COMPARACIÓN DE TÉCNICAS DE PARAMETRIZACIÓN ESPECTRAL PARA RECONOCIMIENTO DE VOZ EN IDIOMA ESPAÑOL

Ing. Manuel Alejandro Soto Murillo[1], Dr. José Ismael de la Rosa Vargas[2],
Dr. Arturo Moreno Báez[3].

***Resumen*—En este artículo, se presenta una comparación de las técnicas clásicas de parametrización; Codificación Predictiva Lineal (LPC) y Coeficientes Cepstrales de Frecuencias-Mel (MFCC), implementadas en la etapa de extracción de características en los Sistemas de Reconocimiento Automático de Voz (SRAV) para obtener los coeficientes que mejor caractericen la señal de voz. Las señales de voz se muestrearon a 8 y 16kHz y se varió el número de coeficientes característicos (8-12 para 8kHz y 16-24 para 16kHz) para encontrar la configuración que brinde la mayor tasa de reconocimiento y el menor consumo de recursos (tiempo y cálculo). En la etapa de modelado se usó la técnica Modelos Ocultos de Markov (HMM). La técnica de parametrización MFCC presentó una tasa de reconocimiento superior que la técnica LPC bajo las mismas condiciones, obteniendo tasas de reconocimiento de hasta 99.66%.**

***Palabras clave*—Reconocimiento de voz, parametrización, LPC, MFCC, HMM.**

Introducción

La comunicación oral es una de las capacidades básicas y más esenciales que poseen los seres humanos, al igual que su sistema auditivo. El desarrollo científico y tecnológico ha permitido que el ser humano se comunique e interactúe de manera muy simple entre sí. Sin embargo, aún se tiene la necesidad de interactuar con las máquinas y dispositivos electrónicos sin tener que utilizar las manos o pies. Por lo que se han implementado sistemas de reconocimiento de voz a dichas máquinas y dispositivos para poder controlarlos mediante comandos de voz y que la interacción humano-máquina sea lo más similar a la comunicación oral entre humanos.

La mayoría de los sistemas de reconocimiento de voz se encuentran en ambientes inmersos de ruido y reverberaciones. Este ruido impide identificar con claridad las palabras del ruido, es decir, el inicio y el final de cada palabra y afecta directamente en el modelado de la señal de voz y a la tasa de reconocimiento de voz.

Con el fin de mejorar la tasa de reconocimiento de voz, la investigación se enfocó en la etapa de extracción de características o parametrización. Las técnicas de parametrización que se empelaron fueron: la Codificación Predictiva Lineal (LPC) y los Coeficientes Cepstarales de Frecuencias Mel (MFCC) con diferente número de parámetros y frecuencia de muestreo. Con el fin de determinar cuál de las técnicas y en qué condiciones presenta la mayor tasa de reconocimiento, y así hacer una mejora significativa en los sistemas de reconocimiento de voz.

La Voz

El habla es el acto por medio del cual una persona hace uso de una lengua (español) con la finalidad de poder comunicarse, elaborando previamente un mensaje según las reglas gramaticales lingüísticas determinadas de la lengua en que se está comunicando. Mientras que, la voz es el sonido que se caracteriza por tres elementos:

- **La intensidad:** es equivalente al volumen y son las vibraciones producidas por aire al pasar por la glotis. Entre mayor sea su amplitud mayor será la fuerza de la voz. Se mide en decibeles (dB).
- **El tono:** es la cantidad de vibraciones que posee una onda de sonido, a mayor número de vibraciones más aguda será la voz. Estas vibraciones se producen en la laringe y se miden en Hertz (Hz).
- **El timbre:** Es lo que permite que distingamos entre dos sonidos de igual intensidad y tono. Tiene peculiaridades únicas en cada persona, dependiendo de su morfología.

La voz sólo contiene información lingüística en el rango de frecuencias de 200Hz a 8kHz y varía entre hombres, mujeres y niños. Mientras que el límite de audición del ser humano va de 20Hz a 20kHz aproximadamente.

