Ingenius. Revista de Ciencia y Tecnología
ISSN: 1390-650X
revistaingenius@ups.edu.ec
Universidad Politécnica Salesiana
Ecuador

Martínez Mascorro, Guillermo Arturo; Aguilar Torres, Gualberto
Reconocimiento de voz basado en MFCC, SBC y Espectrogramas
Ingenius. Revista de Ciencia y Tecnología, núm. 10, julio-diciembre, 2013, pp. 12-20
Universidad Politécnica Salesiana
Cuenca, Ecuador

Disponible en: http://www.redalyc.org/articulo.oa?id=505554816003

- Cómo citar el artículo
- Número completo
- Más información del artículo
- Página de la revista en redalyc.org

redalyc.org
Sistema de Información Científica
Red de Revistas Científicas de América Latina, el Caribe, España y Portugal
Proyecto académico sin fines de lucro, desarrollado bajo la iniciativa de acceso abierto

UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR
INGENIUS
Revista de Ciencia y Tecnología

Reconocimiento de voz basado en MFCC, SBC y Espectrogramas

Guillermo Arturo Martínez Mascorro[1,*] y Gualberto Aguilar Torres[2]

Resumen

Uno de los problemas en los sistemas de reconocimiento automático de hablante son los cambios en la voz. Comúnmente, una persona puede tener cambios voluntarios e involuntarios (también naturales y artificiales) que provocan confusiones en el sistema, los cambios en la voz también pueden ser naturales y artificiales. En el artículo presente se propone un sistema de reconocimiento a través de una identificación en paralelo, usando tres algoritmos: MFCC, SBC y el espectrograma. Empleando una máquina de soporte vectorial como clasificador, cada algoritmo arroja un grupo de personas con las probabilidades más altas y después de una evaluación, se toma una decisión. El objetivo de este artículo es tomar ventaja de los tres algoritmos.

Palabras clave: Reconocimiento del hablante con cambios en la voz, coeficientes cepstrales en la frecuencia de Mel, parámetros cepstrales basados en sub-banda, espectrograma, máquina de soporte vectorial.

Abstract

One of the problems of the Automatic Speech Recognition systems is the voice's changes. Typically, a person can have voluntary and involuntary voice's changes and the system can get confused in these cases, also the changes could be natural and artificial. This paper proposes and recognition system with a parallel identification, using three different algorithms: MFCC, SBC and Spectrogram. Using a Support Vector Machine as a classifier, every algorithm gives a group of persons with the highest likelihood and, after an evaluation, the result is obtained. The aim of this paper is to take advantage of the three algorithms.

Keywords: Speech recognition with voice changes, Mel Frequency Cepstral Coefficients, Subband-Based Cepstral Parameters, Spectrogram, Support Vector Machine.

[1,*]Ingeniero en Electrónica, Estudiante de la Maestría en Ciencias de Ingeniería en Microelectrónica, Instituto Politécnico Nacional, México DF, México. Autor para correspondencia ✉: gmartinezma1103@alumno.ipn.mx
[2]Doctor en Ciencias en Comunicaciones y Electrónica, Maestro en Ciencias de Ingeniería en Microelectrónica, Ingeniero en Comunicaciones y Electrónica, Docente del Instituto Politécnico Nacional en la Sección de Estudios de Posgrado e Investigación de la ESIME Culhuacán, México DF, México.

Recibido: 08-11-2013, Aprobado tras revisión: 18-11-2013.

Forma sugerida de citación: Martínez, G. y Aguilar, G. (2013). "Reconocimiento de voz basado en MFCC, SBC y Espectrogramas". Ingenius. N.°10, (Julio-Diciembre). pp. 12-20. ISSN: 1390-650X.

1. Introducción

Uno de los problemas dentro del reconocimiento de hablantes es la modificación de tono en la voz. En algunos casos, la persona a reconocer no puede controlar algunos cambios en su propia voz, por ejemplo, cuando una persona está enferma. Existen dos tipos de cambios en la voz: voluntarios e involuntarios. El primero ocurre cuando una persona, de manera consciente, hace alteraciones a su propia voz para no ser reconocido, e.g. hablar más grave, cubrir la boca o tapar su nariz. Los cambios involuntarios se dan cuando una persona no puede controlarlos dichos cambios, por ejemplo, cuando tiene un resfriado, tos o está ronco. Pueden haber algunas razones más para cambios involuntarios, sin embargo, estas son las más comunes.

Puede haber otra clasificación para estos cambios, los naturales y los artificiales. Los cambios artificiales en la voz ocurren cuando, además del hecho de querer cambiar la voz, alguien utiliza un dispositivo para realizar este cambio, como un procesador de voz.

