Evaluación comparativa de sistemas de reconocimiento de locutor basados en los algoritmos LPC, CC y MFCC

Comparative Evaluation of Speaker Recognition Systems Based on the LPC, CC and MFCC Algorithms.

Yesenia González[11] Héctor Juárez[2], Oscar Rocha[3], Rubén Hernández[4], Alfredo Bermúdez[5].

Recibido: 06/2019 Aceptado: 09/2019

Resumen. - El presente documento propone realizar la evaluación de sistemas de reconocimiento de locutor basados en los algoritmos LPC (Coeficientes de Predicción Lineal), CC (Coeficientes Cepstrales) y MFCC (Coeficientes Cepstrales en Frecuencias Mel), empleados en la extracción de parámetros de voz. La evaluación, siguiendo una metodología cuantitativa experimental, consiste en determinar el cambio de desempeño cuando la señal de entrada es expuesta a diferentes condiciones de ruido (bullicio y gaussiano), es decir, a distintos niveles de SNR, comparando los resultados de verificación para 2 locutores. Aunque todos los sistemas disminuyen su desempeño en ambientes ruidosos, cada uno posee de forma intrínseca cierto nivel de robustez. Esta evaluación servirá de referencia en la construcción de sistemas de reconocimiento de locutor, los cuales incluyan sistemas de mejora de voz para disminución del ruido.
Palabras clave: Reconocimiento de locutor, ruido de bullicio, algoritmo MFCC, algoritmo CC, algoritmo LPC.

Summary. -*This document proposes the evaluation of speaker recognition systems based on the LPC (Linear Predicting Coding), CC (Cepstral Coefficients) and MFCC (Mel Frequency Cepstral Coefficients) algorithms, used in the extraction of voice parameters. The evaluation, following an experimental quantitative methodology, consists of determining the change in performance when the input signal is exposed to different noise conditions (crowd and Gaussian noise), namely, at different levels of SNR, comparing the verification results for 2 speakers. Although all the systems decrease their performance in noisy environments, each one possesses intrinsically a certain level of robustness. This evaluation will serve as a reference in the construction of speaker recognition systems, which include voice enhancement systems to reduce noise.*
Keywords: *Speaker recognition systems, crowd noise, MFCC algorithm, CC algorithm, LPC algorithm.*

[1] Instituto Politécnico Nacional. ygonzalezn@ipn.mx, ORCID 0000-0003-2370-4660
2 Instituto Politécnico Nacional. hjuarezl1400@alumno.ipn.mx, ORCID 0000-0003-1347-0645
3Instituto Politécnico Nacional. orochaa1400@alumno.ipn.mx, ORCID 0000-0002-1676-6620
4Instituto Politécnico Nacional. rhtovar@ipn.mx, ORCID 0000-0001-7059-6426
5Instituto Politécnico Nacional. jbermudezs@ipn.mx, ORCID 0000-0001-5714-0061

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

1 Introducción. - Los sistemas de procesamiento de voz se han desarrollado a la par que las comunicaciones digitales y su principal uso es en los sistemas de telefonía. Además, existen aplicaciones en otros campos, como en el reconocimiento de palabras y reconocimiento de locutor.

En éste análisis se plantea el método de la verificación de locutor, el cual consiste en obtener un análisis de las características de la voz, las cuales están determinadas por factores físicos del locutor (tráquea, laringe y labios) y por el hábito o manera de hablar (ritmo, entonación y pronunciación), que permiten la comparación con una referencia y que, a su vez, pueda definir si la persona que habla es quien se encuentra en el registro.

En particular, el reconocimiento de locutor tiene utilidad en sistemas de acceso seguro y en cuestiones forenses, es decir, el procesamiento de voz para el reconocimiento de locutor puede auxiliar en la restricción de acceso por medios electrónicos, donde sirve para la identificación de la persona, con base en parámetros característicos de su voz. Asimismo, puede auxiliar en procedimientos en ciencias criminalísticas para determinar si la voz de una persona coincide con algún actor involucrado en algún hecho delictivo o bajo proceso legal [1].

En los sistemas de reconocimiento de locutor se han empleado diversos métodos de extracción de parámetros de voz como lo son: Coeficientes de Predicción Lineal (LPC, por sus siglas en inglés), Coeficientes Cepstrales (CC, por sus siglas en inglés), Coeficientes Cepstrales en Frecuencias Mel (MFCC, por sus siglas en inglés), bajo diferentes circunstancias y en combinación con diferentes sistemas de clasificación incluyendo redes neuronales, distancias específicas y sistemas difusos [2].

Resulta de interés medir la eficiencia, bajo ambientes ruidosos, de los diferentes procedimientos de extracción de parámetros de voz: LPC, CC y MFCC. Estos procedimientos, han sido evaluados con diferentes propósitos, y son utilizados ampliamente en diferentes sistemas modernos. Por ejemplo, los LPC se emplean para la síntesis de voz en telefonía celular y telefonía IP [3], mientras que los MFCC son preferidos para el reconocimiento de voz en sistemas de dictado o de conversación emulada [4] [5].

Un método para poder medir la eficiencia de un sistema de reconocimiento de voz es el porcentaje de éxito al reconocer un locutor, es decir, si de 100 vectores de prueba el sistema identifica al locutor en todos los casos, para este conjunto se tendrá un 100% de éxito. El incorrecto reconocimiento puede ser causa de: los efectos sonoros que hay en el ambiente, el ruido que se agrega a la señal digitalizada, los dispositivos y/o materiales (propiedades acústicas, mala calidad de los materiales) empleados en la captura de la voz, el algoritmo empleado en el sistema, la mala pronunciación del hablante, entre otros.

2 Desarrollo. - La investigación aquí desarrollada pretende brindar información para la selección de los sistemas de reconocimiento de locutor de acuerdo con su porcentaje de reconocimiento, viabilidad de aplicación en ambientes ruidosos y la cantidad de coeficientes que son usados por sus algoritmos de extracción de características.

El trabajo consiste en la programación de tres sistemas de reconocimiento de locutor que difieren en la etapa de extracción de características. Un diagrama a bloques de un sistema de reconocimiento de locutor se puede observar en la Figura I.

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

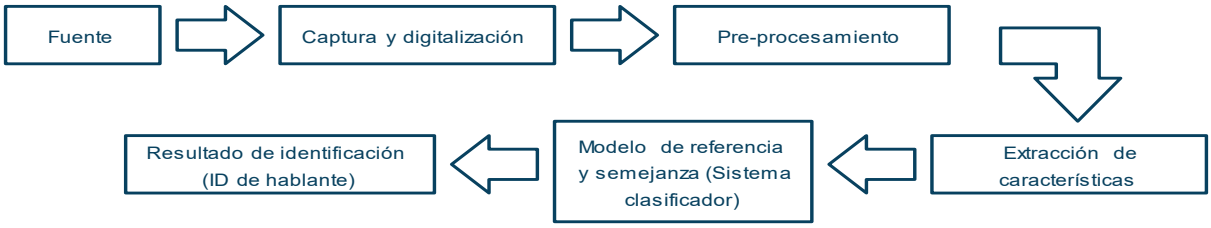

Figura I. Etapas en un sistema de reconocimiento de locutor [6].

