ISSN: 0798-1015

REVISTA ESPACIOS

Vol. 41 (32) 2020 • Art. 16

Recibido/Received: 07/07/2020 • **Aprobado**/Approved: 16/08/2020 • **Publicado**/Published: 27/08/2020

Reconocimiento de voz para el control de una silla de ruedas, basado en MFCC-Entropía Difusa

Speech Recognition for wheelchair control based on MFCC-Fuzzy Entropy

MEDINA, Boris A.[1]
SIERRA, Javier E. [2]
LÓPEZ, José L.[3]

Resumen
Este artículo presenta un enfoque para el reconocimiento de voz basado en el uso de los coeficientes cepstrales de frecuencia de Mel-Entropía difusa (MFCC-FE), cuyo propósito consiste en extraer y seleccionar los coeficientes de mayor relevancia que permitan identificar el comando de voz específico para ordenar el movimiento y direccionamiento de una silla de ruedas. La técnica MFCC extrae las características relevantes de la señal de voz y el criterio de Entropía Difusa reduce su dimensionalidad. Un algoritmo de clasificación proporcionado por Maquina de Vector de Soporte (SVM) recibe el vector de coeficientes resultante para decidir entre cinco palabras (Avanzar, Atrás, Parar, Izquierda y Derecha), la acción que ejecutarán los motores de CC en cada una de las ruedas de la silla de ruedas. Los resultados de la implementación comparan las técnicas MFCC y MFCC-FE, y entregan una precisión de reconocimiento superior al 91%.
Palabras clave: MFCC, entropía difusa, SVM, reconocimiento de voz.

Abstract
This article presents an approach to speech recognition based on the use of the Mel Frequency Cepstral Coefficients – Fuzzy Entropy (MFCC-FE), whose purpose is to extract and select the most relevant coefficients that allow the specific Speech command to be identified to order the movement and direction of a wheelchair. MFCC technique extracts the relevant features of the speech signal and the Fuzzy Entropy criterion reduces its dimensionality. A classifier algorithm provided by Support Vector Machine (SVM) receives the resulting vector of coefficients to decide between five words (Go, Back, Stop, Left and Right), the action that the DC motors will execute in each of the wheelchair wheels. The implementation results compare the MFCC and MFCC-FE techniques, and deliver a recognition accuracy of over 91%.
key words: MFCC, fuzzy entropy, SVM, speech recognition.

[1] Ingeniero Electrónico. Magíster en Automatización y Control Industrial. Docente Investigador Universidad de Sucre, Facultad de Ingenieía. E.mail: boris.medina@unisucre.edu.co
[2] Ingeniero Electrónico; Magíster en Ingeniería. Doctor en Ingeniería. Docente Investigador Universidad de Sucre, Facultad de Ingenieía. E.mail: javier.sierra@unisucre.edu.co
[3] Ingeniero Electrónico; Magíster en Ingeniería. Docente Investigador Universidad de Sucre, Facultad de Ingenieía. E.mail: jose.lopez@unisucre.edu.co

1. Introducción

El reconocimiento de voz es la capacidad que posee una máquina para identificar palabras y frases del lenguaje hablado y convertirlas a un formato legible por el autómata, para su posterior tratamiento. El reconocimiento de voz es un área de investigación incidente en el campo del procesamiento del lenguaje natural, procesamiento de señales, aprendizaje automático, estadísticas, etc. [Gold & Morgan, 2002], con aplicaciones prácticas en traducción automática de voz a otros idiomas, subtitulación de contenidos, herramientas de conversión de voz a texto o sistemas de toma de notas, asistencia de directorio telefónico, consulta de bases de datos habladas para usuarios, robótica, entre otras [Karpagavalli &Chandra, 2016].