A. Producción del habla en el cerebro y en el aparato fonador

[1] El Ingeniero en Comunicaciones y Electrónica Manuel Alejandro Soto Murillo es Alumno de la Maestría en Ciencias de la Ingeniería de la Universidad Autónoma de Zacatecas, Zacatecas, México. ing.alex7@yahoo.com.mx
[2] El Dr. José Ismael de la Rosa Vargas es Docente-Investigador de la Unidad Académica de Ingeniería Eléctrica y director del Doctorado en Ciencias de la Ingeniería de la Universidad Autónoma de Zacatecas, Zacatecas, México. ismaelrv@yahoo.com
[3] El Dr. Arturo Moreno Báez es Docente-Investigador de la Unidad Académica de Ingeniería Eléctrica de la Universidad Autónoma de Zacatecas, Zacatecas, México. morenob20@uaz.edu.mx

El proceso de producción del habla inicia en el cerebro, en el área de Wernicke; encargada de la elaboración del pensamiento, la elección de las palabras y la comprensión del lenguaje. Y en el área Broca; encargada de la programación de las conductas verbales y de coordinar los órganos del aparato fonador para la producción del habla.

El aparato fonador está conformado por los *órganos de respiración* donde se almacena y circula el aire; los *órganos de fonación*, donde el aire se convierte en sonido y los *órganos de articulación*, donde el sonido adquiere sus cualidades que caracterizan a cada voz.

B. Clasificación del sonido de la voz y sus características

Cada idioma tiene un conjunto de sonidos de voz denominados *fonemas, su* sonoridad depende de las características lingüísticas del idioma. En español son 29 fonemas aproximadamente. Las dos primeras formantes F_1 y F_2 permiten clasificar los diferentes fonemas. La frecuencia de las formantes depende de la forma y de las dimensiones del tracto vocal. Las señales de voz pueden clasificarse en tres tipos según su sonido, A. Flores (1993):

- **Sonoras:** se generan por la vibración de las cuerdas vocales, se caracterizan por tener alta energía y un contenido frecuencial 300-4000Hz, como las vocales.
- **No sonoras o fricativas:** se caracterizan por tener un comportamiento aleatorio en forma de ruido blanco (contiene todas las frecuencias), como las consonantes F, S y V.
- **Plosivas:** se generan cuando el tracto vocal se cierra en algún punto, lo que causa que el aire se acumule para después salir expulsado repentinamente, como las consonantes B, D, M, P y T.

C. Proceso de comunicación oral

La comunicación oral se lleva a cabo entre dos personas con el fin de transmitir un mensaje. Este proceso involucra un emisor, un receptor y un medio de transmisión. Al producir el emisor la señal de voz, la señal acústico-fonética se propaga como ondas a través del aire y llega al oído del receptor, el cual capta las ondas sonoras y las transforma en información que el cerebro sea capaz de interpretar como el habla o la música.

El reconocimiento automático del habla pretende realizar este mismo proceso. Sin embargo, el receptor es una máquina que tiene que ser capaz de comprender los comandos de voz que se le indican. Para llevar a cabo este procedimiento se debe de tratar a la señal de voz para que la maquina sea capaz de comprender el mensaje.

Reconocimiento de voz

El Reconocimiento Automático del Habla (RAH) es una tecnología que le permite al ser humano comunicarse con una computadora. El RAH considera diferentes factores al analizar la señal de voz: acústica, fonética, fluctuaciones de la voz, el medio ambiente, los medios de captación, el ruido ambiental, etcétera. El objetivo del RAH es extraer el contenido lingüístico de una locución, siendo la variabilidad del habla y el gasto computacional los principales problemas.

Los sistemas de reconocimiento automático de voz (SRAV) presentan diferentes estructuras y no todos tienen las mismas etapas. Sin embargo, hay etapas que son fundamentales para que sean considerados como sistemas de reconocimiento de voz: adquisición, pre-procesamiento, extracción de características y reconocimiento de patrones:

A. Adquisición de voz

Es la primera etapa, en la que las ondas acústicas de la señal de voz son obtenidas mediante un micrófono. Se deben predefinir las características de la señal de entrada; como el canal (mono- o multi-canal), el formato de codificación de la muestra, la frecuencia de muestreo y el formato del archivo. En esta etapa se define el número de palabras que conformarán el corpus de voz.