Cada sistema es evaluado con las diferentes señales de voz originales, obteniéndose así un conjunto de estadísticas de desempeño. Posteriormente, se evalúan reiteradamente empleando las señales con diferentes niveles de ruido aditivo (ruido de bullicio y ruido gaussiano), lo cual, permite comparar las variaciones en el desempeño de cada sistema respecto al obtenido con las señales originales. Es importante resaltar que no se encontraron trabajos previos donde se realice la comparativa de reconocimiento de locutor entre los 3 algoritmos aquí presentados y que incluyan una variación del SNR (relación señal a ruido) en las señales de entrada al sistema. Si bien, existen estudios comparativos [7] de rendimiento entre los algoritmos LPC y CC que obtienen como conclusión que la carga computacional para LPC es menor en comparación a CC, los cuales coinciden con los resultados obtenidos en esta investigación.

Se decidió utilizar una *metodología cuantitativa* de tipo *experimental* debido a que esta se caracteriza, básicamente, por la manipulación intencional de una o más variables independientes (nivel de SNR y número de coeficientes empleados en los algoritmos de extracción de características), para observar/medir su influencia en una o más variables dependientes (porcentaje de reconocimiento correcto).

Para el desarrollo del proyecto, se recolectaron grabaciones de voz de los dos alumnos participantes en el proyecto, pronunciando 9 palabras con 10 repeticiones cada una. Para la recolección del ruido de bullicio, se realizaron grabaciones de audio dentro de la Unidad Académica de los autores, en lugares donde se encontraban entre 10-30 personas hablando al mismo tiempo.

2.1 Algoritmos de reconocimiento de locutor. - Como se ha mencionado, el trabajo implementa los algoritmos LPC, CC y MFCC. A continuación, se realiza una descripción de cada uno de ellos.

2.1.1 Codificación de predicción lineal (LPC).- El modelo de Codificación de Predicción Lineal es considerado uno de los modelos más próximos (analógicamente hablando) al sistema vocal humano [8].

Lo fundamental de este modelo es representar una señal de voz como una función de excitación constituida por un tren de pulsos cuasi periódicos (para sonidos vocalizados) o una fuente de ruido aleatorio (para sonidos no vocalizados). La idea básica del modelo LPC es que una muestra de voz dada en un tiempo n, se puede aproximar mediante una combinación lineal de las últimas p muestras de voz, de manera que:

$$\hat{s}(n) = -a_1 s(n-1) - a_2 s(n-2) - \ldots - a_p s(n-p), \quad (1)$$

donde los coeficientes a_1, a_2, …, a_p representan los coeficientes LPC [9].

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k), \quad (2)$$

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

se define al error de predicción $e(n)$ como:

$$\boldsymbol{e(n) = s(n) - \hat{s}(n)}, \tag{3}$$

sustituyendo (2) en (3) se tiene:

$$\boldsymbol{e(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k)}; \tag{4}$$

el error de una trama E_T de tamaño N se define como:

$$\boldsymbol{E_T = \sum_n e^2(n) = \sum_n \left(s(n) + \sum_{k=1}^{p} a_k \cdot s(n-k)\right)^2, \quad 0 < n < N}. \tag{5}$$

Para minimizar el error de predicción respecto al conjunto de parámetros a_k, se suponen coeficientes a_j para los cuales:

$$\boldsymbol{\frac{\partial E_T}{\partial a_j} = 0}. \qquad\qquad \boldsymbol{j = 1, 2, .., p}. \tag{6}$$

Siendo la derivada de una sumatoria, la suma de sus derivadas, y teniendo la función de la forma: $y = x^n$; su derivada resulta $y' = n \cdot x^{n-1} dx$, donde dx es la derivada de la suma.

$$\boldsymbol{\frac{\partial E_T}{\partial a_j} = 2 \sum_n \left(s(n) + \sum_{k=1}^{p} a_k \cdot s(n-k)\right) \cdot dx} \tag{7}$$

Dado que $\frac{\partial E_T}{\partial a_j}$ está derivando con respecto a j, el resultado de la derivada será 0 para todos los términos k diferentes de j. Resultando así:

$$\boldsymbol{\frac{\partial E_T}{\partial a_j} = 2 \sum_n \left(s(n) + \sum_{k=1}^{p} a_k \cdot s(n-k)\right) \cdot s(n-j) = 0}\ ; \ . \tag{8}$$

reacomodando términos, con a_k siendo los coeficientes de predicción óptimos

$$\boldsymbol{\sum_{k=1}^{p} a_k \cdot \sum_n s(n-k) \cdot s(n-j) = -\sum_n s(n) \cdot s(n-j)}\ , \tag{9}$$

con la expresión de autocorrelación r_{ss} definida como:

$$\boldsymbol{r_{ss}(j) = \sum_n s(n) \cdot s(n-j)}. \tag{10}$$

Se puede observar que en (9) existe una autocorrelación en la parte derecha de la ecuación y una autocorrelación retrasada k unidades en la parte izquierda. Teniendo así:

$$\boldsymbol{\sum_{k=1}^{p} a_k \cdot r_{ss}(j-k) = -r_{ss}(j)}\ . \tag{11}$$

Lo que constituye un conjunto de p ecuaciones con p incógnitas. Las ecuaciones que deben cumplir los parámetros LPC se pueden expresar como:

$$\begin{bmatrix} r_{ss}(0) & r_{ss}(-1) & \cdots & r_{ss}(-(p-1)) \\ r_{ss}(1) & r_{ss}(0) & \cdots & r_{ss}(-(p-2)) \\ \vdots & \vdots & \cdots & \vdots \\ r_{ss}(p-1) & r_{ss}(p-2) & \cdots & r_{ss}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} r_{ss}(1) \\ r_{ss}(2) \\ \vdots \\ r_{ss}(p) \end{bmatrix}; \tag{12}$$

como la función de autocorrelación es simétrica:

$$\boldsymbol{r_{ss}(-j) = r_{ss}(j)}; \tag{13}$$

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

teniendo así:

$$\begin{bmatrix} r_{ss}(0) & r_{ss}(1) & \cdots & r_{ss}(p-1) \\ r_{ss}(1) & r_{ss}(0) & \cdots & r_{ss}(p-2) \\ \vdots & \vdots & \cdots & \vdots \\ r_{ss}(p-1) & r_{ss}(p-2) & \cdots & r_{ss}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} r_{ss}(1) \\ r_{ss}(2) \\ \vdots \\ r_{ss}(p) \end{bmatrix}. \quad (14)$$

La matriz en (14) se puede resolver mediante el método recursivo de Levinson – Durbin [10] obteniendo así los coeficientes LPC.