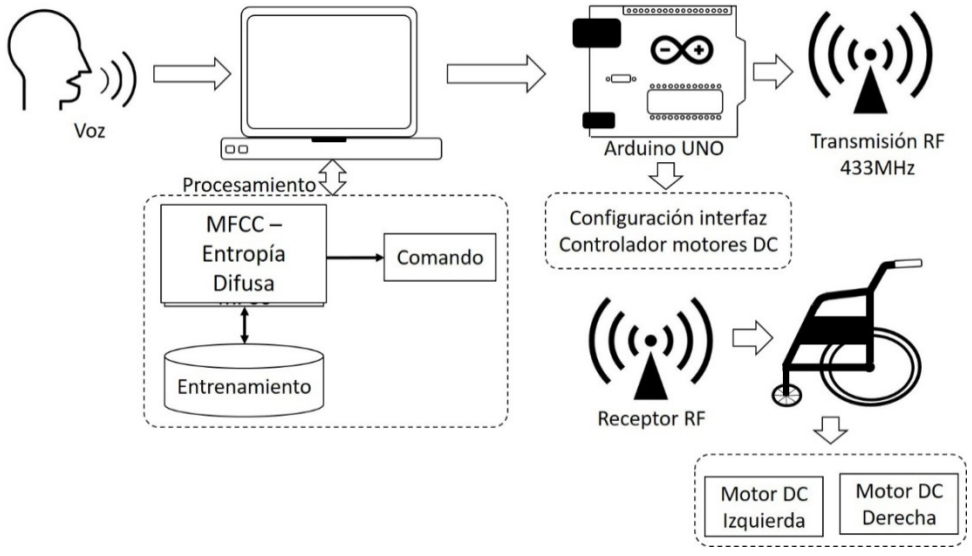
El algoritmo de reconocimiento de voz consta de varias etapas en las que la extracción y clasificación de características son operaciones fundamentales en la efectividad del sistema [Sen et al., 2019]. Usualmente se emplean técnicas de caracterización de las señales de voz que permiten extraer la información más relevante de la señal, para poder distinguir entre una señal de voz y otra, tales, como Coeficientes Predictivos Lineales (LPC) [Gupta, 2016], Coeficientes Cepstrales (MFCC) [Kumar et al., 2017], Coeficientes Cepstrales Predictivos Lineales (LPPC) [Gupta, 2016], Wavelets [Iswarya & Radha, 2017], Predicción lineal Perceptual (PLP) [Prithvi & Kumar, 2016], entre otras.

En nuestro trabajo, usaremos MFCC como método de extracción, y Entropía Difusa como criterio para reducir la dimensionalidad del conjunto de datos y distinguir las características más representativas que aportan alto contenido de información y suponen una reducción del costo computacional [Medina & Álvarez, 2017]. Nuestro enfoque nos permitirá identificar los distintos comandos de voz utilizados para controlar el movimiento de una silla de ruedas de manera remota, orientado a mejorar la calidad de vida en personas que sufren algún tipo de discapacidad funcional en sus extremidades.

2. Metodología

En este trabajo se propone el uso de la técnica reconocimiento de voz basada en la extracción de coeficientes cepstrales y reducción de la dimensionalidad a través del criterio de Entropía Difusa, para controlar el movimiento de una silla de ruedas a través de comandos de voz, como se detalla en la Figura 1.

Figura 1
Esquema funcional del sistema propuesto



En primera instancia se implementó un algoritmo para detectar, procesar y reconocer los comandos de voz. Las características más relevantes producto de la extracción a través de MFCC y reducción de la dimensionalidad a través de Entropía Difusa, se usan como vector característico para reconocer la clase a la que pertenece. Esta acción perrmitirá la interacción remota con la tarjeta Arduino para controlar inalámbricamente el movimiento de los motores que actuarán sobre la silla de ruedas que a su vez, ejecutarán la respectiva acción.

