Visión Electrónica Vol. 10 No. 1 (2016) • Enero-Junio • p.p. 1- • ISSN 1909-9746 • ISSN-E 2248-4728 • Bogotá (Colombia)

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

Visión Electrónica

Más que un estado sólido

http: revistas.udistrital.edu.co/ojs/index.php/visele/index

VISIÓN ELECTRÓNICA

VISIÓN INVESTIGADORA

Reconocimiento de patrones de habla usando MFCC y RNA

Recognition of speech pattern using MFCC and ANN

***Olga L. Ramos*[1], *Diego A. Rojas*[2], *Leonardo A. Góngora*[3]**

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Enviado: 14/03/2016
Recibido: 24/03/2016
Aceptado: 26/04/2016

Palabras clave:
Habla
Inteligencia artificial
MFCC
Reconocimiento de patrones
RNA

RESUMEN

Este estudio presenta los resultados del diseño y desarrollo de un algoritmo basado en inteligencia artificial para el reconocimiento de patrones de vocablos del idioma español utlizando Coeficientes Cepstrales en las Frecuencuas de Mel (MFCC). El uso de MFCC ha permitido caracterizar las señales de voz teniendo en cuenta el ruido presente en el ambiente de grabación, lo que ayuda a la obtención de patrones comunes entre estas señales cuando presentan alteraciones a través de un clasificador basado en Redes Neuronales Artificiales (RNA). Como resultado se obtuvo un reconocimiento superior al 95 % de las tres vocales escogidas: /a/, /e/, /o/; entre un grupo de 22 muestras por vocal de entrenamiento y 11 muestras para la validación. Las muestras se tomaron de 11 personas de género masculino

Keywords:
Speech
Artificial intelligence
MFCC
Pattern recognition
ANN

ABSTRACT

This study shows the results of the designing and development of an algorithm based on artificial intelligence and Mel Frequency Cepstral Coefficients (MFCC) it recognizes Spanish words by using speech patterns. The using of MFCC has allowed to characterize voice signals, taking into account the noise in a record environment which helps with the estimation of common patterns among these signals when presents disturbances through a classifier based in Artificial Neural Networks (ANN). As main result of this study, a recognizing rate between 93 % and 96 % of the selected vowels (/a/, /e/, /o/) was achieved. For the training a number of 22 samples were used and other 11 for the validation process. The samples were obtained from 11 test subjects, all of them were male genre.

[1]Ingeniera electrónica, Universidad Antonio Nariño; M.Sc Teleinformática; PhD Ingeniería, Universidad Distrital Francisco José de Caldas; docente, Universidad Militar Nueva Granada, Colombia. Correo electronico: olga.ramos@unimilitar.edu.co

[2]Ingeniero en Mecatrónica, Universidad Militar Nueva Granada, Colombia.; coinvestigador, Universidad Militar Nueva Granada, Colombia. Correo electronico: u1801620@unimilitar.edu.co

[3]Ingeniero Mecatrónico, Universidad Piloto de Colombia; coinvestigador, Universidad Militar Nueva Granada, Colombia. Correo electronico: tmp.leonardo.gongora@unimilitar.edu.co

Citar este artículo como: O. Ramos, D. Rojas, L. Góngora. "Reconocimiento de patrones de habla usando MFCC y RNA." Visión Electrónica, algo más que un estado sólido, Vol. 10, No. 1, 1-, enero-junio 2016.

1. Introducción

El habla es la forma verbal de la comunicación humana, está basada en la combinación sintáctica, entre léxico y nombres, que es extraída de un compendio de palabras llamado vocabulario. Cada palabra pronunciada es creada gracias a la combinación fonética de unidades sonoras conocidas como vocales y consonantes; el habla, es un proceso formado por cuatro eventos: producción, percepción, repetición y error. Lingüísticamente, la producción del habla describe cómo se posiciona la lengua, los labios, la mandíbula y otros órganos involucrados en este proceso para lograr la producción de un sonido [1]; la percepción [2] hace referencia al proceso a través del cual el ser humano es capaz de interpretar y entender los sonidos propios de un lenguaje; la repetición [3] es cuando el sonido relacionado a una palabra se trasforma en un movimiento involuntario de los órganos relacionados al habla, de esta manera el habla se convierte en un proceso mecánico.