2.1.2 Cepstrum y coeficientes cepstrales (CC).- De acuerdo con el modelo mostrado en Figura II, el habla está compuesta de una secuencia de excitación convolucionada con la respuesta al impulso del modelo del sistema vocal. Solo se tiene acceso a la salida, por lo que muchas veces es deseable eliminar algunas de sus componentes para que otras puedan ser examinadas, codificadas, modeladas, o usadas en un algoritmo de reconocimiento.

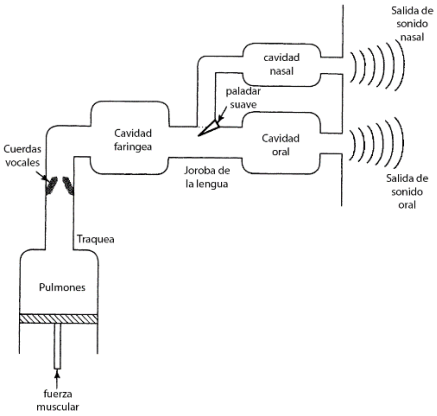

Figura II. Diagrama a bloques de producción de habla humano [11].

El cepstrum viene de aplicar una operación no lineal al espectro (función logaritmo) de una señal, y aplicar la operación contraria de la transformada de Fourier, de aquí su analogía al dominio espectral, teniendo así términos similares con silabas invertidas, por ejemplo, del inglés spectral, se invierte la primera silaba spec, a ceps, y se combina con la segunda silaba teniendo la palabra cepstral. Así como el espectro de una señal permite ver cada una de sus componentes en el dominio de la frecuencia, el cepstrum permite representar señales combinadas por convolución en las sumas de sus componentes cepstrales [9]. El proceso de extracción de los coeficientes cepstrales se muestra en el diagrama a bloques de la Figura III.

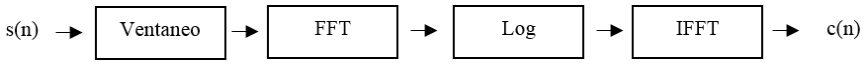

Figura III. Proceso de extracción de coeficientes cepstrales. siendo s(n) la señal de entrada y c(n) los coeficientes cepstrales [12].

La producción de la voz puede ser representada como un sistema con entradas y salidas. Por lo que, si la señal de voz es $s(n)$, la cual se compone de una excitación $v(n)$ que pasa a través de la respuesta al impulso de un sistema $h(n)$, este sistema se puede representar como:

$$\boldsymbol{s(n) = h(n) * v(n)}, \quad (15)$$

donde * representa la operación de convolución. Por propiedades de la transformada de Fourier, se puede obtener lo siguiente:

$$\boldsymbol{S(n) = H(n)V(n)}. \quad (16)$$

Ahora, al aplicar la función absoluto y logaritmo en (16), se puede separar la señal correspondiente a la excitación $v(n)$ con respecto al sistema $h(n)$:

$$\boldsymbol{log|S(n)| = log|H(n)| + log|V(n)|}, \quad (17)$$

sin embargo, aún se encuentra en el dominio de la frecuencia, para regresar al dominio del tiempo se aplica la transformada de Fourier inversa, llegando así a un nuevo dominio llamado dominio cepstral.

$$\boldsymbol{c_s(n) = c_h(n) + c_v(n)}, \quad (18)$$

donde los valores de $c_s(n)$ son los coeficientes cepstrales.

Siendo que el modelo representativo de una persona no es algo que varíe mucho, mientras que la excitación sí lo es, los primeros $c_s(n)$ coeficientes corresponden al modelo que identifica a la persona, mientras que los siguientes corresponden a la excitación [19].

2.1.3 Coeficientes cepstrales en la frecuencia de Mel (MFCC). - En el procesamiento de sonido, el cepstrum en frecuencias Mel es una representación del espectro de potencia a corto plazo de un sonido. El cepstrum en frecuencias Mel (MFC), realmente es un cepstrum con su espectro mapeado en la escala Mel. Los coeficientes cepstrales en frecuencias Mel (MFCC) son coeficientes que colectivamente forman un MFC. La diferencia entre el "cepstrum" y el "cepstrum en frecuencia de Mel" es que, en la MFC, las bandas de frecuencia están igualmente espaciadas en la escala de Mel, esto se aproxima más a la respuesta del sistema auditivo humano que las bandas de frecuencia linealmente utilizadas en el cepstrum normal. Esta distorsión de frecuencia puede permitir una mejor representación del sonido.

Ventaneo y tramas. - En todas las aplicaciones de procesamiento de señales, es necesario trabajar con términos cortos o tramas de la señal, necesitando hacer la selección de esas tramas o intervalos elegidos; por lo que se requerirá hacer uso de una función de ventana.

Una función de ventana $w(n)$, es una secuencia real de longitud finita N usada para seleccionar una trama deseada de la señal, y que tiene un valor cero fuera del intervalo elegido.

Escala de Mel. - La escala de Mel relaciona la frecuencia percibida, o el tono, de un tono puro con su frecuencia real medida. Por ejemplo, los seres humanos son mucho mejores para discernir pequeños cambios en el tono en las frecuencias bajas a comparación de las frecuencias altas. Por lo tanto, esta escala hace que las características coincidan más estrechamente con lo que escuchan los humanos.

La escala Mel fue desarrollada por Stanley Smith Stevens, John Volkman y Edwin Newman [13]. El nombre "Mel" viene de la palabra "melodía", para indicar que la escala está basada en comparaciones de tonos. El punto de referencia entre esta escala y la frecuencia normal se define equiparando un tono de 1000 Hz, 40 dB por encima del umbral de audición

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

del oyente, con un tono de 1000 Mels. La fórmula empleada para calcular alguna frecuencia f en Hertz a m de la escala Mel [13], se describe en (19). La operación inversa se describe en (20). La Figura IV describe el comportamiento de las ecuaciones (18) y (19).

$$\boldsymbol{m = 2595 log_{10}\left[1 + \frac{f}{700}\right] = 1127 ln\left[1 + \frac{f}{700}\right]}, \tag{19}$$

$$\mathbf{f = 700\left(e^{\frac{m}{1127}} - 1\right)}. \tag{20}$$

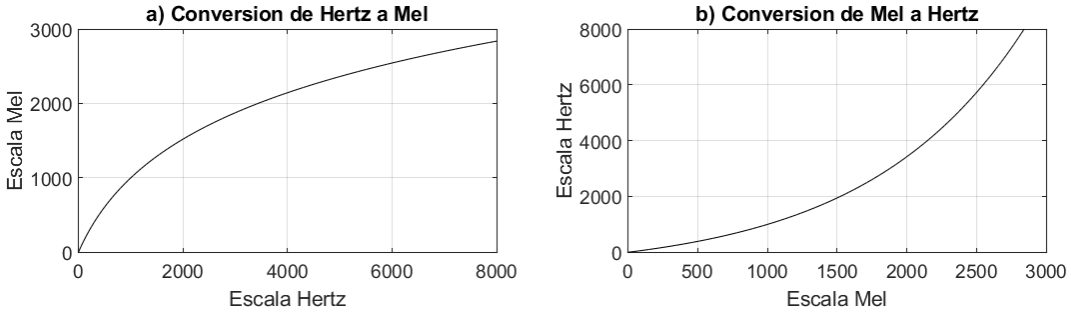

Figura IV. a) Gráfica de escala Mel versus escala Hertz. b) Gráfica de escala de Hertz versus escala de Mel.