2.1. Conjunto de datos

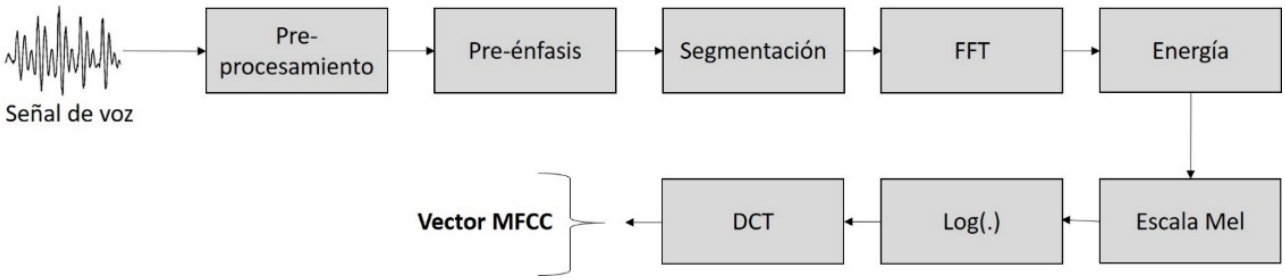
El conjunto de datos de entrenamiento se registró a partir de un sujeto normal durante una sesión sin retroalimentación. El sujeto se sentó en una silla normal, los brazos relajados descansando sobre la mesa. La tarea consistía en pronunciar una palabra de acuerdo a las indicaciones mostradas en una pantalla de computador; por ejemplo, mediante el texto: “Pronuncie la palabra IZQUIERDA una vez presione la tecla Enter” se pronunciaba la palabra “izquierda”. El experimento consistió en 1 sesión, llevada a cabo el mismo día.

Se obtuvieron 50 iteraciones de 2000 ms de longitud cada una, 10 por cada una de las cinco palabras empleadas en nuestro trabajo (clase 1: “Adelante”, clase 2: “Atrás”, clase 3: “Stop”, clase 4: “Izquierda”, clase 5: “Derecha”). Los datos se proporcionan a una frecuencia de muestreo de 44100 Hz.

2.2. Extracción de características

MFCC es una de las técnicas más populares para el reconocimiento de voz; utiliza la base espectral como parámetros para el reconocimiento de la señal de voz; MFCC está basado en la percepción de los sistemas auditivos humanos, debido al uso del posicionamiento de sus bandas de frecuencia en escalas logarítmicas (escala de Mel). La Figura 2 muestra el diagrama de bloques del proceso de extracción de características de MFCC.

Figura 2
Proceso de extracción de características MFCC



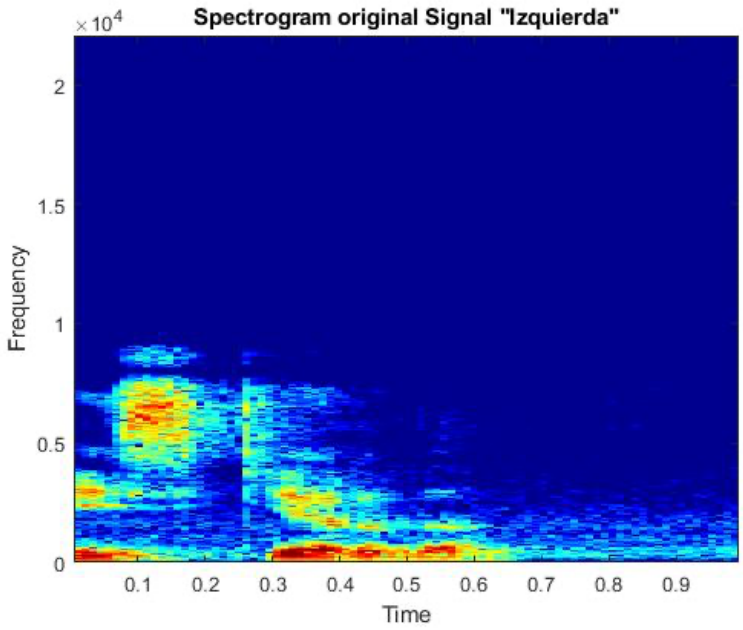
Pre-procesamiento: Inicialmente, se realiza un alistamiento a la señal de voz, eliminando los componentes de la señal donde aún no hay componentes de voz, y a partir del momento en que la señal detecta componentes de voz, se selecciona el primer segundo de duración, tiempo suficiente para recoger toda la información de la palabra pronunciada. La Figura 3 muestra un espectrograma de una señal preprocesada.