Estos procesos involucrados en la producción de habla han sido de gran interés por parte de investigadores del área de procesamiento de señales, que, a partir de los sistemas computacionales modernos, han llegado a construir sistemas complejos de reconocimiento automático de voz (ASR), síntesis de texto a voz, identificación del hablante, compresión de datos de habla, detección de emociones, entre otros desarrollos [4]. Los sistemas ASR, han sido de particular interés en el medio académico, formando parte del principal foco de trabajo en el tratamiento de señales de voz.

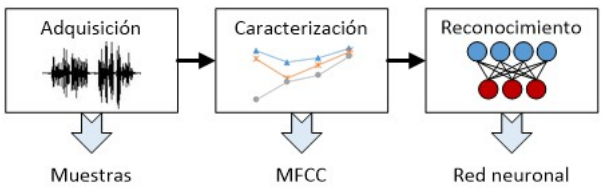
En los sistemas de reconocimiento automático de voz, se identifican tres fases importantes para el procesamiento de habla y tienen que ver con el análisis de las características de las señales de habla, la clasificación y reconocimiento de patrones y la verificación de pronunciación de las palabras reconocidas por el sistema [5], [6]. Como principal método de extracción de características de señales de voz, se tienen los coeficientes cepstrales de las frecuencias de Mel [7], este tipo de técnica ha sido usado en numerosos intentos para realizar la identificación de voz y habla, como el presentado en [8], donde combinan MFCC con una técnica que considera el corrimiento temporal conocida como DTW para realizar la comparaciones de patrones de voz. Algunas técnicas que complementan de manera exitosa los métodos de extracción de características como el MFCC, con los algoritmos de reconocimiento de patrones que utilizan redes neuronales como se ilustra en [9], donde detectan trastornos de la voz a través de los coeficientes cepstrales de la misma y un perceptrón multicapa.

Basado en lo anterior, en este trabajo se presentan los resultados de la extracción de características a partir de señales del habla, utilizando la técnica de MFCC; de igual modo se realiza la clasificación e identificación de los patrones obtenidos de estas características, utilizando un algoritmo de inteligencia artificial, basado en redes neuronales.

2. Formulación del problema

El desarrollo de un sistema para la identificación de unidades fonéticas definidas, como palabras, morfemas o fonemas, es una temática ampliamente trabajada desde el desarrollo de la teoría de señales; este trabajo, evidencia el desarrollo de un sistema para el reconocimiento de los fonemas vocálicos abiertos del español, que corresponden a: /a/, /e/ y /o/. Como metodología generalizada de trabajo, se plantea el desarrollo de un sistema por etapas, como el que se muestra en la Figura 1.

Figura 1: Diagrama de bloques del proceso planteado



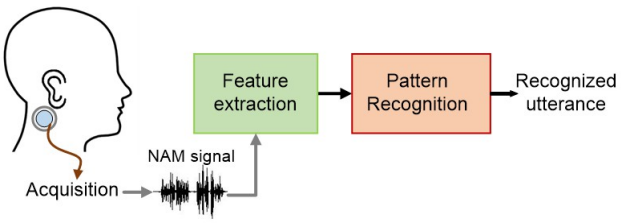
Fuente: elaboración propia.

3. Metodología

Las aplicaciones de la teoría de señales en el procesamiento de habla, han sido ampliamente trabajadas desde hace más de cincuenta años. Sistemas para el reconocimiento automático de voz (ASR), la compresión de datos de habla *(speech coding)*, síntesis de texto a voz o la identificación/verificación del hablante, son algunos de los ejemplos en donde la teoría de procesamiento de señales es aplicada al tratamiento de voz [4].

Dentro de estas aplicaciones, los sistemas ASR quizá han sido los más investigados debido a lo que su desarrollo representa; la construcción de un sistema que sea capaz de interactuar con el hombre, al reconocer y entender con fluidez la voz. Un sistema ASR está constituido básicamente por dos elementos, el primero de ellos corresponde a la extracción de características de la señal de voz que ingresa al sistema; el segundo hace referencia al reconocimiento de patrones para la comparación e identificación de las palabras pronunciadas. La Figura 2 muestra en detalle los elementos constituyentes de un sistema ASR.

Figura 2: Diagrama de bloques generalizado de un sistema ASR



Fuente: elaboración propia.