Una vez filtrada la señal, se obtiene la potencia promedio de cada vector resultante.

$$\mathbf{\frac{1}{N}\sum_{k=0}^{N}|x(k)|^2}\,, \tag{21}$$

y después, se aplica la función logaritmo y por último la transformada discreta de coseno.

$$\boldsymbol{c_s(n) = \sum_{k=1}^{N} S(k) cos\left[\pi n\left(\frac{k-\frac{1}{2}}{N}\right)\right]}, \qquad \boldsymbol{1 \le n \le N} \tag{22}$$

El resultado de la aplicación de las ecuaciones (20) y (21) es la obtención del espectro de potencia de la señal en los intervalos de frecuencia establecidos por la escala Mel. El cepstrum en frecuencias Mel (MFC), realmente es un cepstrum con su espectro mapeado en la escala Mel antes de aplicar la operación logarítmica y la transformada inversa de Fourier, como lo mostrado en la Figura V.

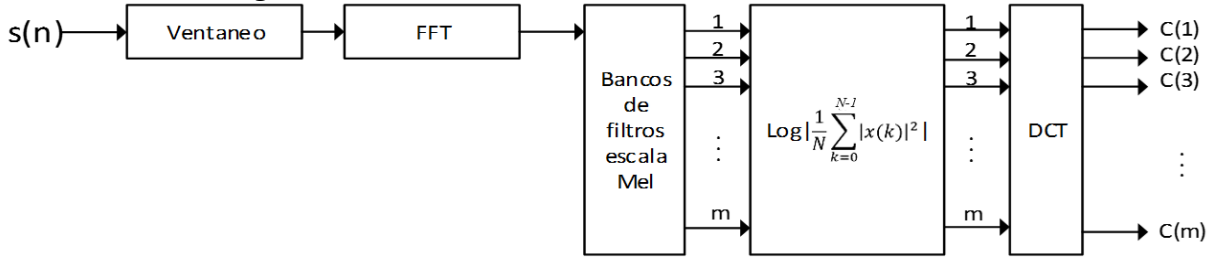

Figura V. Proceso de extracción de MFCC [20].

Al multiplicar la señal por un banco de filtros triangulares (Figura VI) se logra obtener las componentes de frecuencia que le aporta la señal analizada a cada banda del banco de filtros. Comúnmente se usan 24 filtros triangulares espaciados de acuerdo con la escala de frecuencias Mel. Con estos filtros se calcula el promedio del espectro alrededor de cada frecuencia central.

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

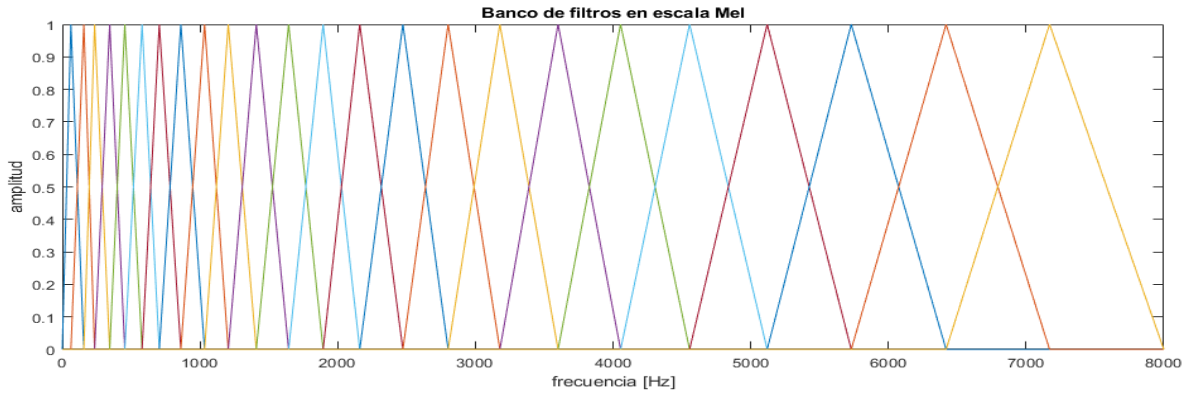

Figura VI. Banco de filtros triangulares superpuestos en escala de Mel.

Cada banco de filtros corresponde a la función (23), donde H_m corresponde al m-ésimo filtro. $1 \leq m \leq M+2$; f_m corresponde a las $M+2$ frecuencias Mel-espaciadas de los M filtros y k es la frecuencia en Hertz.

$$H_m(k) \begin{cases} 0 & k < f_{m-1} \\ \dfrac{k - f_{m-1}}{f_m - f_{m-1}} & f_{m-1} \leq k \leq f_m \\ \dfrac{f_{m+1} - k}{f_{m+1} - f_m} & f_m \leq k \leq f_{m+1} \\ 0 & k > f_{m+1} \end{cases} \tag{23}$$

2.2 Diseño del sistema de reconocimiento de locutor.- Tomando como base el diagrama a bloques de la Figura I, en la Tabla I se describen cada una de las etapas de los sistemas de reconocimiento de locutor propuestos.

Tabla I. Descripción de las etapas de los sistemas de reconocimiento de locutor basados en los Algoritmos LPC, CC y MFCC.

Etapa	Descripción
Fuente	La fuente corresponde a las señales de audio grabadas de diferentes locutores, estas constan de emisiones de audio en idioma español. Se propone obtener muestras originales consideradas sin ruido y versiones diferentes de las mismas, pero con diferentes niveles de ruido, es decir con variaciones en el nivel de relación señal a ruido (SNR, por sus siglas en inglés). Las señales de ruido que se agregan a las muestras originales se pueden generar de forma artificial para el caso de ruido gaussiano y por la mezcla de diferentes voces para el caso de ruido de bullicio.
Captura y digitalización	Esta etapa consta de un transductor (micrófono) que convierte las vibraciones sonoras en variaciones de voltaje, después, mediante un convertidor analógico digital (ADC, por sus siglas en inglés) se realiza un muestreo de la señal.
Pre procesamiento	En la etapa de pre procesamiento se realiza el ventaneo y se aplican filtros para adecuar la señal antes de ser procesada. En esta investigación se propone utilizar un filtro preénfasis, ya que los segmentos de voz sonoros tienen una pendiente espectral negativa, este filtro permite contrarrestar esta pendiente, mejorando así la eficiencia de las etapas posteriores. Se utiliza una ventana de Hamming

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

	(proceso de ventaneo) para poder eliminar problemas causados por los cambios rápidos en los extremos de las tramas de voz.
Extracción de características	Para la extracción de las características del hablante se aplican los algoritmos LPC, CC y MFCC en los que se basa esta investigación. Cada algoritmo se implementó en Matlab el cual a comparación de Python permite un manejo sencillo de vectores en tiempo real gracias a su Workspace. Cabe resaltar que no se usaron las funciones nativas para una posible implementación en cualquier otro lenguaje de programación; se hizo la validación de estos algoritmos comparándolos con los de Matlab y, también se corroboraron con la reconstrucción del espectro de la señal original.
Modelo de referencia y semejanza	El modelo de referencia utilizado es una red neuronal, debido a su capacidad de clasificación y aprendizaje para determinadas tareas, con solamente datos y sin necesidad de conocer la fuente. Las redes neuronales se componen de dos etapas: etapa de entrenamiento y etapa de validación [14]. En la etapa de entrenamiento la red neuronal se ajusta de tal manera que relaciona los valores de entrada a su salida correspondiente, en este caso, relaciona cada grabación de voz con el usuario a verificar o rechazar. En la etapa de validación se introducen grabaciones que no se hayan utilizado en la etapa de entrenamiento, pudiendo tener un resultado exitoso o erróneo en la verificación de locutor, calculando así el porcentaje de eficiencia de la red.