Pre-énfasis: Para compensar la atenuación de alta frecuencia causada por el proceso fisiológico del mecanismo de producción del habla, la voz de entrada cuantificada se enfatiza previamente mediante un filtro pasa alto de la forma $y[n] = x[n] – a.\ x[n-1]$ para equilibrar las amplitudes de baja frecuencia y alta frecuencia, donde "a" está dado por un valor 0.9 y 1; el valor seleccionado de “a” para nuestro experimento fue de 0.97 de acuerdo a [Young, 1997].

Segmentación: La señal de voz es un proceso aleatorio y no estacionario. Esto supone un inconveniente a la hora de analizarla. No obstante, es posible salvar este problema si se tiene en cuenta que a corto plazo de tiempo (del orden de milisegundos) la señal es casi-estacionaria. Para su análisis, procedimos a “*enventanar*” la señal,

segmentándola en tramas consecutivas de 20ms de duración, con un solapamiento de 10 ms por cada segmento, para asegurar la continuidad de la información de la señal, de tal manera de obtener información distintiva entre los diferentes tipos de sonidos producidos por la voz, mediante el posterior análisis espectral de la señal.

Figura 3
Espectograma de una señal de voz



FFT: A partir de aquí, se lleva a cabo la Transformada rápida de Fourier (FFT) en cada segmento de señal del dominio del tiempo al dominio de la frecuencia, para obtener un vector característico *x[k].*

Escala Mel: Teniendo en cuenta que la señal de voz es un proceso aleatorio y no estacionario, y no siguen una escala lineal, en MFCC, para cada tono con una frecuencia real *f*, se encuentra un tono subjetivo en una escala llamada escala "Mel". La escala de frecuencia de *Mel* realiza un espaciado de frecuencia lineal por debajo de 1000 Hz y un espaciado logarítmico por encima de 1000Hz. Basados en esta escala, se construye un banco 20 filtros triangulares, espaciados de acuerdo a la escala de Mel [Bala, 2010] mostrada en la ecuación (1) para una frecuencia dada "*f*" en Hz.

DCT: Finalmente, los coeficientes del espectro *Mel* son convertidos al dominio del tiempo utilizando la transformada discreta del coseno (DCT) a través de la ecuación (2). El resultado de este proceso es llamado MFFC. El conjunto de coeficientes obtenidos es nuestro vector acústico que servirá de entrada en el algoritmo de clasificación.

$$Mel(f) = 2595*log10(1 + f/700) \quad (1)$$

$$C_n = \sum_{k=1}^{k} (log\, S_k)\, cos\, \left\{ n\left(k - \frac{1}{2}\right)\frac{\pi}{k} \right\} \quad (2)$$

2.3. Selección de características mediante Entropía Difusa (FE)

En el enfoque probabilístico, la entropía de Shannon es una medida bien conocida de la incertidumbre y se cubre ampliamente en la literatura [Khushaba, 2007]. Una extensión de la entropía de Shannon es el concepto de

entropía difusa, en la que los conjuntos difusos se utilizan para ayudar a la estimación de la entropía. La entropía difusa se diferencia de la entropía de Shannon clásica desde el punto en que la entropía difusa contiene incertidumbres difusas (posibilista), mientras que la entropía de Shannon contiene incertidumbres con aleatoriedad (probabilística).

La entropía difusa, al igual que la entropía de Shannon satisface los cuatro axiomas de De Luca-Termini. La entropía de Shannon se define a partir de una variable aleatoria discreta (x) con función de probabilidad $p(x_i)$, dada por (3).

$$H(X) = -\sum_i p(xi) log_2 p(xi) \quad (3)$$

A partir de la entropía de Shannon, Khushaba et al., (2007), define la entropía difusa conjunta de los elementos de la clase *i*, denotada como $H(f, c_i)$, dada por (4), donde $P(f,c_i)$ puede ser interpretada como el grado en que la muestra predefinida para pertenecer a la clase i, realmente contribuye a esa clase específica.