En [1] se muestra una comparativa de distintos métodos de parametrización de voz, teniendo como base de comparación los Coeficientes Cepstrales en la Frecuencia de Mel, siendo estos los más usados debido a su bajo costo computacional y a su robustez. Con base en los resultados, se decidió analizar estas técnicas dentro del reconocimiento del habla robusto a cambios de tono en la voz, utilizando técnicas con distintas transformaciones, como la Transformada Discreta de Fourier (DFT) y la Transformada Discreta de Paquetes Wavelet (DWPT).

2. Desarrollo

2.1. Base de datos

La base de datos cuenta con grabaciones de 19 personas, cada una pronunciando 17 oraciones en 4 tonos distintos, de estas oraciones se obtiene un total de 168 palabras de las cuales se extrajeron sus características y se realizaron pruebas. Para la creación de esta base de datos, se utilizó un micrófono externo de clip, las oraciones fueron grabadas directamente en MATLAB® a una frecuencia de 8 kHz, la dinámica consiste en decir las oraciones establecidas en distintos tonos, de modo que se pueda tener más información de la voz de la persona aún con cambios en la misma.

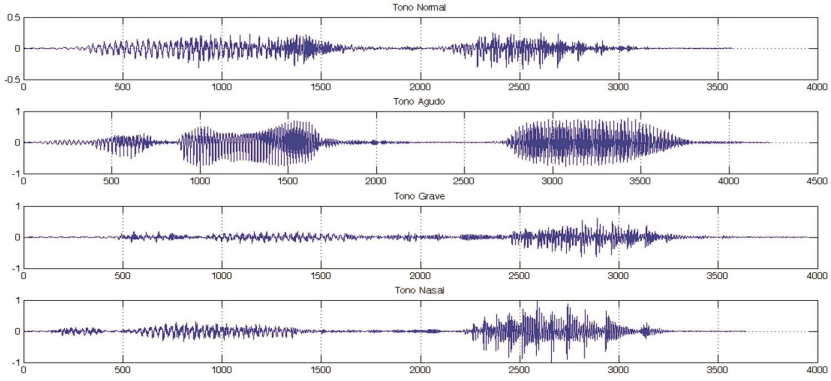

Figura 1. Comparación de la palabra "fijo" en cuatro tonos.

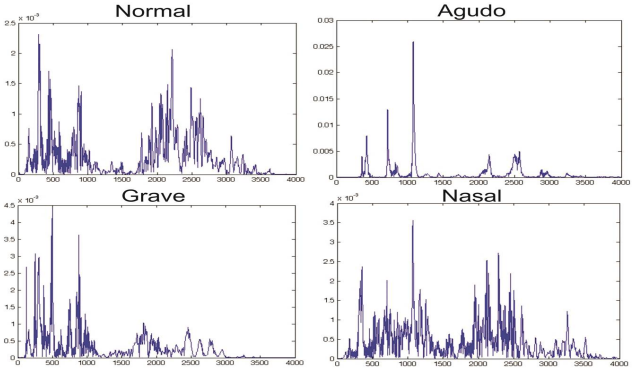

Figura 2. Componentes del espectro de frecuencias de la palabra "fijo" en cuatro tonos.

Los tonos utilizados son normal, agudo, grave y nasal, este último se simula tapando la nariz del hablante. En la Figura 1, se muestra la amplitud de la palabra "fijo" grabada en los cuatro tonos mencionados, mientras que en la Figura 2, se observa el espectro de frecuencias de cada una de las grabaciones.

2.2. Algoritmos previos

Previo a la selección de los tres métodos de extracción de características que serían utilizados para el sistema propuesto, se realizó una comparativa con las técnicas siguientes:

2.2.1. MFCC

Los Coeficientes Cepstrales en la Escala de Mel (MFCC) representan la amplitud del espectro del habla de manera compacta, esto los ha vuelto la técnica de extracción de características más usada en reconocimiento del habla [2]. En la Figura 3, se muestra el proceso para la elaboración de un vector característico de MFCC. Primeramente, se aplica un filtro de pre-énfasis a la señal y posteriormente se divide la misma en tramas y se le aplica una función de ventaneo, en este caso una ventana de Hamming de 20 ms. El ventaneo sirve para eliminar los bordes de la señal y darle una acentuación a la parte central de la trama para su análisis.

Al obtener la Transformada Discreta de Fourier (DFT) de cada trama se utiliza la amplitud del espectro, y esta información es pasada al dominio de Mel mediante el Banco de Filtros. La escala Mel se basa en mapear entre la frecuencia actual al *pitch* que percibe, un escucha humano simulado, esta escala es lineal por debajo de 1 kHz y logarítmica por encima de este umbral. Después se obtiene el logaritmo de la señal y finalmente se aplica la Transformada de Coseno Discreta (DCT), de este vector obtenido se toman la cantidad de coeficientes deseados por trama.