2.2.1 Red neuronal .- Como modelo de referencia se utilizó una red neuronal tipo perceptrón multicapa con el algoritmo de entrenamiento de retropropagación, este tipo de red es el más usado [15]. Para encontrar la topología de la red que diera mejores resultados, se hizo variar entre el número de capas y número de neuronas, probando redes con 1, 3 y 5 capas ocultas, y usando valores de 16, 64 y 256 en el número de neuronas de cada una de las redes; como resultado, la topología con el mejor rendimiento fue de 5 capas ocultas y 16 neuronas en cada capa.

2.3 Variables de estado. - La Tabla II describe las diferentes variables de estado aplicadas a los sistemas de reconocimiento de locutor. Para determinar el número de bits de resolución, se utilizó la siguiente ecuación de error de cuantificación $SQNR_{dB}$:

$$\boldsymbol{SQNR_{dB} = 20\log_{10}(2^{n})}, \tag{24}$$

donde n es el número de bits. Para determinar el número de muestras por trama se utilizó:

$$\boldsymbol{T = fs * tamaño}. \tag{25}$$

Tabla II. Descripción de variables de estado utilizadas.

Variable	Descripción
Frecuencia de muestreo	Se eligió la frecuencia de 16000 Hz. Se conoce como audio de banda ancha o audio de alta definición, debido a que permite una mejor calidad al grabar la voz, permitiendo grabar así hasta 8 hasta fonemas. Esta frecuencia de muestreo es utilizada en aplicaciones modernas de VoIP [16].
Bits de resolución	Se utilizaron 16 bits/muestra para minimizar el error de cuantificación. Con esta resolución utilizando la ecuación (23), el error de cuantificación $SQNR_{dB}$ máximo es 96.32 dB.
Tamaño de la ventana/trama	El tamaño elegido es de $32\,ms$, ya que en el artículo [17] se realizó un experimento para comparar distintos tamaños de ventana al grabar 6 consonantes oclusivas (consonantes cuyo tipo de sonido consonántico obstruyente es producido por una detención del flujo de aire y por su posterior

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

	liberación. Ejemplo: [b, d, g, p, t, k].). El resultado dictó que $20 - 40\,ms$ es un tamaño apropiado para el correcto entendimiento de las palabras. Con una $fs = 16000\,Hz$, aplicando (24), el número de muestras en una trama T será de $T = fs * tamaño = 16000 * 0.32 = 512$. Es decir, cada trama contendrá 512 muestras por segundo, que es una potencia de 2, facilitando así el cálculo de la FFT.
Número de puntos en FFT	Se utilizan 512 puntos, correspondiente con el tamaño de la trama.
Número de repeticiones	Se eligieron 10 repeticiones, de las cuales 8 se utilizan para entrenar la red neuronal artificial y 2 repeticiones se utilizan para probar la red.
Palabras a utilizar	Estas palabras se seleccionaron basándose en [18], debido a que sus características fonéticas proporcionan más información (en estas palabras las cuerdas vocales vibran), que caracteriza a cada locutor que pertenecerá al sistema: "gama, mano, llave, barra, niño, lobo, mayo, daga, lava".
Grabaciones con ruido de bullicio	Se han realizado grabaciones dentro de la Unidad Académica, buscando los lugares con mayor cantidad de personas hablando, destacando la cafetería de la Unidad.
Formato de audio	Se hizo la elección del formato ".wav", tomando en cuenta que es un formato sin compresión, que es de los más usados y es soportado por una gran cantidad de reproductores y programas.
Número de coeficientes en los algoritmos	El número de coeficientes se puede seleccionar dentro del sistema, sin embargo, siguiendo las recomendaciones de [2], se han utilizado entre 16-40 coeficientes en cada algoritmo de extracción de características.

3 Resultados. - Esta sección describe la validación de los algoritmos utilizados y los distintos escenarios de prueba que se llevaron a cabo.

3.1 Validación de algoritmos. - Para corroborar la implementación de los algoritmos, es decir, que extraen de forma correcta las características de la voz, se realizó una aproximación al espectro de voz del locutor, basándose en los coeficientes de cada algoritmo. La Figura VII muestra la reconstrucción del espectro de voz para los algoritmos LPC, CC y MFCC respectivamente usando 8, 16, 32, 64, 128 y 256 coeficientes.

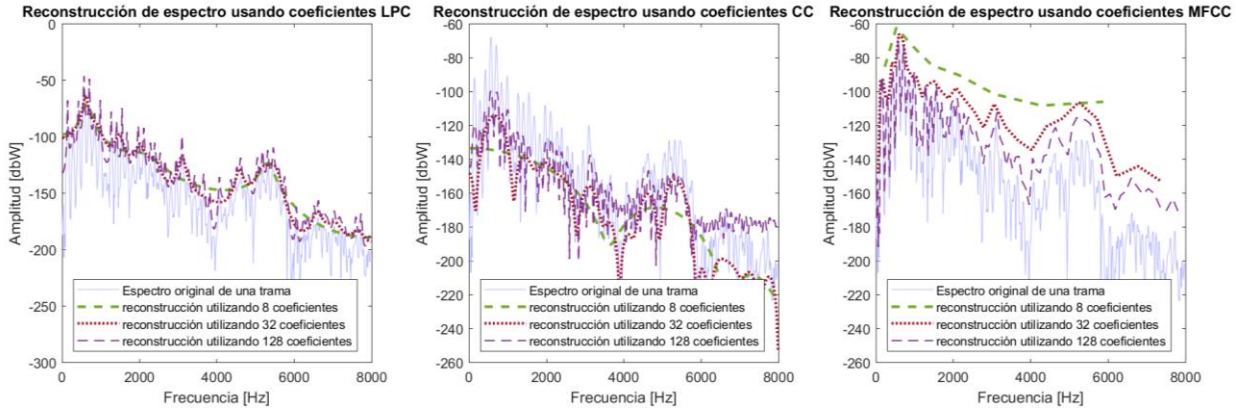

Figura VII. Espectro aproximado de voz variando los coeficientes LPC, CC y MFCC.