$$H(f, c_i) = -P(f, c_i) log P(f, c_i) \quad (4)$$

La entropía difusa completa a lo largo de las c-clases está dada por (5):

$$H(f, C) = \sum_i^c H(f, c_i) \quad (5)$$

El equivalente difuso para la probabilidad conjunta de los patrones de entrenamiento que pertenecen a la clase *i,* está dada por (6).

$$P(f, c_i) = \frac{\sum_{k \in A_i} \mu_{ik}}{NP} \quad (6)$$

A_i es el conjunto de índices de los patrones de entrenamiento que pertenecen a la clase *i, NP* es el número total de patrones, y μ_{ik} es el *k*-ésimo valor de membresía difusa perteneciente a la clase *i*.

2.4. Clasificación

Las máquinas de vector de soporte (SVM) son esencialmente un clasificador binario no lineal, capaz de determinar si un vector de entrada "*x*" pertenece a una clase 1, donde la salida deseada sería $y = 1$, o a una clase 2 donde $y = -1$. El clasificador SVM se asigna a un nuevo espacio de mayor dimensionalidad que depende de una función no lineal y busca un hiperplano en ese nuevo espacio. El hiperplano separador es optimizado por la maximización del margen.

La clasificación consta de dos pasos: entrenamiento y evaluación. En la fase de entrenamiento, SVM recibe el vector característico MFCC como entrada. Este vector característico de voz representado por N parámetros de características pueden verse como puntos en el espacio N-dimensional. En este estudio, se forman 12 coeficientes MFCC para la matriz de características de entrada. Luego, la máquina de clasificación puede encontrar las etiquetas de los nuevos vectores comparándolos con los utilizados en la fase de entrenamiento.

2.5. Interfaz de comunicación

Como interfaz de comunicación para establecer la conexión entre el computador y la silla de ruedas, se empleó una tarjeta Arduino UNO. Un algoritmo dentro de la tarjeta Arduino recibe la señal de la clase identificada, procesa la información y envía una señal a dos motores de corriente directa ubicados en cada una de las ruedas

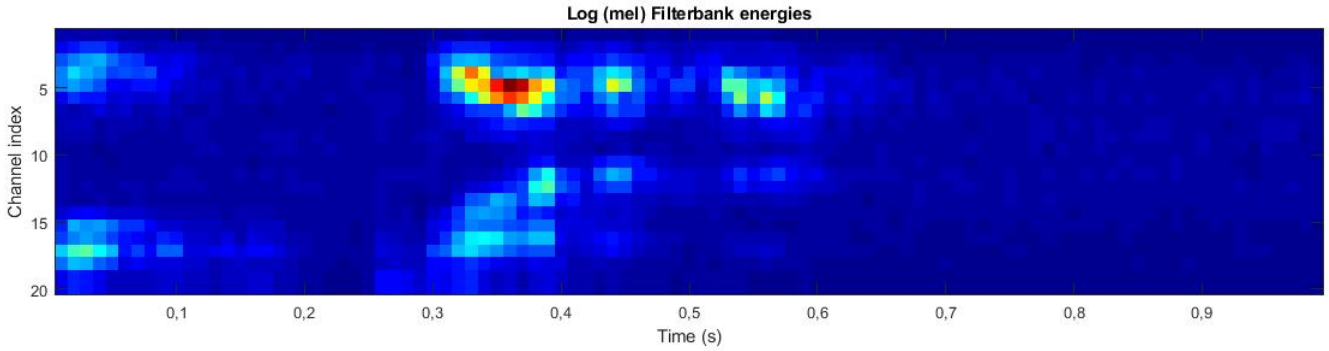
traseras de una silla de ruedas, para su correspondiente activación a través de los módulos inalámbricos de transmisión FS1000A y recepciones XY-MK-5V compatibles con Arduino.

3. Resultados

Texto La digitalización de la señal de audio se muestrea a una frecuencia de 44100 Hz con una duración de 1 segundo en cada grabación, con lo que se produce un vector de 44100 muestras para cada señal. Luego de aplicar el filtro pasa alta como preénfasis para cada trama, usamos la Transformada rápida de Fourier en cada una de las tramas segmentadas de la señal, para obtener un vector característico *x[k]*.