B. Pre-procesamiento

Es la segunda etapa, donde se pre-procesan las señales de voz que conforman el corpus de voz del sistema para reducir el gasto computacional. Esta etapa incluye típicamente estas subetapas (J. Xu et al. 2005):

- **Filtro pasa bajas;** elimina las frecuencias altas que contienen particularmente ruido ambiental.
- **Filtro pasa altas;** elimina las interferencias de baja frecuencia comúnmente introducidas por el micrófono.
- **Detección de la señal de voz:** se identifica el inicio y fin de cada palabra para sólo analizar esa información. Para ello es necesario calcular la energía promedio de la señal de voz filtrada y la energía de corto plazo, además se debe definir un umbral de energía. La energía promedio está dada por:

$$E_{promedio} = \frac{1}{N}\sum_{n=0}^{N-1} s(n)^2,$$

donde $s(n)$ es la señal de voz y N el número total de muestras. La energía de corto plazo está dada por:

$$E_{trama} = \frac{1}{T}\sum_{n=0}^{T-1} s(n)^2\, t(m-n),$$

donde $t(m-n)$ es la trama actual a la que se está calculando la energía y T el tamaño de la trama.

- **Pre-énfasis;** es un filtro digital pasa altas de primer orden que hace que el espectro de voz tenga un rango similar en todas las frecuencias. El filtro está dado por:

$$S_{pp}[n] = S_{bp}[n] - \alpha S_{bp}[n-1],$$

donde $S_{pp}[n]$ denota la señal de salida actual del filtro, $S_{bp}[n]$ es la señal de entrada actual, $S_{bp}[n-1]$ es la señal de entrada previa y α es la constante de suavizado que toma un valor entre 0.9 y 1.

- **Segmentación y ventaneo;** las señales de voz son estocásticas, cuyo contenido frecuencial y nivel de energía varían en largos periodos de tiempo, esto impide su análisis. Para poder analizar la señal de voz es necesario que sea estacionaria, por lo que se debe segmentar en "tramas" de 10-30mseg para considerar a la señal de voz como una señal cuasi-estacionaria. Cada trama se debe traslapar con la ventana adyacente para generar transiciones suaves entre tramas. Posteriormente se debe aplicar el ventaneo, para eliminar los problemas causados por los cambios rápidos de la señal de voz en los extremos de cada trama.

C. Extracción de características

Es la tercera etapa, donde la señal de voz es transformada en una serie de parámetros (coeficientes). El problema fundamental de la parametrización es la elección de un modelo adecuado que estimé la envolvente espectral de la señal de voz. Esta envolvente debe ser representada con un número reducido de parámetros y su cálculo debe exigir el mínimo gasto computacional posible. Las técnicas de parametrización empleadas fueron LPC y MFCC.

LPC se basa en que cada muestra de voz puede predecirse o representarse mediante una combinación lineal de varias muestras pasadas, es decir, que cada muestra de voz $s(n)$ en un tiempo n, puede ser aproximada como una combinación lineal de las muestras de voz anteriores:

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p),$$

donde p es el orden de predicción y $a_1, a_2, \cdots, a_p$ son los coeficientes de predicción que se deben calcular. El orden de predicción se elige a partir de la frecuencia de muestreo y la longitud del tracto vocal. El esquema básico para el cálculo de LPC se representa en el siguiente diagrama según (S. Feraru y M. Zbancioc 2013):

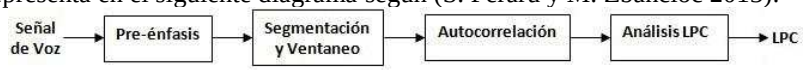


Figura 2. Diagrama de bloques para el cálculo de LPC.

Los dos primeros bloques forman parte del pre-procesamiento de la señal de voz, de los cuales se obtiene una matriz que contiene cada una de las tramas traslapadas y ventaneadas. Cada una de estas tramas se autocorrelaciona para analizar la periodicidad de las muestras que la conforman. En el último bloque, se convierte a cada una de las tramas autocorrelacionadas en un conjunto de parámetros LPC mediante el método auto regresivo Levinson-Durbin. Los coeficientes LPC pueden representar de forma eficiente la información de la envolvente espectral de corto tiempo de las señales de voz según L. Wang et al. (2015).