En la Figura 2, el parámetro $s[n]$, corresponde a la señal de voz digitalizada a través de la tarjeta de audio de un pc portátil a una frecuencia de , 8KHz y $\bar{w}$, es la unidad fonológica reconocida.
Para el caso de estudio que se presenta en este documento, se desarrolló un sistema para el reconocimiento de fonemas vocálicos de tipo abierto (/a/, /e/ y /o/); para esto, se contó con un grupo de trabajo de once personas de quienes fueron grabadas tres muestras por vocal. El conjunto de grabaciones resultante fue ingresado como parámetro de procesamiento para la etapa de extracción de características, las cuales fueron representadas a través de los coeficientes cepstrales de las frecuencias de Mel; finalmente, estos MFCC fueron ingresados a una red neuronal para el entrenamiento del sistema y posterior validación del proceso de identificación de los fonemas vocálicos trabajados.

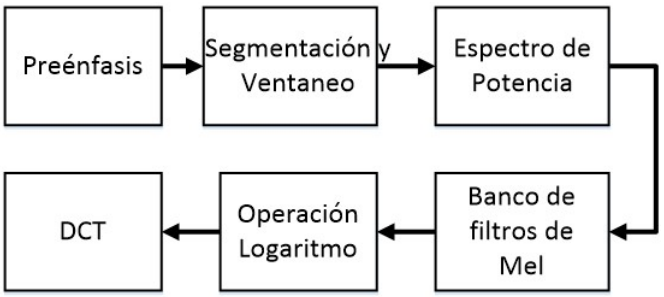
3.1. Metodología para la extracción de los MFCC

Son numerosos los métodos de estimación de características aplicados al procesamiento de voz como el análisis cepstral [10], [11], el Linear Predictive Coding (LPC, por sus siglas en inglés) [12], [13], el MFCC [14], además de aproximaciones clásicas basadas en el análisis temporal de las señales y las representaciones en frecuencia de las mismas.

A pesar del gran número de metodologías disponibles, los coeficientes cepstrales en las frecuencias de Mel, representan la herramienta más utilizada para la extracción de características, por su robustez y calidad en la información representativa que entrega [15], [16]. La Figura 3, muestra los procesos involucrados para la extracción de los MFCC de una señal de voz.

Como primera etapa en la estimación de los coeficientes MFCC, se tiene el preénfasis de la señal de audio digitalizada, esto es logrado a partir de la aplicación de un filtro pasa altos de primer orden, definido como se muestra en (1).

Figura 3: Procesos involucrados para el cálculo de los MFCC



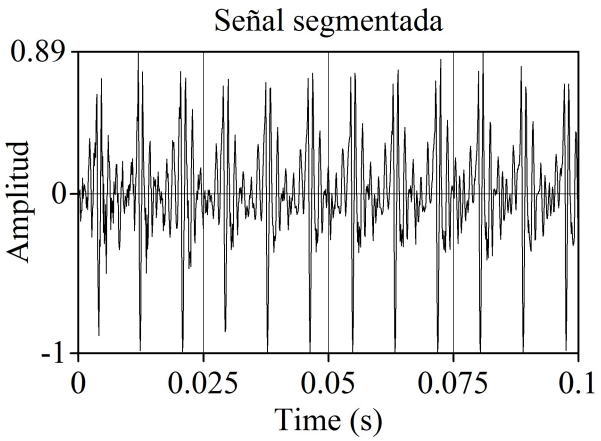
Fuente: elaboración propia.

$$p(z) = 1 - 0{,}97z^{-1} \tag{1}$$

El preénfasis es aplicado a las señales de voz, para incrementar la magnitud de la cantidad de energía presente en las altas frecuencias de la señal y hacer de este parámetro información detectable en posteriores fases de procesamiento [17].

El ventaneo de la señal, se realiza al dividir la señal de entrada ($s[n]$) en tramas cortas ($x_i[n]$). Se escogió como ancho de ventana un tiempo de 25 ms , y un tiempo de desplazamiento no superior a 10 ms [18]. La Figura 4 representa una de las muestras utilizadas para el desarrollo del sistema ASR.

Figura 4: Señal representativa de la vocal /a/ junto con los segmentos de ventana escogidos



Fuente: elaboración propia.