2.2.2. SBC

Los parámetros Cepstrales Basados en Sub-banda (SBC) son derivados de tomar la DCT de la energía del logaritmo de la sub-banda. En este aspecto son similares a los MFCC, pero en el método subyacente para descomponer la señal en sub-bandas, es diferente [3].

El análisis de sub-bandas utiliza una Transformada Wavelet bi-ortogonal para realizar una descomposición jerárquica de tiempo-frecuencia de la señal de voz. El modo más simple para realizar la descomposición es en sus componentes pasa-baja y pasa-alta. La ventaja de este esquema es que la señal original puede ser reconstruida perfectamente de los bloques obtenidos, al realizar la operación inversa.

En la Figura 4, se muestra el árbol de descomposición que utiliza el paquete de *wavelets* que se ocupó para las pruebas realizadas. Este esquema de descomposición fue diseñado por Sarikaya & Hansen y utiliza un rango de frecuencias entre 0 y 4 kHz. El análisis propuesto enfatiza las frecuencias bajas y medias, asignando más sub-bandas en estas partes. Una vez que se realizó la descomposición se calcula la energía en cada sub-banda y es escalado por un número de coeficientes establecidos para cada banda. La energía se calcula mediante la ecuación 1.

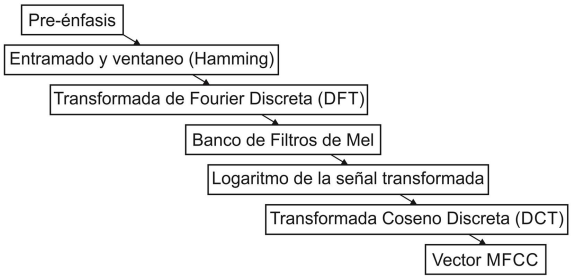

Figura 3. Proceso de obtención de los coeficientes MFCC.

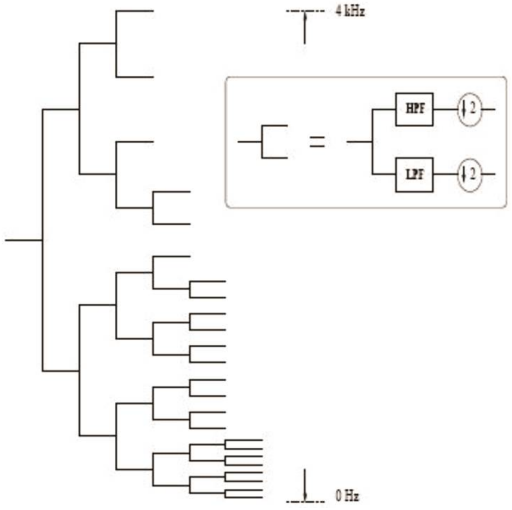

Figura 4. Proceso de obtención de los coeficientes MFCC.

$$E_i = \frac{\sum_{m \in i} \left[\left(W_\Psi x \right) (i), m \right]^2}{N_i} \tag{1}$$

Donde $W_\Psi x$ es la transformada de paquete *wavelet* de la señal x, i es el índice de frecuencia de sub-banda y N_i es el número de coeficientes en la sub-banda i.

La transformada de paquete *wavelet* fue implementada usando el filtro *wavelet* de Daubechie de orden 32. Las características de habla resultantes son nombradas Parámetros Cepstrales Basados en Sub-banda (SBC), son derivados de la aplicación de la transformada DCT en la energía de la sub-banda:

$$SBC(j) = \sum_{i=1}^{M} \log\, E_i \cos\left(\frac{j\,(i-0.5)}{M}\pi\right)^2 \qquad (2)$$

Para $j = 1, \ldots, J$, donde J es el número de parámetros SBC y M es el número total de bandas de frecuencia.

2.2.3. Espectrograma

El espectrograma consiste en la representación gráfica del espectro de frecuencias de la emisión sonora. El espectrograma puede revelar rasgos, como altas frecuencias o modulaciones de amplitud, que no pueden apreciarse incluso aunque estén dentro de los límites de frecuencia del oído humano.

Normalmente, un espectrograma representa el tiempo sobre el eje horizontal, la frecuencia sobre el eje vertical y la amplitud de las señales mediante una escala de grises o de colores. Con la ayuda de la función "*spectrogram*" de MATLAB® es más fácil obtener estos gráficos de una señal determinada de voz, y da opción a establecer un "mapa de colores" (*colormap*), el cual será utilizado para determinar el color del eje encargado de la amplitud. En la Figura 5, se muestra un ejemplo de espectrograma, el *colormap* es conocido como "*Jet*" y es el que maneja por defecto MATLAB® en su función "*spectrogram*".

Para la comparativa realizada, se utilizó la primera matriz del espectrograma, utilizando un *colormap Jet*, la cual corresponde a la capa de color rojo de la imagen.

