets and Filter Banks*, Ed. Wellesley-Cambridge Press. 1996.

[4] P. Lieberman, S. Blumstein. *Speech physiology, speech perception and acoustic phonetics,* Ed. Cambridge University Press. 1988

[5] Y. Meyer, *Wavelets. Algorithms & Applications*, Ed. SIAM 1993.

[6] I. Chávez, "Sistema Automático de Reconocimiento de Voz" Tesis de Maestría en Ingeniería Eléctrica, División de Estudios de Postgrado, Univ. Michoacana de San Nicolás de Hidalgo, 2009.

[7] I. Daubechies, *Ten lectures on wavelets*. Ed. Philadelphia, SIAM. 1992.

[8] M. Misiti, Y. Misiti,, G. Oppenheim, and J. Poggi, *Wavelet Toolbox For Use With Matlab*, The Math Works, Inc., 1996.

[9] Oracle TechRep 2010, [en línea]. Disponible en http://www.oracle.com/technetwork/java/index-139508.html

[10] J. Mariani, "Recent advances in speech processing," *in Proc 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing.*

[11] L. Rabiner, A Rosenberg. "Considerations in Dynamic Time Warping algorithms for discrete word recognition". *IEEE Trans on Acoustics, Speech and Signal Processing*, Vol. 26. pp 575-582, Dec. 1978.

[12] R. Duda, P. Hart and D. Stork, *Pattern Classification 2nd Edition*. Ed John Wiley and Sons , 2001

[13] T. Fawcett. "Roc graphs: Notes and practical considerations for researchers". Kluwer Academic Publishers. Netherlands. HP Laboratories. 2004, [en línea]. Disponible en http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf

VI. Biografía

Ismael Chávez Nació en la ciudad de Morelia, Michoacán México el 31 de julio de 1967. Obtuvo el grado de Licenciatura en Ingeniería Eléctrica por la Escuela de Ingeniería Eléctrica de la Universidad Michoacana de San Nicolás de Hidalgo E.I.E-U.M.S.N.H. (1990). Posee el grado de Maestro en Ingeniería en la opción de Sistemas Computacionales por la División de Estudios de posgrado de la Facultad de Ingeniería Eléctrica F.I.E.) de la U.M.S.N.H. (2009). Actualmente se desempeña como Profesor e investigador de tiempo completo en la F.I.E.-U.M.S.N.H., cuerpo académico al que ingresó en 1990. Sus áreas de interés principales son entre otras: el reconocimiento automático de formas/patrones, y la instrucción/entrenamiento asistidos por computadora.

Antonio Camarena Nació en la ciudad de Morelia el 11 de Julio de 1964. Se graduó como Ingeniero Electricista en 1987 en la Universidad Michoacana. Obtuvo el grado de Maestro en Ciencias Computacionales en el Instituto Tecnológico de Toluca en 1996. Recibió el grado de Doctor en Ciencias en Ingeniería Eléctrica, opción Sistemas Computacionales. Es profesor miembro de la División de Estudios de Postgrado de la Facultad de Ingeniería Eléctrica y Jefe del Laboratorio de Sistemas Computacionales en la misma División.

ISBN: 978-607-95476-3-9