MFCC es la técnica de extracción de características más empleada en el reconocimiento del habla. Se basa en la percepción del sistema auditivo humano, en la variación de los anchos de banda de las frecuencias críticas del oído humano. El esquema básico para la extracción de los MFCC según G. Martínez y G. Aguilar (2013) es:

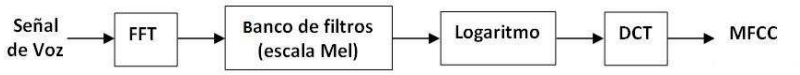


Figura 3. Diagrama de bloques para el cálculo de MFCC.

En el primer bloque se calcula la transformada rápida de Fourier (FFT) a cada una de las tramas de la señal de voz previamente pre-procesadas y se obtiene la magnitud y la densidad espectral de potencia (DEP) de cada trama. Esta transformación se hace para identificar qué frecuencias contiene cada trama. Las frecuencias de la DSP se deben agrupar en regiones y sumar para saber cuánta energía existe en esas regiones. Esto se realiza mediante el banco de filtros Mel, compuesto de filtros triangulares que están distribuidos en escala Mel.

El banco de filtros se calcula con las siguientes ecuaciones:

$$B(m,k) = \begin{cases} 0 & si \quad k > f(m-1), \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & si \quad f(m-1) \le k \le f(m), \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & si \quad f(m) \le k \le f(m+1), \\ 0 & si \quad k > f(m+1), \end{cases}$$

donde $B(m,k)$ es la matriz del banco de filtros, m el número de filtros del banco y $k$el número de ventanas de análisis. Se obtiene un banco de filtros por cada ventana de la señal de voz. El primer filtro que se obtiene es muy estrecho e indica cuánta energía existe cerca de 0 Hz. A medida que las frecuencias aumentan, los filtros se amplían y las variaciones son menores.

Para conocer la energía de los bancos de filtros se debe multiplicar cada banco de filtros con las ventanas de densidad espectral de potencia y posteriormente sumar los coeficientes:

$$E(m,k) = \sum_{m=1}^{M} B(m,k)P(k) \qquad k = 1,2,\cdots,K,$$

donde $P(k)$ es la densidad espectral de potencia y $k$representa el número de la ventana.

Posteriormente se calcula el logaritmo de las energías de los bancos de filtros. Esta operación hace que las características obtenidas coincidan más estrechamente con lo que los humanos realmente escuchan.

$$E_{log}(m,k) = \sum_{m=1}^{M} B(m,k)P(k) \qquad k = 1,2,\cdots,K,$$

En el último bloque se calcula la transformada de coseno discreto del logaritmo de las energías de los bancos de filtros para obtener los MFCC (M. Bezoui et al. 2016). Se hace uso de la transformada de coseno discreto (DCT) para disminuir el gasto computacional (D. Albiñana 2014). La DCT se define por:

$$MFCC(n) = \sum_{m=1}^{M} E_{log}(m,k)\ cos\left[n\left(m-\frac{1}{2}\right)\frac{\pi}{M}\right] \qquad m = 1,2,\cdots,M,$$

Donde M representa el número total de coeficientes MFCC que varía con respecto a n, m representa el número de filtros del banco y k el número de la ventana de análisis.

D. Reconocimiento de patrones

Es la última etapa de un SRAV, donde se hace el entrenamiento de los vectores característicos de las palabras del corpus de voz y la prueba o reconocimiento de la señal de voz de entrada. La técnica de reconocimiento de voz que se aplicó fue Modelos Ocultos de Markov (HMM). Esta técnica se basa en que la señal de voz se puede caracterizar como un proceso estocástico paramétrico y que los parámetros del proceso pueden ser estimados de manera precisa y definida según Rabiner (1993).

Los HMM constan de dos procesos estocásticos anidados, donde cada vector característico (observación) es a su vez una función estocástica de cada estado del modelo. La función estocástica subyacente está oculta y sólo se puede observarse a través de otro conjunto de procesos estocásticos que producen la secuencia de observación O_t en un instante t. Un modelo HMM tiene estados finitos y cada estado tiene asociada una Función de Densidad de Probabilidad (FDP) para cada vector característico.