Se realizó una comparación entre los algoritmos implementados y los nativos de Matlab, para una palabra con 30 tramas de 32ms los resultados se muestran en la Tabla III. Para los algoritmos LPC y CC las funciones nativas muestran un menor tiempo, mientras que en MFCC es lo contrario, al revisar la documentación de Matlab se justifica el mayor tiempo por el banco de filtros el cual presenta además de estar en intervalos logarítmicos, presenta

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

amplitudes logarítmicas. Otra explicación, es que la función de Matlab nativa calcule el banco de filtros por cada vez que se manda a llamar la función, mientras que el algoritmo que se implementó permite guardar el banco de filtros en una variable para evitar tener que calcularlo en cada iteración. Por otra parte, al usar la función FFT nativa de Matlab se explica que el error cuadrático medio del algoritmo CC sea 0.

Para los algoritmos LPC y CC las funciones nativas muestran un menor tiempo, mientras que en MFCC es lo contrario, al revisar la documentación de Matlab se justifica el mayor tiempo por el banco de filtros el cual presenta además de estar en intervalos logarítmicos, presenta amplitudes logarítmicas. Otra explicación, es que la función de Matlab nativa calcule el banco de filtros por cada vez que se manda a llamar la función, mientras que el algoritmo que se implementó permite guardar el banco de filtros en una variable para evitar tener que calcularlo en cada iteración. Por otra parte, al usar la función FFT nativa de Matlab se explica que el error cuadrático medio del algoritmo CC sea 0.

Tabla III. Resultados con coeficientes de orden 30.

	Algoritmo LPC	**Algoritmo CC**	**Algoritmo MFCC**
Complejidad computacional	O(Nm)	O($Nlog_2N$)	O($Nm + Nlog_2N$)
Tamaño de la matriz resultante	30x32	30x512	30x32
Tiempo de ejecución Matlab [s]	0.016263	0.006297	0.297807
Tiempo de ejecución propio [s]	0.020659	0.010825	0.063452
Diferencia de tiempos	-0.004396	-0.004528	0.234355
Error cuadrático medio. [Amplitud^2]	7.2166e-20	0	0.8799
Error cuadrático medio [dB]	-881.5064	-Inf	-2.5596

Al realizar la prueba de las 9 propuestas con 10 repeticiones (Figura VIII.a), se observó que existen puntos (representaciones de cada muestra) que están sobrepuestos entre cada locutor, a diferencia de realizar la prueba para una palabra con 100 repeticiones (Figura VIII.b).

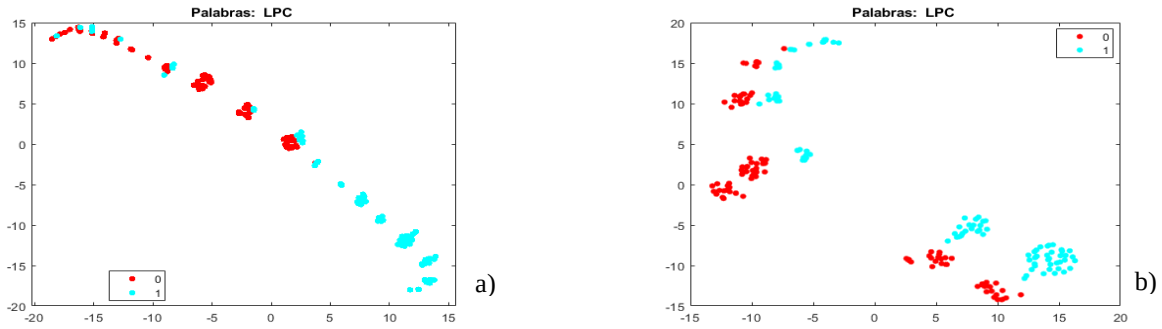

Figura VIII. a) 9 palabras con 10 repeticiones, b) 1 palabra con 100 repeticiones.

3.2 Escenarios de prueba. - Para la selección de los escenarios de prueba, se utilizaron 32 coeficientes en los algoritmos de extracción, se programó una red neuronal con 5 capas ocultas y 16 neuronas por capa y un entrenamiento con 10,000 épocas para cada una de las palabras. El resultado del porcentaje de reconocimiento por cada locutor es mostrado en la Tabla III.

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

Tabla IV. Porcentaje de reconocimiento de locutor de una red neuronal entrenada por palabra.

Palabras	**Locutor 1**			**Locutor 2**		
	LPC	CC	MFCC	LPC	CC	MFCC
Barra	75	75	75	75	75	75
Daga	50	50	75	75	75	75
Gama	100	100	100	100	100	100
Lava	75	75	75	75	75	75
Llave	75	50	75	75	75	75
Lobo	100	100	100	100	100	100
Mano	75	75	75	50	50	50
Mayo	100	100	100	75	75	75
Niño	100	75	75	75	100	100
Murciélago	75	75	75	75	50	100
Aurelion	100	75	75	75	100	100

Tomando como punto de partida la palabra “Gama” en la que se obtuvo mayor porcentaje en los tres algoritmos de reconocimiento, se realizaron pruebas utilizando un mayor número de repeticiones, procurando tener los mismos parámetros (entonación y rapidez) en ambos usuarios para después corromper las muestras de voz con ruido (bullicio y gaussiano) y observar los resultados de eficiencia de estos sistemas.

3.3 Discusión de resultados.- El porcentaje de reconocimiento de locutor para las muestras contaminadas con ruido gaussiano con una variación en la SNR de 60 dB a -20 dB se muestra en la Tabla V. La Figura IX y la Figura X muestran las gráficas respectivas. Para ambos locutores, el algoritmo que presentó mejor desempeño fue el de MFCC, seguido del algoritmo LPC (el porcentaje de reconocimiento correcto logra mantenerse para valores más pequeños de SNR). Para el caso del algoritmo CC, incluso con valores altos de SNR mantuvo un bajo desempeño.

Tabla V. Porcentaje de reconocimiento de locutor para la palabra “Gama” obtenido al variar la SNR con ruido gaussiano y utilizando 16 y 32 coeficientes.