El siguiente paso para la obtención de los coeficientes de Mel, corresponde al cálculo del periodograma de la señal, con esto se busca encontrar la cantidad de energía

presente en cada una de las bandas de frecuencia en las que la señal se encuentra ubicada; se utiliza entonces (2), para realizar este proceso.

$$P_i[k] = \frac{1}{N} \mid S_i[k] \mid^2 \tag{2}$$

Donde es el periodograma resultante, es el número de muestras presentes en cada una de las ventanas de la señal y representa la transformada de Fourier en tiempo discreto (3).

$$S_i[k] = \sum_{n=1}^{N} x_i[n]h[n]e^{-j2\pi kn/N} \tag{3}$$

La ecuación (3), representa la FFT del segmento $x_i[n]$, el cual es multiplicado escalarmente con un vector que corresponde a la definición matricial de una ventana de tipo Hamming.
Después de hallar la densidad de potencia presente en las bandas de frecuencia de la señal, se filtra el vector de datos resultante, utilizando un banco de filtros de Mel, que son definidos a partir de la relación representada en (4).

$$mel(f) = 2595Ln\left(1 + \frac{f}{700}\right) \tag{4}$$

Por último, se aplica la operación logarítmica al resultado del proceso de filtrado, para extraer los coeficientes de Mel utilizando la transformada discreta de coseno (DCT), cuyo objetivo es de correlacionar las cantidades estimadas a partir de la aplicación de los filtros de Mel [14].
El resultado final después de aplicar la DCT, condensa 13 valores de coeficientes por ventana, de acuerdo al número de filtros definidos para la extracción. Estos 13 datos, corresponden a los coeficientes estáticos de las frecuencias de Mel, y fueron utilizados como parámetros de entrenamiento de la red neuronal para el reconocimiento de las vocales.

3.2. *Reconocimiento de patrones*

El reconocimiento de patrones es una derivación del aprendizaje de máquina que se centra en identificar los comportamientos regulares en series de datos. En esencia, estos sistemas describen y clasifican patrones recopilados a partir de objetos, eventos, personas, o señales, para este caso provenientes del habla, teniendo presente características únicas para cada patrón que se desea reconocer. El clasificador usado para el desarrollo de este trabajo está basado en redes neuronales, para lo cual se necesita un conjunto que contenga todos los patrones a reconocer debidamente etiquetados para realizar el sistema de reconocimiento de vocales se propuso el algoritmo de clasificación descrito a continuación.

3.2.1. *Arquitectura de la red neuronal*

Para la elaboración de este trabajo se utilizó una red neuronal de tipo feed forward, la cual consta de tres capas: una primera para el ingreso de los patrones a la red; la segunda es la capa oculta que realiza la mayoría del procesamiento de los datos y, por último, se encuentra la capa de salida que recibe los datos procesados de la capa oculta y los ajusta para arrojar los valores de salida de la red entera. La función de activación usada para la capa oculta fue la función sigmoidal representada por (5).

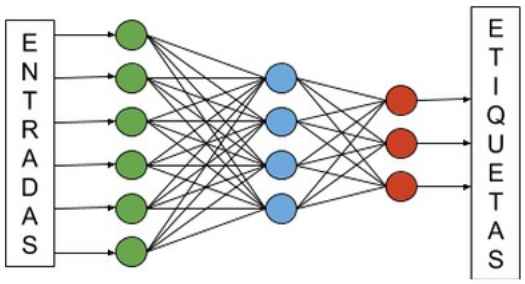
$$f(n) = \frac{1}{1 + e^{-n}} \tag{5}$$

Y la función de activación para la capa de salida fue la función softmax tal cual se ilustra en (6)

$$g(n) = \frac{e^n}{\sum_{i=1}^{k} e^n} \tag{6}$$

La topología de la red es como se ilustra en la Figura 5, además las funciones mencionadas anteriormente se seleccionaron debido a las ventajas que ofrecen en el entrenamiento y, por ende, para la tarea de clasificación, por ejemplo, a derivada de la función sigmoide puede expresarse en términos de la misma función.

Figura 5: Topología de la red Feed-Forward



Fuente: elaboración propia.