Diseño del sistema de Reconocimiento de Voz

A. Adquisición.

Se seleccionaron a diez locutores (cinco hombres y cinco mujeres) con edades entre 23 y 61 años de forma aleatoria. Las palabras grabadas fueron; los dígitos del 0-9, encender, apagara, subir, bajar, volumen, enviar, mensaje, llamar, colgar y buscar. Los locutores grabaron diez repeticiones de cada palabra con un solo canal de entrada, a una frecuencia de muestreo de 44.1 kHZ (calidad CD) y extensión ".wav". Las señales de voz de entrada se adquirieron con la herramienta computacional WaveSurfer 1.8.8p4. Con este mismo software se sub-muestrearon las grabaciones a 16kHZ y a 8kHZ para su posterior procesamiento.

B. Pre-procesamiento.

Se diseñaron dos filtros pasa bajas tipo Butterworth IIR, uno para las señales de voz muestreadas a 16khz con una frecuencia de corte de 6kHz y otro para las señales muestreadas a 8khz con una frecuencia de corte de 3kHz. También se diseñaron dos filtros pasa altas tipo Butterworth IIR, uno para las señales muestreadas a 16kHz y otro para las muestreadas a 8kHz ambos con frecuencia de corte de 200Hz. Cada una de las repeticiones se filtró primeramente con el pasa bajas y después con el pasa altas según su frecuencia de muestreo.

Para la detección de actividad de voz se realizó el cálculo de la energía promedio de la señal de voz filtrada, se definió el umbral de decisión igual 0.01 y se calculó la energía de corto plazo en tramas de 25 mseg. Las tramas cuya energía superó el producto de la energía promedio y el umbral de decisión fueron se consideraron segmentos de voz. A partir de la distancia entre cada una de las tramas subsecuentes que contienen información lingüística (tramas vocalizadas), se definió el inicio y el fin de la palabra.

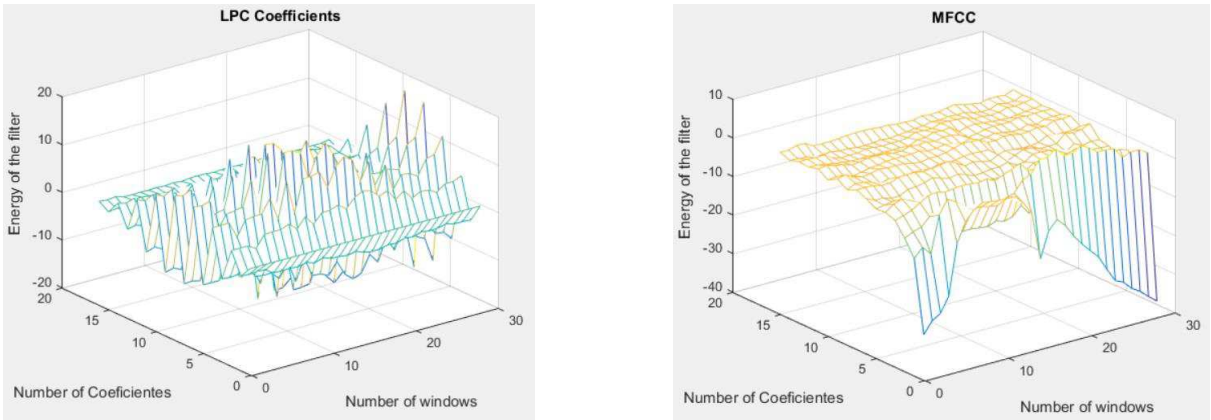
En el pre-énfasis se definió el factor de suavizado $\alpha = 0.96875$ por los resultados que obtuvo (J. Xu et al. 2005).

Las señales de voz muestreada a 8kHz se segmentaron en tramas de 256 muestras y las señales voz muestreadas a 16kHz se segmentaron en tramas de 512 muestras. Las tramas de ambas frecuencias se traslaparon al 50% y fueron ventaneadas con una ventana Hamming. Todas las ventanas de cada una de las repeticiones se almacenaron en una matriz (una ventana por fila y una matriz por repetición).

C. Extracción de características.

Para obtener los coeficientes LPC de las señales de voz muestreadas a $8kHz$ se varió el orden de predicción entre 8-12 coeficientes por trama, ya que según (N. Wankhede y M. Shah 2013) son los coeficientes que modelan correctamente la longitud del tracto vocal. Para las señales de voz con $Fm = 6kHz$ se varió el orden de predicción entre 16-24 coeficientes por ventana. Se seleccionar el doble de coeficientes con el fin de que se tuviera la misma distribución en las bandas de 0 a 4kHz, para que fuera una comparación objetiva de la existencia de una mejora al duplicar la frecuencia de muestreo. Los coeficientes LPC obtenidos se ven gráficamente como en la figura 4a.