En la Figura 6, se muestra la gama de colores que utiliza el *colormap Jet* y la escala de grises en la que queda la primera capa del mismo, mientras que en la Figura 7, se observa de manera aplicada, estos mapas de colores en el cálculo del espectrograma de un segmento vocalizado.

2.2.4. Características de voz

Para la elaboración de estos vectores se revisaron diversas propiedades que se encuentran presentes en las señales de voz y proporcionan información sobre la calidad de las mismas. Estas características se encuentran en el dominio del tiempo:

- *Pitch.*
- Contorno de energía.
- *Jitter.*
- *Shimmer.*
- PMR.

La información de estas características se encuentra en [4].

2.2.5. Filtros de Gabor

Los filtros de Gabor son un ejemplo de filtros *wavelet* ampliamente utilizados en aplicaciones de procesamiento de imágenes, como el análisis de textura, segmentación y clasificación.

Principalmente analiza los componentes de frecuencia espacial de una imagen de manera localizada; para esto se crea una envolvente gaussiana cuya anchura se ajusta a la frecuencia de las sinusoidales complejas. Son comúnmente utilizados por su capacidad de eliminar ruidos y mejorar la cresta y valle de las estructuras.

Las funciones de Gabor quedan determinadas por cuatro parámetros, dos que expresan su localización en el dominio espacial (x, y) y otros dos que expresan la frecuencia espacial de sintonía (F) y orientación (Φ), así que ésta se puede expresar como:

$$h\,(x, y) = g\,(x', y') \exp\,(2\pi j F x') \qquad (3)$$

La señal elemental de Gabor bidimensional espacial, está en función de la respuesta gaussiana bidimensional $g'(x', y')$, la frecuencia espacial (F) y la rotación aplicada (Φ). La respuesta gaussiana bidimensional puede expresarse mediante la siguiente ecuación:

$$g\,(x'y') = \left(\frac{1}{2\pi\lambda\sigma^2}\right) \exp\left[\frac{(x/\lambda)^2 + y^2}{2\sigma^2}\right] \qquad (4)$$

Estas funciones operan en el conjunto de números complejos, cuya parte real es la función de Gabor simétrica y la parte imaginaria es la función de Gabor asimétrica.

$$(x'y') = (x\cos\Phi + y\sin\Phi - x\sin\Phi + y\cos\Phi) \qquad (5)$$

$$h\,(x, y) = h_c\,(x, y) - j h_s\,(x, y) \qquad (6)$$

$$h_c\,(x, y) = g\,(x', y') - \cos\,(2\pi F x') \qquad (7)$$

$$h_s\,(x, y) = g\,(x', y') - \sin\,(2\pi F x') \qquad (8)$$

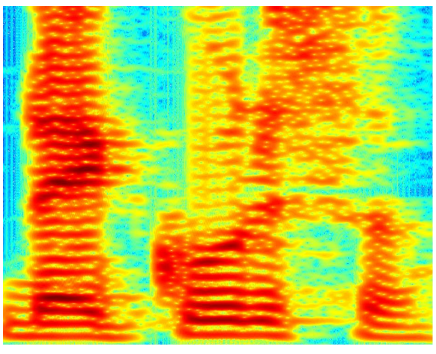

Figura 5. Ejemplo de un espectrograma en escala de colores, con un *colormap Jet*.

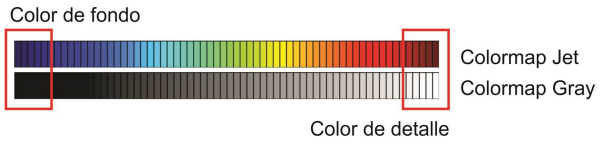

Figura 6. Distintas versiones de *colormaps* utilizados para la comparativa.

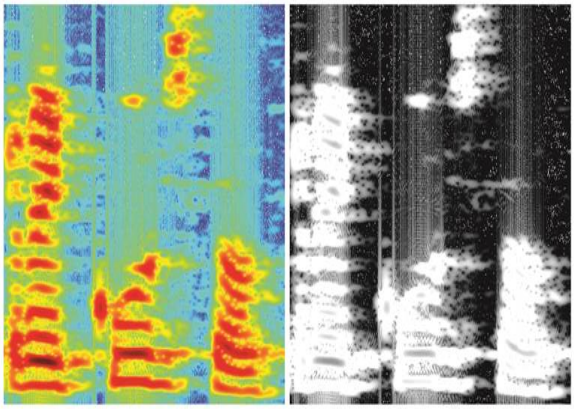

Figura 7. *Colormaps* aplicados al espectrograma de una señal, a la izquierda *colormap Jet*, a la derecha escala de grises de la primera capa.