3.2.2. *Aprendizaje*

El entrenamiento de la red se llevó a cabo por medio del algoritmo de propagación hacia atrás o back propagation, con una diferencia en la manera de realizar el computo del error para iterar dicho algoritmo, ya que en una red neuronal usada para ajustar datos o realizar regresiones es común utilizar el valor del error cuadrático medio, pero en tareas de clasificación se recomienda utilizar el error calculado por entropía cruzada tal cual se muestra en (7).

$$E = -\frac{1}{N}\sum_{n=1}^{N}[y_n log(\widehat{y_n}) + (1 - y_n) log(1 - \widehat{y_n})] \quad (7)$$

Donde:
y_n Salida Deseada
$\widehat{y_n}$ Salida Obtenida
N Cantidad de muestras

En la Tabla 1 se encuentra un ejemplo que ilustra la ventaja de utilizar el error por entropía cruzada en lugar del error cuadrático medio, debido a que este primero representa de una mejor manera el error total de aprendizaje que va presentando el sistema de reconocimiento a medida que las épocas van incrementándose, y es de acuerdo a este valor que el algoritmo de entrenamiento irá actualizando y ajustando los valores de los pesos de la red hasta obtener las salidas deseadas que proporcionen un error despreciable o idealmente cero.

Tabla 1: Diferencia entre el error cuadrático medio (ECM) y el error por entropía cruzada (EEC) en tareas de clasificación

Salidas			Etiquetas			ECM	EEC
A	E	O	A	E	O		
0.7	0.45	0.2	1	0	0	0.333	0.357
0.15	0.5	0.2	0	1	0	0.313	0.693
0.1	0.05	0.6	0	0	1	0.173	0.511
Total						0.164	0.312

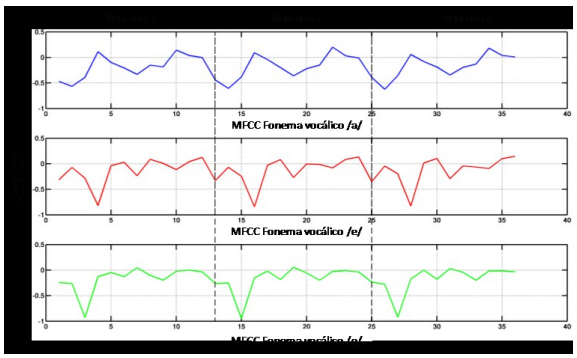
Fuente: elaboración propia.

4. Resultados

Para el caso de estudio que se presenta en este documento se desarrolló un sistema para el reconocimiento de fonemas vocálicos de tipo abierto (/a/, /e/ y /o/); para esto, se contó con un grupo de trabajo de once personas de quienes fueron grabadas tres muestras por vocal. El conjunto de grabaciones resultante, fue ingresado como parámetro de procesamiento para la etapa de extracción de características, las cuales fueron representadas a través de los coeficientes cepstrales de las frecuencias de Mel. Finalmente, estos MFCC fueron ingresados a una red neuronal para el entrenamiento del sistema y posterior validación del proceso de identificación de los fonemas vocálicos trabajados.

Como primer resultado en el desarrollo de este sistema de reconocimiento de los fonemas vocálicos abiertos del español se tienen de acuerdo a la metodología, los coeficientes extraídos a partir de la metodología de MFCC. La Figura 6 condensa gráficamente los valores de los coeficientes de las tres primeras ventanas para tres de las muestras grabadas.

Figura 6: Coeficientes de Mel para las tres primeras ventanas de los fonemas /a/, /e/ y /o/

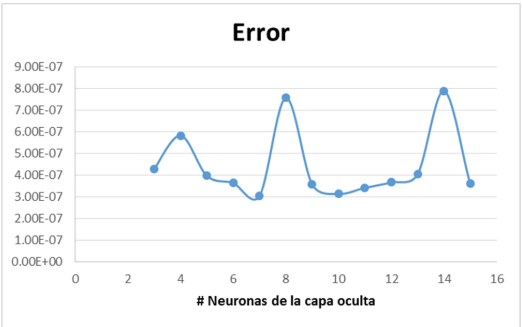


Fuente: elaboración propia.

Como se explicó anteriormente, el resultado final de la metodología MFCC arroja doce coeficientes denominados estáticos en relación a su representación característica de la señal. En la Figura 6 se puede notar a simple vista, que existe una relación entre los coeficientes de cada una de las ventanas para cada una de las muestras de los fonemas vocálicos. Esto se debe al método de extracción de los segmentos de audio a trabajar, en donde se adquirió un vector de datos de señal sin tratar del intervalo de tiempo neto en donde se produce la excitación de las cuerdas vocales para la pronunciación de las vocales.