Para obtener los coeficientes MFCC de las señales de voz se varió de la misma manera el orden de predicción para ambas frecuencias de muestreo. Los coeficientes MFCC obtenidos se ven gráficamente como en la figura 4b.



a) Matriz de coeficientes LPC. b) Matriz de coeficientes MFCC.

Figura 4. Matrices de coeficientes LPC y MFCC de la primera repetición de la palabra "cero" del locutor "Lulú" con $Fm = 16kHz$.

D. Reconocimiento de patrones

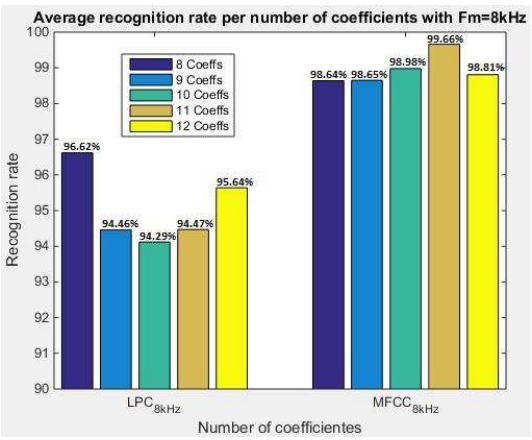
La etapa de reconocimiento de voz se realizó mediante los HMM desarrollados por (Kevin Murphey 2017).

A partir de las matrices de coeficientes LPC y MFCC, se crearon los modelos de entrenamiento mediante el algoritmo de estimación máxima Baum-Welch. Se creó un modelo por cada una de las 20 palabras que conforman el corpus de voz. En la prueba se hizo una comparación de la palabra que se deseaba reconocer con respecto a los modelos creados en la etapa de entrenamiento, mediante el algoritmo de Viterbi. Se entrenó el 70\% de las repeticiones de cada palabra y se probó el 30\% de las repeticiones restantes de las mismas palabras.

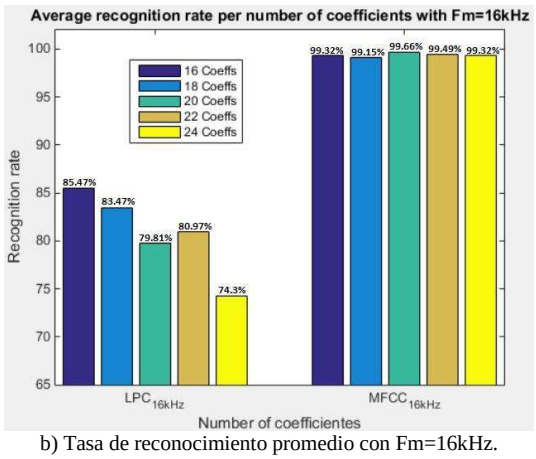
Los modelos fueron de 5 estados con una secuencia de transición left-right y 3 mezclas gaussianas por estado. El número de observaciones del modelo se fue variando según la cantidad de coeficientes que caracterizaron a la palabra, usando el mismo número de observaciones que coeficientes. El modelo que arrojó la mayor probabilidad logarítmica fue la palabra reconocida y después se definió si la palabra fue reconocida correcta o incorrectamente.

Resultados

Con las técnicas de parametrización LPC y MFCC se obtuvieron altas tasas de reconocimiento de voz, algunas de las palabras se reconocieron al 100% con ambas frecuencias de muestreo y diferente número de coeficientes. En la figura 5 se muestran las tasas de reconocimiento promedio por número de coeficientes con $Fm = 8kHz$ y con $Fm = 16kHz$ para LPC y MFCC. En la gráfica 5a los coeficientes se variaron de 8-12 y en la gráfica 5b se variaron los coeficientes de dos en dos de 16-24. Se puede observar en ambas gráficas que la técnica MFCC presentó tasas de reconocimiento superiores a las de LPC en las diez diferentes configuraciones de coeficientes. Se puede observar que las tasas de reconocimiento con la técnica LPC disminuyeron al aumentar la frecuencia de muestreo y el número de coeficientes. La técnica de extracción de características MFCC presentó tasas de reconocimiento mayores a las de LPC para ambas frecuencias de muestreo y bajo las condiciones de prueba desarrolladas en la investigación; ambiente de grabación, la frecuencia de muestreo y el número de coeficientes con los que se caracterizó cada palabra.



a) Tasa de reconocimiento promedio con Fm=8kHz.



b) Tasa de reconocimiento promedio con Fm=16kHz.