Donde $h_c(x, y)$ es la señal elemental de Gabor con componentes reales y $h_s(x, y)$ es la señal elemental de Gabor con componentes imaginarios, simetría par e impar respectivamente. La información aportada por este par en cuadratura de fase corresponde al contraste de energía en un punto dado. El contraste de energía $M(x, y)$ de un par en cuadratura se obtiene mediante la ecuación 9.

$$M\left(x,y\right) = \sqrt{h_c^2 + h_s^2} \tag{9}$$

Esta función presenta gran similitud con el comportamiento de las células complejas y proporciona una medida de la respuesta del canal, que es independiente del cambio de fase local. Al promediar cada una de estas amplitudes de la señal resultante, se obtiene los vectores característicos de la respuesta de la imagen:

$$M = \frac{\sum_{p=1}^{B} M_p\left(x,y\right)}{B} \tag{10}$$

Donde B es el número de bancos de filtros de Gabor [5].

2.3. SVM

Una máquina de soporte vectorial es esencialmente un clasificador binario no lineal, capaz de determinar si un vector de entrada "x" pertenece a una clase 1, donde la salida deseada sería $y = 1$, o a una clase 2 donde $y = -1$.

Este algoritmo fue propuesto en 1992 y es una versión no lineal de un algoritmo lineal mucho más antiguo, la regla de decisión del hiperplano óptimo, que fue introducido en los años sesenta [6]. En la ecuación 10 se muestra la fórmula general de la SVM y en [6] se encuentran los kernels más usados.

$$f\left(x\right) = \sum_{i=1}^{N} \propto_i y_i K\left(x_i, x\right) + b \tag{11}$$

3. Comparativa y resultados

Para comparar el funcionamiento de cada una de las técnicas antes mencionadas, se utilizó dos terceras partes de la base de datos, 112 palabras, para entrenar los modelos de la SVM y 56 palabras como prueba de cada persona.

En esta etapa se usó la extracción de características directo de la señal y, con las mismas técnicas se elaboraron imágenes para después procesarlas con los filtros de Gabor y obtener un vector característico.

Se describirán las nueve variantes utilizadas a partir de las técnicas descritas, las primeras cuatro corresponden a extracción directa de la señal, y las últimas cinco corresponden a imágenes generadas con estas técnicas para después filtrarlas con Gabor.

3.1. SBC - 8 primeros segmentos (8PS)

El análisis SBC utiliza una ventana de 24 ms de la cual obtiene 12 coeficientes, debido a la variación en la duración de las palabras, se tomó como medida de prueba los primeros 8 segmentos de cualquier palabra, generando un vector de 96 coeficientes.

3.2. SBC - 10 segmentos iguales (10SI)

La diferencia en esta variante es que en lugar de utilizar la ventana de 24 ms, se toma una de un décimo de la duración de la palabra a caracterizar. Con este cambio siempre se obtiene un vector de 120 coeficientes.

3.3. Carvoz

Se denominó así, ya que utiliza las características de voz antes mencionadas, y es el vector desarrollado en [4] de 408 coeficientes.

3.4. Carvoz + SBC

Utilizando el vector anterior, se le concatenan los coeficientes del vector SBC - 8PS y nos da un vector de 504 coeficientes.

Hasta aquí son las variantes que trabajan directamente con la señal. Las siguientes cinco, como se

mencionó previamente, se caracterizan y con esa información se genera una imagen que después se filtra con Gabor. Todas estas variantes tendrán vectores de 108 coeficientes.

3.5. Matriz MFCC

Se usa el análisis tal cual se explicó previamente, la matriz resultante se imprime en pantalla, como se muestra en la Figura 8. Posteriormente se filtra con Gabor. Como la imagen siempre tendrá el mismo tamaño no importa la diferencia de duraciones de cada palabra.

3.6. Matriz SBC

Utiliza nuevamente su ventana de 24 ms, pero igual que la matriz MFCC, como la imagen generada es del mismo tamaño, no importa la diferencia de duraciones (Figura 9).

3.7. Espectrograma

Se elabora un espectrograma con un *colormap Jet*, al momento de guardarlo en una variable se tiene una matriz de tres dimensiones, correspondiendo cada una a las capas de color del RGB (rojo, verde y azul), solo se utiliza la primera capa de color para obtener el vector característico. La Figura 10, muestra un ejemplo.

Una vez que se entrenaron los modelos de la SVM, 112 vectores de entrenamiento y 56 de prueba, se realizaron las pruebas de reconocimiento con la ayuda de la SVM. Los resultados obtenidos se muestran en la Tabla 1.

4. Sistema propuesto

Con base en los resultados de la Tabla 1, se observa que la matriz MFCC, los SBC-10SI y el espectrograma son las tres técnicas que tienen un mayor porcentaje de reconocimiento.

La propuesta de reconocimiento es que cada uno de las identificaciones, basadas en cada algoritmo de extracción de características, realice una evaluación de rango tres, de manera que se tengan tres posibles autores por cada evaluación y una matriz de probables locutores, la cual será analizada para determinar la decisión final.