Los patrones reflejados en los coeficientes, justifican la implementación de redes neuronales para la identificación aislada de las unidades fonéticas analizadas.

Figura 7: Error de entrenamiento para la red neuronal

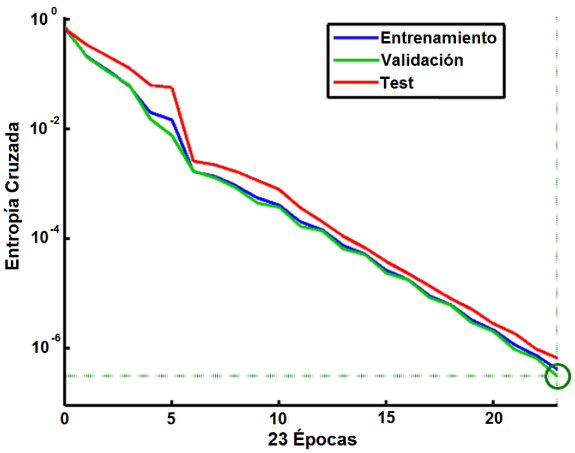


Fuente: elaboración propia.

En la Figura 7 se puede apreciar el comportamiento que presenta el error de la red neuronal programada en función de la cantidad de capas ocultas escogidas para el procesamiento de los patrones.

Teniendo en cuenta los datos mostrados en la Figura 5, la cantidad de neuronas que se escogió para implementar la red fue de siete, ya que para esta cantidad el error presentó su valor más bajo. El proceso de aprendizaje es como se muestra en la Figura 8, en esta se aprecia que fueron necesarias veintitrés épocas para alcanzar un valor de error adecuado para clasificar correctamente los patrones que le fueron enseñados a la red.

Figura 8: Error obtenido al finalizar el aprendizaje



Fuente: elaboración propia.

El proceso de entrenamiento mostrado en la Figura 8 se realizó con los datos de veintidós muestras para cada vocal, lo que hace un total de 66 muestras para las tres categorías. Para la etapa de validación de utilizaron once muestras no tenidas en cuenta en la etapa de aprendizaje, las cuales fueron tratadas con el mismo sistema de extracción de características que se usó para las otras 66; los datos arrojados por la red neuronal en la etapa de evaluación se encuentran consignados en la Tabla 2, es importante destacar que los datos utilizados para comprobar el correcto funcionamiento de la red fueron muestras que no se tuvieron en cuenta en la etapa de aprendizaje; esto permite verificar la capacidad de generalización adquirida por la red para clasificar e identificar patrones nuevos.

Los valores consignados en la Tabla 2 indican que para una cantidad de siete neuronas en la capa oculta la generalización de patrones que es capaz de realizar la red neuronal es buena, ya que el porcentaje promedio de identificación para las tres vocales oscila entre 93 y 96 %, obteniendo una mayor precisión para el caso de la vocal /a/ con un 96.1 %.

Tabla 2: Resultados de la red neuronal en la tarea de clasificación

Vocal	Numero de muestra											Promedio
	1	2	3	4	5	6	7	8	9	10	11	
A	1	0.98	0.95	0.92	0.96	0.97	0.92	0.96	0.94	0.97	1	961
E	0.92	0.97	0.94	0.97	0.96	0.91	0.98	0.9	0.93	0.89	0.92	935
O	0.93	0.9	0.95	0.97	0.95	0.92	0.95	1	0.9	1	0.93	945

Fuente: elaboración propia.

5. Conclusiones

Los coeficientes cepstrales de las frecuencias de Mel, demuestran ser los principales entes de información en el desarrollo de aplicaciones como la que se planteó en este documento; los datos representativos extraídos son relevantes y muestran ser importantes para condensar los patrones variables de la señal de voz, tal y como se demostró anteriormente. También, a pesar de que su uso fue limitado a la extracción de los doce primeros coeficientes, esto le fue suficiente a la red neuronal para lograr la identificación de las señales de voz de los fonemas vocálicos /a/, /e/ y /o/.