Figura 5. Tasas de reconocimiento por número de coeficientes LPC y MFCC.

Comentarios Finales

Se puede concluir que en un sistema de reconocimiento de voz cada una de las etapas que lo conforman es indispensable y omitir alguna de ellas afecta directamente la tasa de reconocimiento. Se concluye que la técnica MFCC a mayor frecuencia de muestreo y número de coeficientes, presenta mayor tasa de reconocimiento de voz. Caso contrario a la técnica LPC, en la que, a mayor frecuencia de muestreo y número de coeficientes, menor es la tasa de reconocimiento. El tiempo procesamiento de voz con la técnica MFCC es más rápido con respecto a LPC. A mayor frecuencia de muestreo y número de coeficientes, el tiempo de procesamiento es menor.

Como trabajo futuro, se puede aumentar el número de locutores, el vocabulario y el corpus de voz. También se puede implementar un sistema de reconocimiento de voz con estas técnicas de extracción de características LPC y MFCC con frecuencias de muestreo de 8y16kHz y sus respectivas configuraciones de coeficientes, en alguna tarjeta digital y comparar si se obtienen los mismos resultados que los obtenidos en las simulaciones de esta investigación.

Referencias

Albiñana, D., *Implementación de algoritmos para la extracción de patrones característicos en Sistemas de Reconocimiento de Voz en Matlab*, Universidad Politécnica de Valencia, 2014.

Bezoui, M., A. Elmoutaouakkil and A.Beni-hssane, *Feature Extraction of some Quranic Recitation using Mel-Frequency Cepstral Coeficients (MFCC)*, 5th International Conference on Multimedia Computing and Systems (ICMCS), Marrakech, Morocco, sept. 29 - oct.1, 2016.

Boussaa, M., I. Atouf, M. Atibi and A. Bennis, *Comparison of MFCC and DWT features extractors applied to PCG classification*, 11th International Conference on Intelligent Systems: Theories and Applications (SITA), Mohammedia, Morocco, oct. 19-20, 2016.

Duque, C. y M. Morales, *Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones*, Universidad Tecnol_ogica de Pereira, Abril 2007.

Flores, A., *Reconocimiento de Palabras Aisladas en Castellano*, Inictel, Direcci_on de Investigaci_on y Desarrollo, 1993.

Furui, S., *50 years of progress in speech and speaker recognition*, Dept. of Computer Science, Tokyo Institute of Technology, pp 1-9, 2004.

Gruhn, R. et al., *Statistical Pronunciation Modeling for Non-Native Speech Processing, Signals and Communication Technology*, Springer-Verlag, Berlin Heidelberg, 2011.

Martínez, G. y G. Aguilar, *Reconocimiento de voz basado en MFCC, SBC y Espectrogramas*, Ingenius. N.10, ISSN: 1390-650X, pp. 12-20, 2013.

Murphey, K., HMM toolkit, Massachusetts EE.UU., noviembre 2017, disponible en: https://www.cs.ubc.ca/ murphyk/Software/HMM.zip

Rabiner, L., *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

Wang, F. y W. Xu, *A Comparison of Algorithms for the Calculation of LPC Coeficients*, Communication University of China, Beijing, Information Science, Electronics and Electrical Engineering (ISEEE), International Conference on vol.3, pp. 26-28, 2014.

Wang, L., Z. Chen and F. Yin, *A Novel Hierarchical Decomposition Vector Quantization Method for High-Order LPC Parameters*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, VOL. 23, No. 1, 2015.
Xu, J., A. Ariyaeeinia, R. Sotudeh and Z. Ahmad, *Pre-processing Speech Signal in FPGAs*, University of Herrtsfordshire, UK, IEEE Xplore, 2005.