Para este sistema se introducen los segmentos vocalizados de una señal de voz, cada uno de estos segmentos es evaluado y da un puntaje a los autores más probables. Al finalizar la evaluación de todos los segmentos vocalizados se tiene una lista de los autores con sus puntajes y de aquí se toma el que tenga un mayor puntaje para ser el más probable autor.

Después de las evaluaciones se crea lo que llamaremos "matriz de probables", esta es una matriz de 3x3 donde cada columna representa un método de extracción utilizado y el número de fila representa el orden de mayor a menor probabilidad de ser el autor del segmento vocalizado.

De acuerdo a la posición en el reconocimiento, se le da una cantidad de puntos a la persona, al más probable se le otorgan tres puntos, al segundo dos y al tercero un punto. En la Figura 11, se muestra un ejemplo de la matriz de probables de un segmento vocalizado y posteriormente se explicará paso a paso como se evaluará este segmento.

Como se observa en la Figura 11, el posible autor se puede repetir en dos o más columnas pero no en una misma. El que un locutor se repita en los reconocimientos también es importante, debido a que coinciden los reconocimientos. Si la persona se repite en los tres reconocimientos recibe un punto, si se repite en dos, recibe medio punto, y si aparece una sola vez en la matriz, no es tomado en cuenta y se le asigna 0 puntos (Figura 12).

Posteriormente el sistema crea una matriz donde la primera columna son los posibles autores, la segunda los puntos por posición, la tercera los puntos por repetición y en la cuarta una puntuación total, la cual se obtiene de multiplicar los puntos de posición por los de repetición (Tabla 2) y finalmente se busca a la persona con el total más alto; en el caso de este ejemplo la persona 1 sería aquella que el sistema reportaría como autor del segmento.

Para poder evaluar una oración se tendrán n segmentos vocalizados y se repetirá este proceso de evaluación en cada uno, los puntos totales se irán acumulando en otra matriz dónde se sumaran de acuerdo a la evaluación de cada segmento vocalizado.

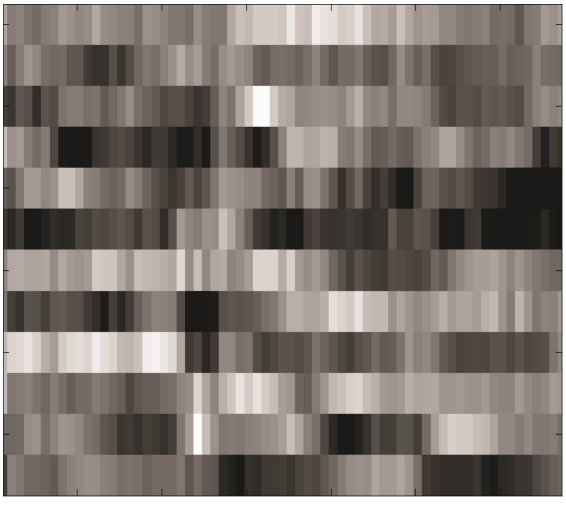

Figura 8. Matriz MFCC impresa en pantalla.

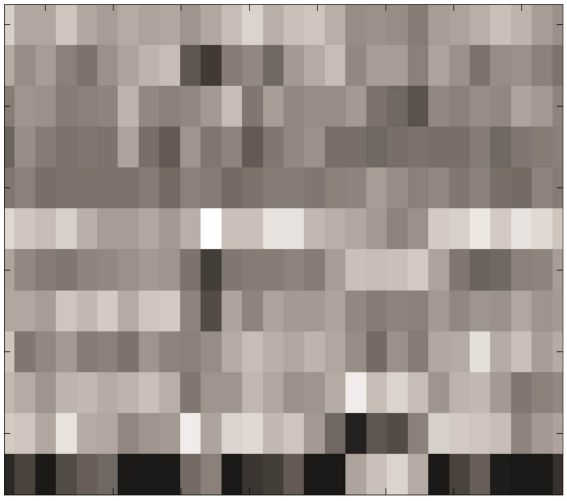

Figura 9. Matriz SBC impresa en pantalla.

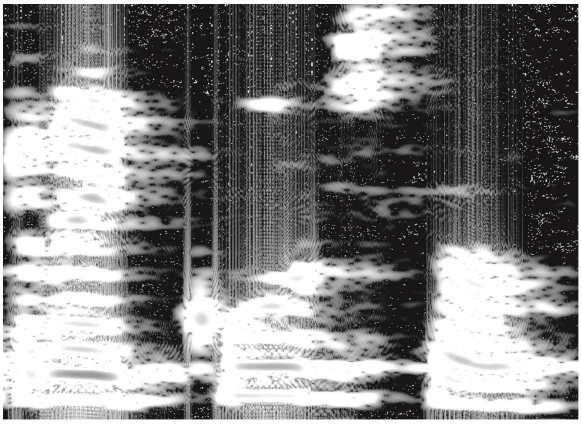

Figura 10. Matriz SBC impresa en pantalla.