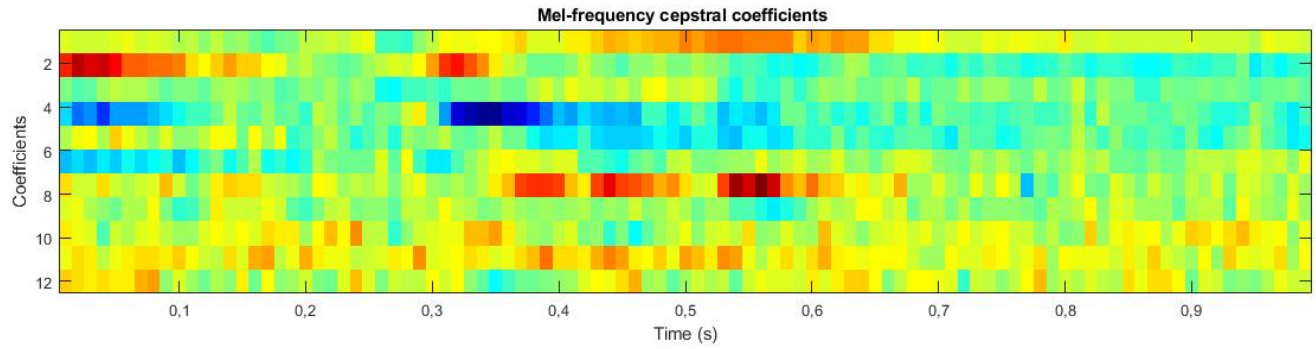
El banco de filtros *Mel* es multiplicado por el espectro de energía del vector característico *x[k]*; Un total de 20 coeficientes nos indican la energía en cada banco de filtros. La Figura 4 muestra la energía del banco de filtros aplicado a la señal de voz de referencia con la palabra "*Izquierda*".

Figura 4
Energía del Banco de filtros,
palabra "Izquierda"



Posteriormente usamos la Transformada Discreta del Coseno (DCT) para eliminar la dependencia y correlación estadística en las bandas adyacentes de los filtros. La DCT lleva los coeficientes espectrales resultantes al dominio de la *quefrencia*, convirtiéndolos en coeficientes cepstrales (MFCC). DCT es calculada usando la ecuación mostrada en (2) [Muda et al., 2010], donde *n = 1, 2, … k,* y *Sk* son los coeficientes FFT. Los resultados de este procedimiento se muestran en la Figura 5 para la señal de referencia. Este procedimiento es usado para extraer las características de la señal, tanto para el entrenamiento como para su validación.

Figura 5
Un vector característico de coeficientes
MFCC, de la palabra "Izquierda"

