UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR
INGENIUS
Revista de Ciencia y Tecnología

Sistema para identificación de hablantes robusto a cambios en la voz

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

Resumen

Los sistemas de reconocimiento de hablante se componen de tres partes principales: preprocesamiento, extracción de características y clasificación de vectores. En el trabajo presente se considera la cuestión de los cambios en la voz, voluntarios e involuntarios, y cómo esto afecta al reconocimiento de hablante. Para este proyecto se detalla todo el pre procesamiento que se realiza sobre la señal y cómo se obtienen los segmentos vocalizados de la misma. También se aplica un modelo de elaboración de vectores característicos basados en ciertas propiedades de la voz, y en Coeficientes Cepstrales en la Frecuencia de Mel (MFCC), así como una Máquina de Soporte Vectorial (SVM) y una Red Neuronal Artificial (ANN) como clasificadores, posteriormente se comparan los resultados obtenidos. Las pruebas realizadas consisten en analizar la trama que se le presenta al sistema, detectar el segmento vocalizado e indicarle al sistema de qué vocal se trata, para posteriormente, identificar a qué persona pertenece dicha vocal. Los resultados muestran que la elaboración de estos vectores conjuntando propiedades y coeficientes MFCC tienen un alto índice de reconocimiento.

Palabras clave: características de voz, coeficientes cepstrales en la frecuencia de Mel, máquina de soporte vectorial, reconocimiento automático del habla, red neuronal artificial.

Abstract

The speech recognition systems consist of three principal parts: preprocessing, features extraction and vectors classification. This paper considers the voice changes, voluntary and involuntary changes, and how this affects the speaker recognition. In this project is explained the preprocessing of the signal and how are obtained the voiced segments. Also is applied a features vector based in voice properties and Mel Frequency Cepstrum Coefficients (MFCC), and a Support Vector Machine (SVM) and an Artificial Neural Network as classifiers. The experiments consist in analyzing the frame presented to the system, detect the voiced segment and indicate to the system which vowel has been said, for an identification of which person pronounce the vowel. The results show the features-MFCC vectors have high rate on recognition.

Keywords: voice features, Mel frequency cepstrum coefficients, support vector machine, speech recognition, artificial neural network.

[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 D.F., México.*

Recibido: 15 - Octubre - 2012, *Aprobado tras revisión: 11 - Noviembre - 2012*

Forma sugerida de citación: Martínez Mascorro, M. y Aguilar Torres, G. (2012). "Sistema para identificación de hablantes robusto a cambios en la voz". *INGENIUS*. N.°8, (Julio/Diciembre). pp. 45-53. ISSN: 1390-650X

1. Introducción

El Reconocimiento Automático del Habla (RAH) se puede dividir en dos áreas: identificación y verificación. La identificación de hablante consiste en asociar la voz de un nuevo individuo, presentado al sistema, con alguna de las voces previamente registradas dentro del mismo. A su vez, los sistemas de identificación pueden dividirse en dependientes e independientes del texto. Los dependientes del texto necesitan que se repita un texto en específico para poder reconocer a la persona, mientras que los independientes deben reconocer cualquier segmento de voz [1].

Uno de los principales problemas que busca resolver este trabajo es el reconocimiento robusto con cambios en la voz, los cuales pueden ser voluntarios e involuntarios. Un cambio de voz voluntario es aquel que se hace de manera consciente, con la intención de no ser identificado por la persona que le escucha, mientras que un cambio involuntario es aquel que ocurre de manera ajena a la intención de la persona, algunos ejemplos: son el enronquecimiento de la voz, o el tono de voz afectado por la constipación nasal.

Dentro de los clasificadores más usados para el reconocimiento de hablante se encuentran los Modelos de Mezclas Gaussianas (GMM), las Redes Neuronales Artificiales (ANN), los Modelos Ocultos de Markov (HMM) y las Máquinas de Soporte Vectorial (SVM).

Los HMM son indudablemente la técnica más empleada para RAH. Durante las últimas décadas, la investigación en HMM para RAH ha dado avances significativos y, consecuentemente los HMM son actualmente muy exactos para esta aplicación. Sin embargo, se está lejos de lograr un sistema de reconocimiento con alto rendimiento [2].