	Locutor 1						**Locutor 2**					
	LPC		CC		MFCC		LPC		CC		MFCC	
SNR [dB]	16 c.	32 c.	16 c.	32 c.	16 c.	32 c.	16 c.	32 c.	16 c.	32 c.	16 c.	32 c.
60	100	100	5	5	97.5	100	100	100	7.5	5	100	100
50	97.5	100	5	7.5	97.5	100	100	100	7.5	5	100	100
40	92.5	92.5	5	7.5	97.5	100	97.5	97.5	7.5	5	100	100
30	52.5	62.5	5	5	97.5	97.5	70	82.5	7.5	5	100	100
20	50	50	5	5	95	95	40	50	7.5	5	100	85

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

10	50	32.5	5	5	70	70	42.5	45	7.5	5	72.5	55
0	50	2.5	5	5	55	52.5	45	45	7.5	5	55	50
-10	50	7.5	5	5	50	47.5	37.5	35	7.5	5	55	40

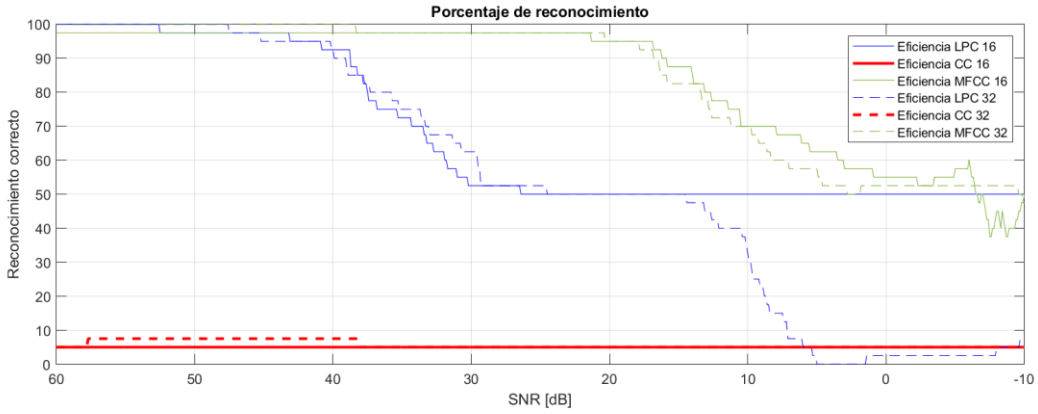

Figura IX. Porcentaje de reconocimiento en locutor 1 con ruido gaussiano.

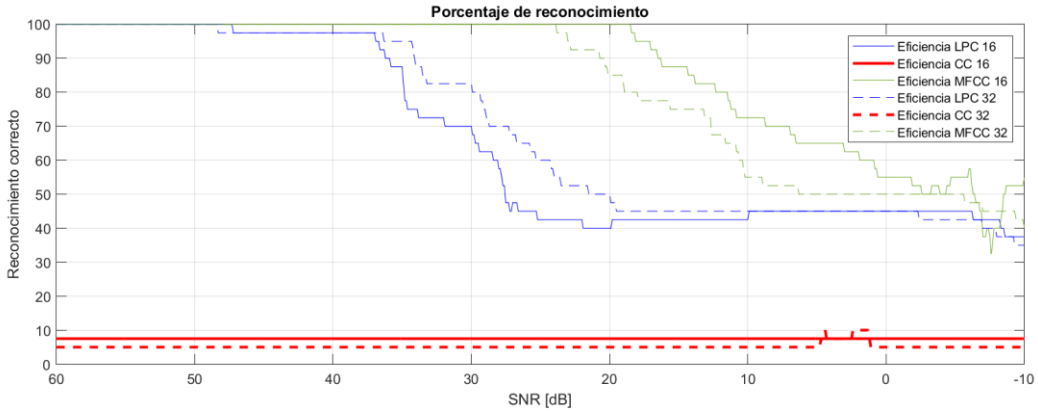

Figura X. Porcentaje de reconocimiento de locutor 2 con ruido gaussiano.

El porcentaje de reconocimiento de locutor para las muestras contaminadas con ruido de bullicio con una variación en la SNR de 60 dB a -20 dB se muestra en la Tabla VI. La Figura XI y la Figura XII muestran las gráficas correspondientes. Los algoritmos presentaron comportamientos similares a los obtenidos al aplicar ruido gaussiano, es decir, el algoritmo MFCC obtuvo el mejor desempeño, seguido por el algoritmo LPC y muy por debajo el desempeño del algoritmo CC.

Tabla VI. Porcentaje de reconocimiento de locutor para la palabra "Gama" obtenido al variar la SNR con ruido de bullicio y utilizando 16 y 32 coeficientes.

	Locutor 1						**Locutor 2**					
	LPC		CC		MFCC		LPC		CC		MFCC	
SNR [dB]	16 c.	32 c.	16 c.	32 c.	16 c.	32 c.	16 c.	32 c.	16 c.	32 c.	16 c.	32 c.
60	100	100	5	5	97.5	100	100	100	7.5	5	100	100
50	97.5	100	5	7.5	97.5	100	100	100	7.5	5	100	100

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

40	100	100	5	5	97.5	100	97.5	100	7.5	5	100	100
30	92.5	92.5	5	5	97.5	97.5	95	100	7.5	5	100	100
20	72.5	57.5	5	5	95	95	70	87.5	7.5	5	95	87.5
10	50	50	5	5	80	65	50	57.5	7.5	5	82.5	62.5
0	50	42.5	5	5	65	60	40	45	7.5	5	50	52.5
-10	50	37.5	5	7.5	40	50	42.5	35	7.5	7.5	50	32.5

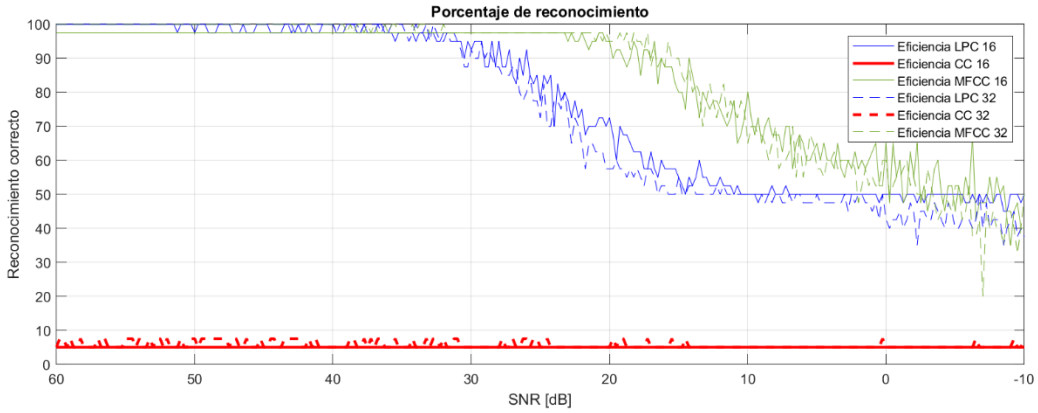

Figura XI. Porcentaje de reconocimiento de locutor 1 con ruido de bullicio.