Tabla 1. Resultados de pruebas de reconocimiento con las diferentes técnicas utilizadas.

Técnica	Número de aciertos (3192)	Porcentaje de reconocimiento
M. MFCC (5)	3121	97.77%
SBC - 10SI (2)	3118	97.68%
Espectrograma (7)	3086	96.68%
Carvoz (3)	3077	96.40%
Carvoz+SBC (4)	3074	96.30%
M. SBC (6)	3021	94.67%
SBC - 8PS (1)	2983	93.45%

5. Pruebas

Para probar la eficiencia del sistema, tres personas grabaron cinco oraciones completamente distintas a las de entrenamiento, las primeras cuatro manteniendo un tono en cada oración pero distinto el tono en cada grabación, la quinta involucra cambios en la voz a lo largo de la frase.

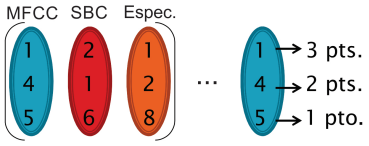

Figura 11. Matriz de probables y modo de puntuación por posición.

Figura 12. Matriz de probables y modo de puntuación por repetición.

Tabla 2. Resultado de la evaluación del ejemplo con el sistema propuesto.

Persona	Pts. posición	Pts. repetición	Total
1	8 (3+2+3)	1	8
2	5 (0+3+2)	0.5	2.5
4	2 (2+0+0)	0	0
5	1 (1+0+0)	0	0
6	1 (0+1+0)	0	0
8	1 (0+0+1)	0	0

Como se explicó previamente, la dinámica es la siguiente: se evalúa cada segmento vocalizado, alrededor de 20 por oración, se suman los puntos de todas las evaluaciones y la persona con más puntos es la que se considera autora de la frase.

Durante los reconocimientos de segmentos se presenciaron algunos casos sobre cómo podía ubicarse la persona correcta en la matriz de probables. En la Figura 13, se muestran una serie de pantallas sobre estos casos, en todos ellos la persona correcta es la persona 1, sin embargo, no en todos aparece de la misma manera.

En la primera pantalla se muestra el caso ideal, los tres sistemas reconocieron a la persona correcta como la más probable, lo que le da una puntuación perfecta. En el segundo caso uno de los reconocimientos la ubica en segundo lugar, bajando su puntuación. El tercer caso es que solo aparece en dos reconocimientos y esto afecta sus puntos por repetición lo cual la pone en el segundo lugar en el resultado.

La cuarta pantalla muestra a la persona correcta una sola vez en la matriz de probables por lo cual no es tomada en cuenta y finalmente en la quinta pantalla ya no aparece en la matriz de probables. Todos estos

casos son probables y por ello se toman en cuenta los puntos al final de todas las evaluaciones.

Una vez evaluados todos los segmentos se genera una pantalla de puntuaciones y da un resultado el sistema (Figura 14).

6. Resultados

Al evaluar todas las oraciones se obtuvo una sola grabación con un resultado no deseado, por lo cual el sistema demostró un 93.33% de acierto en el reconocimiento. En la Tabla 3, se muestran los resultados de la evaluación de cada una de las oraciones de prueba con el sistema propuesto, donde se observa que la oración cuatro de la persona tres es la que presenta el resultado no deseado. En la Figura 15, se muestra la evaluación errada y se observa que, aunque el primer puesto lo tiene la persona 9, la persona 3 tiene un alto puntaje y es la segunda opción en la lista.

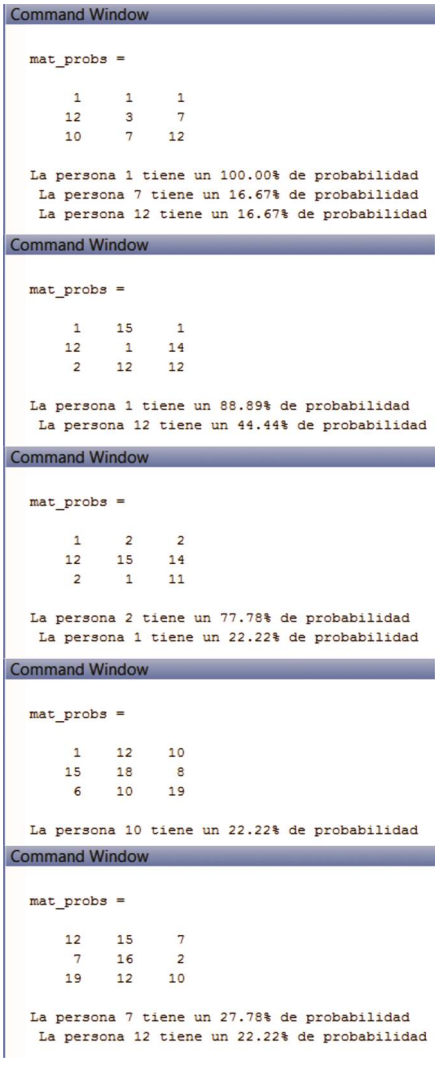

Figura 13. Matriz de probables y modo de puntuación por repetición.