El método de resolución de las SVM se basa en la maximización de la distancia entre las muestras y el borde de clasificación. Esta solución de máximo margen permite a las SVM superar a la mayoría de los clasificadores no lineales en presencia de ruido, que es uno de los problemas duraderos en RAH. También las SVM no tienen problemas de convergencia y estabilidad típicos de otros clasificadores como las redes neuronales artificiales (ANN) [2].

Para poder realizar una correcta identificación, es importante extraer las características principales de la señal de voz. Algunos de los análisis que se realizan son el análisis Cepstral, los Coeficientes Cepstrales en la Frecuencia de Mel (MFCC), la Codificación Predictiva Lineal (LPC), el análisis temporal, la estimación de energía, entre otros [3].

En el presente trabajo se muestran los resultados de las pruebas de reconocimiento, usando vectores con diversas características, como los MFCC y el análisis temporal de la señal. Para las primeras pruebas de reconocimiento se usó una SVM.

2. Marco teórico

2.1 Características de voz

Para la creación de los vectores característicos, se revisaron diversas propiedades que se encuentran presentes en las señales de voz y proporcionan información sobre la calidad de las mismas. En esta sección se presentan las cualidades más significativas que fueron usadas.

2.1.1 Pitch

El pitch, o frecuencia fundamental (F0), es una de las características para modelado de voz usada en muchas áreas de investigación del habla [4]. Al obtener esta característica, se consigue el periodo de pitch por segmentos. El algoritmo utilizado se encuentra en [5].

2.1.2 Contorno de energía en dB

Los parámetros de energía describen características de la amplitud de la señal. Al analizar la energía dentro de una señal de voz se obtiene si esta ha sido modificada por un ruido externo al hablante. Para la obtención del contorno de energía total, se realizó el cálculo de la energía en cada periodo, utilizando el teorema de Parseval, primero se eleva cada término de la señal al cuadrado y luego se convierte a decibelios (dB), y por último se suman. Este proceso se aplica para cada periodo de la señal y se guarda el resultado en un vector. Después de calcular la energía para todos los periodos, se calcula su valor promedio y la variación relativa de este, según la Ecuación 1

$$\Delta en = \frac{\left|\Delta en_{(i)} - \overline{en}\right|}{\overline{en}} \cdot 100 \qquad (1)$$

2.1.3 Coeficientes Cepstrales en la Frecuencia de Mel

Los MFCC son coeficientes para la representación del habla basados en la percepción auditiva humana. Sus bandas de frecuencia están ubicadas logarítmicamente, lo que modela la respuesta auditiva humana más apropiadamente que las bandas espaciadas linealmente.

Aunque los MFCC importan el modelo auditivo del ser humano y mejoran la resolución de la banda de frecuencia, que es sensible a la percepción humana, tiene una desviación orientada entre los formantes y el

área sensible de escucha [1]. En la Figura 1, se muestra el esquema para la obtención de los MFCC.

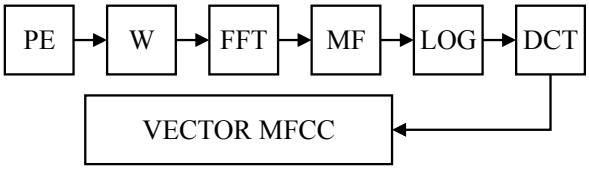

Figura 1. Esquema para la obtención de los MFCC.

Donde PE es el filtrado de pre énfasis, W es el ventaneo, continúa con la transformada rápida de Fourier (FFT), después se convierte a la escala logarítmica (LOG), y se aplica la Transformada Discreta Coseno (DCT).

2.1.4 Jitter

Es la variación de la frecuencia fundamental ciclo a ciclo. Por ejemplo la diferencia promedio absoluta entre periodos consecutivos se expresa como:

$$Jitter = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{2}$$

Donde T_i son las longitudes del periodo de pitch extraídas y N es el número de periodos de pitch extraídos [6].