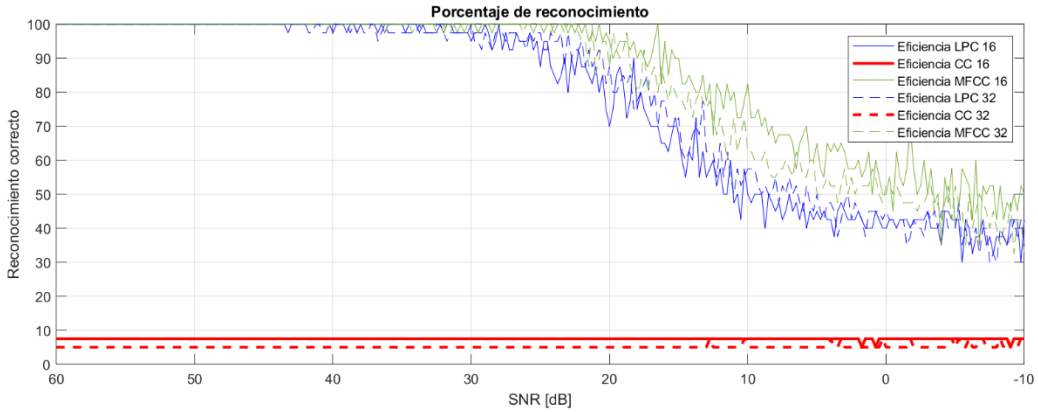

Figura XII. Porcentaje de reconocimiento de locutor 2 con ruido de bullicio.

Al variar la SNR de las señales de entrada, a partir de los 40 dB's se observó un decremento en la eficiencia de los sistemas LPC mientras que los sistemas MFCC mostraron soportar hasta 20 dB's antes de disminuir su eficiencia. En la variación de los coeficientes, las gráficas tienen un comportamiento similar, por lo que no es un factor para mejorar la robustez del algoritmo frente al ruido. En cuanto a las diferencias de ruido gaussiano y de bullicio, en este último las gráficas presentan un comportamiento más inestable, esto podría ser por una mayor presencia de falsos positivos, debido a que se el ruido de bullicio proviene del mismo tipo de fuente que se desea reconocer (locutores). Por otro lado, los sistemas con LPC mostraron un decremento mayor frente al ruido gaussiano, llegando al 50% de reconocimiento en los 25 dB.

Con respecto a la red neuronal, la topología influyó en el porcentaje de reconocimiento que se obtuvo en los sistemas antes de probarlos con señales con ruido. Para poder escoger la topología de red a utilizar, se realizaron pruebas variando el número de capas y el número de

https://doi.org/10.36561/ING.17.6
ISSN 2301-1092 • ISSN (en línea) 2301-1106

neuronas de la red. Finalmente, se observó una mejor relación entre porcentaje de reconocimiento y tiempo de procesamiento al usar 5 capas ocultas con 16 neuronas en cada capa.

4 Conclusiones. - En este trabajo se han implementado sistemas de verificación de locutor utilizando los algoritmos LPC, CC y MFCC de extracción de características de la voz. Para llevar a cabo estos sistemas, se desarrolló e implementó una red neuronal tipo perceptrón multicapa entrenada con el algoritmo de retropropagación.

Los resultados muestran que el ocupar más coeficientes en los algoritmos, no implica que estos sean más resistentes al ruido, sin embargo, la necesidad de mayor número de coeficientes para obtener un mayor porcentaje de reconocimiento se hace presente en el algoritmo CC. Con respecto al tipo de palabras a utilizar, el usar palabras con todas las vocales, no hace al sistema más robusto. Un punto para resaltar es que la utilización de una palabra con varias repeticiones, a diferencia de varias palabras con pocas repeticiones, presenta una mejora en la verificación de locutor (reconocimiento). Las gráficas de porcentaje de reconocimiento cuando están bajo el ruido de bullicio presentan un comportamiento más caótico que cuando están bajo el ruido gaussiano, sin embargo, el desempeño de los algoritmos con ambos tipos de ruido utilizados se mantuvo.

5 Referencias

[1] T. Kinnunen y H. Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12-40, 2010.

[2] H. Beigi, *Fundamentals of speaker recognition*. New York, USA: Springer, 2011.

[3] M. Ray, M. Chandra y B. Patil, "Speech Coding Techniques for VoIP Applications: A Technical Review," *World Applied Sciences Journal*, vol. 33 no. 5, pp. 736-743, 2015.

[4] C. Ittichaichareon, S. Suksri y T. Yingthawornsuk, "Speech Recognition using MFCC," en International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012), Pattaya (Thailand), 2012, pp. 135-138.

[5] X. Jing, J. Ma, J. Zhao y H. Yang, "Speaker recognition based on principal component analysis of LPCC and MFCC.," en International Conference on Trends in Automation, Communications and Computing Technology, 2015, December, pp. 1-5.

[6] A. Charisma, M. Reza Hidayat y Y. Bakti Zainal, "Speaker Recognition Using Mel-Frequency Cepstrum Coefficients and Sum Squar," Engineering Faculty of Universitas Jenderal Achmad Yani, Cimahi, Indonesia, 2017.

[7] A. N. Chadha, J. H. Nirmal y P. Kachare, A comparative performance of various speech analysis-synthesis techniques, International Journal of Signal Processing Systems, vol. 2, no. 1, Jun., pp. 17-22, 2014.

[8] P. Manrique Ramírez y M. A. Meléndez Velázquez, «Diseño de un sistema de codificación de predicción lineal (LPC),» Ciudad de México, 1999.

[9] L. Rabiner y B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs: Petrince Hall International, 1993.

[10] C. Collomb, "Tutorial on linear prediction and Levinson Durbin," *Empty Loop*, febrero 3, 2009. [En línea]. Disponible en: http://www.emptyloop.com/technotes/A%20tutorial%20on%20linear%20prediction%20and%20Levinson-Durbin.pdf. [Último acceso: 12 02 2019].

https://doi.org/10.36561/ING.17.6

ISSN 2301-1092 • ISSN (en línea) 2301-1106

[11] J. R. Deller, J. H. L. Hansen y J. G. Proakis, *Discrete-Time Processing of Speech Signals*. New York: Macmillan, 1993.

[12] J. L. Cheang Loong, K. S. Subari, M. Kamil Abdullah, N. N. Ahmad y R. Besar, «Comparison of MFCC and Cepstral Coefficients as a Feature Set for PCG Biometric Systems,» 2010.

[13] S. S. Stevens, Volkmann y E. B. John & Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *Acoustical Society of America,* vol. 8, nº 3, p. 6, 1937.

[14] V. G. Vílchez García, "Estimación y clasificación de daños en materiales utilizando modelos AR," Sep. 2010. [En línea]. Disponible en: http://ceres.ugr.es/~alumnos/esclas/. [Último acceso: 12 05 18].

[15] W. Gevaert, G. Tsenov y V. Mladenov, "Neural Networks used for Speech Recognition," *Journal of Automatic Control, University of Belgrade*, vol. 20 pp. 1-7, 2010.

[16] VoIP Supply, «Cisco hd voice,» VoIP Supply, [En línea]. Disponible en: https://www.voipsupply.com/cisco-hd-voice. [Último acceso: 22 03 2019].

[17] K. K. Paliwal, J. G. Lyons y K. K. Wojcicki, "Preference for 20-40 ms window duration in speech analysis," en 2010 4th International Conference on signal Processing and Communication Systems (ICSPCS).

[18] C. S. Aguilar Orozco y M. W. Marín Benítez, *Sistema certificador de locutor por voz*. México: Instituto Politécnico Nacional, 2003.

https://doi.org/10.36561/ING.17.6

ISSN 2301-1092 • ISSN (en línea) 2301-1106