El buen desempeño de los algoritmos de clasificación basados en inteligencia artificial como las redes neuronales, depende en gran medida del método de extracción de características usado para encontrar los comportamientos comunes entre patrones, ya que si la extracción define lo suficientemente bien las clases a identificar el trabajo de la red neuronal será más eficaz a nivel de procesamiento y costo computacional. Otro de los factores o parámetros que pueden ser variados en función de mejorar el desempeño de la red, son las funciones de activación programadas en la capa oculta y de salida, ya que la selección de éstas depende del comportamiento que tengan los datos que representan los patrones a identificar. Como objeto de investigación en el futuro se propone algoritmos que complementen los valores arrojados del MFCC, como lo son los modelos ocultos de Markov y técnicas que consideren corrimientos temporales como Dynamic Time Warping.

6. Reconocimientos

A la vicerrectoría de investigaciones de la Universidad Militar Nueva Granada, por la financiación del proyecto ING/INV 1762, año 2015.

Referencias

[1] K. E. Watkins, A. P. Strafella, and T. Paus, "Seeing and hearing speech excites the motor system involved in speech production". *Neuropsychologia*, vol. 41, no. 8, pp. 989–994, Jan. 2003.

[2] N. Kazanina, C. Phillips, and W. Idsardi, "The influence of meaning on the perception of speech sounds". *Proc. Natl. Acad. Sci.* U. S. A., vol. 103, no. 30, pp. 11381–6, Jul. 2006.

[3] A. D. Friederici and S. M. E. Gierhan, "The language network". *Curr. Opin. Neurobiol.*, vol. 23, no. 2, pp. 250–4, Apr. 2013.

[4] L. R. Rabiner and R. W. Schafer, "Theory and Applications of Digital Speech Processing". 1st ed. Pearson, 2011.

[5] H. Veisi and H. Sameti, "Speech enhancement using hidden Markov models in Mel-frequency domain". *Speech Commun.*, vol. 55, no. 2, pp. 205–220, Feb. 2013.

[6] Q. Bao Nguyen, T. Thang Vu, and C. Mai Luong, "Improving acoustic model for English ASR System using deep neural network". in The 2015 *IEEE RIVF International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pp. 25–29, 2015.

[7] L. D. Vignolo, H. L. Rufiner, D. H. Milone, and J. C. Goddard, "Evolutionary cepstral coefficients". *Appl. Soft Comput.*, vol. 11, no. 4, pp. 3419–3428, Jun. 2011.

[8] L. Muda, M. Begam, and I. Elamvazuthi, " Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" . *Journal of computing*, vol 2, Issue 3 Mar. 2010.

[9] J. I. Godino-Llorente and P. Gómez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors". *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–4, Feb. 2004.

[10] R. Maia, M. Akamine, and M. J. F. Gales, "Complex cepstrum for statistical parametric speech synthesis". *Speech Commun.*, vol. 55, no. 5, pp. 606–618, Jun. 2013.

[11] H. Hong, Z. Zhao, X. Wang, and Z. Tao, "Detection of Dynamic Structures of Speech Fundamental Frequency in Tonal Languages". *IEEE Signal Process. Lett.*, vol. 17, no. 10, pp. 843–846, Oct. 2010.

[12] S. Sunny, D. P. S., and K. P. Jacob, "Feature Extraction Methods Based on Linear Predictive Coding and Wavelet Packet Decomposition for Recognizing Spoken Words in Malayalam". in 2012 *International Conference on Advances in Computing and Communications*, 2012, pp. 27–30.

[13] J. D. Wu and B. F. Lin, "Speaker identification based on the frame linear predictive coding spectrum technique". *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8056–8063, May 2009.

[14] X. C. Yuan, C. M. Pun, and C. L. Philip Chen, "Robust Mel-Frequency Cepstral coefficients feature detection and dual-tree complex wavelet transform for digital audio watermarking". *Inf. Sci.* (Ny)., vol. 298, pp. 159–179, Mar. 2015.

[15] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction". in 2010 *4th International Conference on Signal Processing and Communication Systems*, pp. 1–5, 2010

[16] X. Zhou, D. Garcia, R. Duraiswami, C. Espy Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition". in 2011 *IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 559–564, 2011

[17] Y. Zhang, C. He, Y. Luo, K. Chen, and W. Xing, "Improved perceptually non-uniform spectral compression for robust speech recognition". *J. China Univ. Posts Telecommun.*, vol. 20, no. 4, pp. 122–126, Aug. 2013.

[18] D. Jurafsky and J. H. Martin, " Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition" . 2nd ed. Prentice hall, 2009.