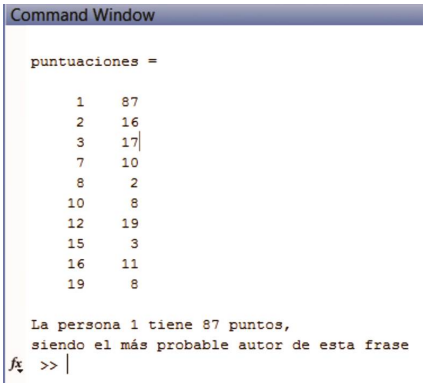

Figura 14. Matriz de probables y modo de puntuación por repetición.

Tabla 3. Resultados de reconocimiento de oraciones de prueba con el sistema propuesto.

	Persona 1	Persona 2	Persona 3
Oración 1	1	2	3
Oración 2	1	2	3
Oración 3	1	2	3
Oración 4	1	2	9
Oración 5	1	2	3

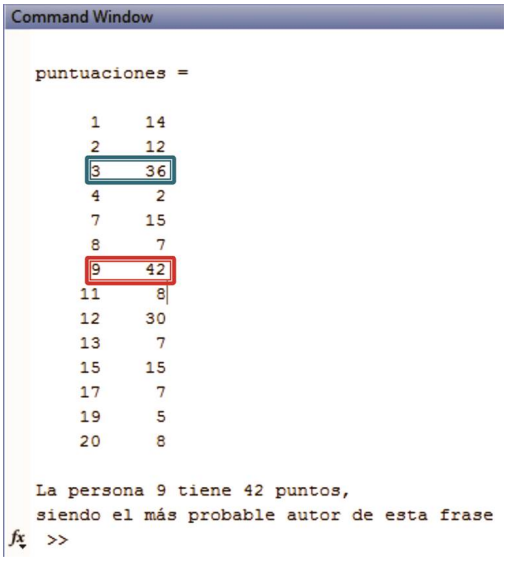

Figura 15. Evaluación de la cuarta oración de la tercera persona.

7. Conclusiones

En el presente artículo se estudiaron varias técnicas de extracción de características y, tras una evaluación de éstas, se escogieron tres para elaborar un sistema de reconocimiento en paralelo. También se explicó el funcionamiento del sistema propuesto y se evaluaron

oraciones nuevas. El sistema presenta un 93% de reconocimiento.

Como propuesta para trabajo futuro, se seguirá probando contra nuevas grabaciones, además de buscar el reconocimiento con cambios artificiales en la voz. También el uso de otro clasificador puede ayudar a tener una mejor tasa de reconocimiento, por lo cual, se propone crear una comparativa de clasificadores con base al sistema propuesto.

Una de las mejoras que se obtuvo con la comparativa de técnicas de parametrización es el cambio en el porcentaje de reconocimiento de utilizar los Filtros de Gabor sobre los MFCC a simplemente el análisis Mel. El reconocimiento mediante análisis directo a MFCC para cambios en voz presenta un 90.8,%, mientras que los MFCC con Gabor arrojan un 97.77%

Agradecimientos

Los autores agradecen el apoyo brindado por el Instituto Politécnico Nacional (IPN) y por el Consejo Nacional de Ciencia y Tecnología (CONACYT), para la elaboración de este documento.

Referencias

[1] I. Mporas, T. Ganchev, M. Siafarikas, and N. Fakotakis, "Comparison of speech features on the speech recognition task," *Journal of Computer Science*, vol. 3, no. 8, pp. 608–616, 2007.

[2] B. Logan, "Mel frequency cepstral coefficients for music modeling." in *International Symposium on Music Information Retrieval*, 2000.

[3] R. Sarikaya and J. H. Hansen, "High resolution speech feature parametrization for monophone-based stressed speech recognition," *Signal Processing Letters, IEEE*, vol. 7, no. 7, pp. 182–185, 2000.

[4] G. A. Martínez and G. Aguilar, "Sistema para identificación de hablantes robusto a cambios en la voz," *Ingenius*, no. 8, pp. 45–53, 2012.

[5] T. Acharya and A. K. Ray, *Image processing: principles and applications.* Wiley, 2005.

[6] R. Solera-Urena, J. Padrell-Sendra, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-De-María, "Svms for automatic speech recognition: a survey," *Progress in nonlinear speech processing*, pp. 190–216, 2007.