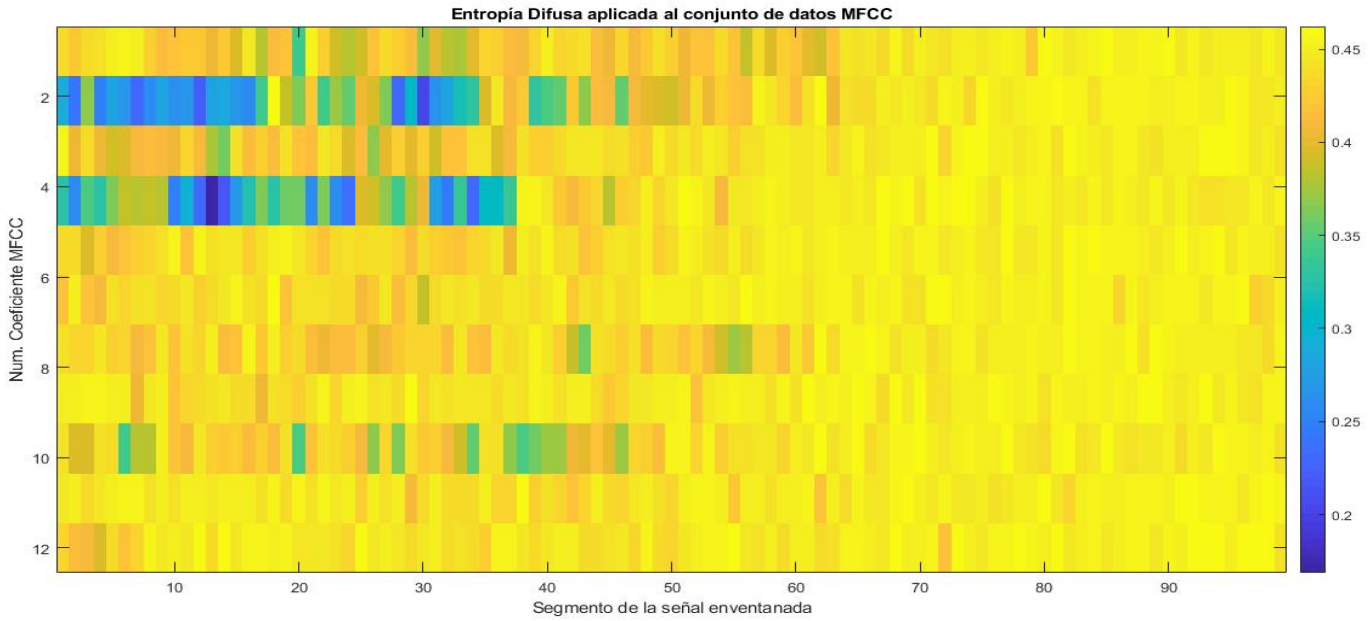


Los 12 coeficientes MFCC en cada una de las 99 segmentaciones de 20ms con solapamiento de 10ms resultantes del enventanado de la señal, producen un vector característico con un total de 1188 coeficientes.

Usando la función *fuzzy c-means* de Matlab, se construyen los valores de membresía difusos μ_{ik} para cada una de las características obtenidas en el conjunto de datos; cada una de estas refleja el grado de pertenencia de la muestra para cada una de las cinco clases.

Combinando los valores de membresía μ_{ik} obtenidos y las ecuaciones (4), (5) y (6), se calculó la entropía difusa sobre el espacio de características específicas. Un valor alto de entropía contribuye poco a la desviación entre las clases y los valores de entropía bajos presentan características más informativas. De esta manera logramos discriminar los coeficientes característicos más relevantes para reducir la dimensionalidad del vector. La Figura 6 detalla en escala de colores, los valores de Entropía Difusa obtenidos del conjunto de datos MFCC. Podemos observar que los cuadros en color azul representan los coeficientes que contribuyen con mayor información dentro del conjunto de datos, donde se destacan los coeficientes cepstrales 2 y 4 para los segmentos de la señal menores a 40.

Figura 6
Entropía Difusa aplicada
al conjunto de datos MFCC



Para evaluar el rendimiento del algoritmo propuesto, se aplicó el criterio de validación cruzada de diez particiones; para cada clasificador de entrenamiento, las validaciones cruzadas se han ejecutado 100 veces.

Empleando el algoritmo de clasificación SVM, como herramienta para clasificar la matriz generada del entrenamiento de nuestro algoritmo, y distinguir entre cada una de las clases, obtuvimos una precisión por encima del 91% para cada una de las clases. La Tabla 1 muestra los porcentajes de acierto al ejecutar el entrenamiento y validación del conjunto de datos, para diferente número de características, Los resultados de la primera columna se obtuvieron del conjunto de datos producto de la extracción MFCC; Los resultados de la segunda columna se obtuvieron del conjunto de datos producto de la extracción MFCC y la selección de los coeficientes más informativos resultantes de la aplicación del algoritmo de Entropía Difusa cuyos valores fueron menores a 0.25 (39 coeficientes).

Tabla 1
Porcentajes de acierto en la distinción de las clases aplicando MFCC/EF

	MFCC (1188 Coef)	MFCC/FE (39 Coef)
Clase 1	90%	90%
Clase 2	100%	100%
Clase 3	100%	100%
Clase 4	100%	100%
Clase 5	100%	100%
Promedio	98%	98%

Es de destacar que la reducción de la dimensionalidad del vector característico en MFCC-FE permitió reducir el costo computacional en la ejecución del algoritmo en aproximadamente 55.5% respecto al algoritmo MFCC, en la etapa de clasificación SVM. La ejecución de los algoritmos se realizó sobre un PC de 64 bits con procesador Intel core i5 de 2.3 GHz y 6 GB de RAM.capítulo 3 Estilo Normal. Calibri 11. Espaciado 1,08

4. Conclusiones

En este estudio representamos un enfoque para el reconocimiento de comandos de voz basado en MFCC-FE. El uso de MFCC permitió extraer los coeficientes cepstrales representantivos de la señal, y la Entropía Difusa ayudó a distinguir las características de relevancia y reducir el espacio de dimensionalidad, sin perder la precisión de clasificación, respecto al número inicial de características y lograr una reducción del costo computacional. Los coeficientes resultantes de MFCC-FE alimentaron como vector característico, el clasificador SVM para lograr la identificación de la palabra o comando de voz pronunciado.

El algoritmo es entrenado para identificar y reconocer entre cinco palabras. Los resultados mostraron que guardando 10 plantillas por cada palabra, se obtienen porcentajes de precisión superior al 91%, tanto en MFCC como en MFCC-FE.

Como trabajo futuro, podemos incluir el uso de otras técnicas no lineales de extracción y discriminación de características como métodos de inteligencia artificial para mejorar la precisión del reconocimiento y velocidad de procesamiento. Así mismo implementaremos otras interfaces de comunicación para evaluar la latencia de las unidades de hardware.

Referencias bibliográficas

Bala A., Kumar A., Birla N. (2010). “Voice Command Recognition System Based On MFCC And DTW”, International Journal of Engineering Science and Technology, Vol. 2, No. 12, pp.7335- 7342.

De Luca, A. and Termini S. (1972). “A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory”. Inf. Control, vol. 20, núm. 4, 301–312.

Gold B., Morgan N. (2002). “Speech and Audio Signal Processing”, New York, John Wiley and Sons.

Gupta H., Gupta D. (2016). "LPC and LPCC method of feature extraction in Speech Recognition System", 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), pp. 498-502, Noida.

Iswarya P., Radha V. (2017). “Speech Query Recognition for Tamil Language Using Wavelet and Wavelet Packets”, Journal of Information Processing System, Vol. 13, No. 5, pp.1135-1148.

Karpagavalli S., Chandra E. (2016). “Review on Automatic Speech Recognition Architecture and Approaches”, International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 9, No. 4, pp.393-404.

Khushaba RN., Al-Jumaily, A., Al-Ani A. (2007). “Novel feature extraction method based on fuzzy entropy and

Wavelet Packet transform for myoelectric Control”. 2007 Int Symp Commun Inf Technol, 352–357.

Kumar A., Rout S.S., Goel V. (2017). "Speech Mel Frequency Cepstral Coefficient feature classification using multi level support vector machine", 4th IEEE Uttar Pradesh Section International Conference on Electrical,

Computer and Electronics (UPCON), pp. 134-138.

Medina B., Alvarez R. (2017). “Characterization of EEG Signals Using Wavelet Packet and Fuzzy Entropy in Motor Imagination Tasks” Ingeniería, vol. 22, no. 2, pp. 226-238.

Muda L., Begam M., Elamvazuthi I. (2010). “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”, Journal of Computing, Vol. 2, No. 3, pp. 138-143.

Prithvi P., Kumar TK. (2016). “Comparative Analysis of MFCC, LFCC, RASTA –PLP”. International Journal of Scientific Engineering and Research (IJSER). Vol. 4, No. 5, pp. 4-7.

Sen S., Dutta A., Dey N. (2019). “Feature Extraction. In: Audio Processing and Speech Recognition”, SpringerBriefs in Applied Sciences and Technology. Springer, Singapore.

Young S. et al. (1997). “The HTK book”. Vol. 2. Cambridge, Massachusetts: Cambridge University.