2.1.5 Shimmer

Está definido como la diferencia promedio absoluta entre las amplitudes de periodos consecutivos, dividido entre la amplitud promedio, y expresado como un porcentaje:

$$Shimmer = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N-1}A_i} \tag{3}$$

Donde A_i es la información de las amplitudes pico a pico extraídas y N es el número de periodos de pitch extraídos [6].

2.1.6 PMR

Es un parámetro de calidad de la señal en el dominio de la frecuencia y consiste en el cálculo de la relación valor pico a valor medio (PMR) del espectro de la amplitud de la señal de voz:

$$PMR = \frac{max\,|S(\omega)|}{mean\,|S(\omega)|} \tag{4}$$

Donde $S(\omega)$ es el espectro de amplitud de la señal de voz calculado mediante la transformada discreta de Fourier [7].

2.2 Segmentos vocalizados

Un segmento de la señal de voz se considera vocalizado si durante su generación el flujo de aire es alterado por las vibraciones de las cuerdas vocales del locutor.

Dicho proceso aporta a la señal resultante características que permiten diferenciarla de las secciones no vocalizadas de la voz. En particular, las componentes vocalizadas presentan un comportamiento periódico y transportan una mayor cantidad de energía que los sonidos no vocalizados [8].

La utilización de un solo criterio de decisión para determinar si un segmento de voz es vocalizado o no, no es suficiente, la mayoría de los trabajos previos en esta área utilizan la combinación de varios criterios, en algunos casos de manera ponderada [9]. Este trabajo se utiliza dos criterios de decisión: análisis de energía de la señal y cantidad de cruces por cero.

2.2.1 Análisis de energía de la señal

El primer criterio se basa en el cálculo de la potencia promedio de la señal la cual se define como:

$$P_{prom} = \frac{1}{N}\sum_{i=1}^{N}|x[i]|^2 \tag{5}$$

Donde x representa la señal de voz a analizar y N el número de muestras que lo forman. A partir de este dato, se obtiene un umbral que de manera empírica fue ubicado como el 50 % de la potencia promedio.

Posteriormente se segmenta la señal y se obtiene la potencia promedio del segmento. Si la señal es mayor al umbral obtenido, el segmento se considera como vocalizado, de lo contrario no es considerado como tal. La Figura 2 muestra un ejemplo de cómo funciona el criterio.

2.2.2 Cantidad de cruces por cero

El segundo criterio está basado en la periodicidad de un segmento vocalizado. Las frecuencias altas tienden a tener un mayor número de cruces por cero. Al tener una mayor amplitud y ser periódico, un segmento vocalizado tiene un menor número de cruces por cero.

La dinámica del criterio consiste en establecer un umbral de cruces, y contar por segmento la cantidad de cruces por cero, en este caso si la cantidad obtenida es menor al umbral se considera un segmento vocalizado.

3. Desarollo del sistema

Para la realización de esta prueba primeramente fueron grabadas diversas voces para la elaboración

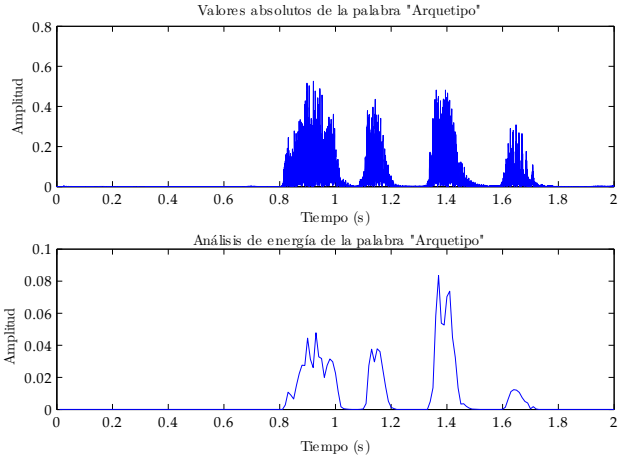

Figura 2. Análisis de energía de la palabra *arquetipo* por segmentos.

de la base de datos, cada una de las grabaciones fue tratada por dos procesos principales: la extracción de vocales y la creación de vectores característicos. Un porcentaje de estos vectores fue utilizado para la producción de los modelos de decisión, mientras que el resto se usaron para pruebas de reconocimiento. En la Figura 3, se muestra el diagrama general del sistema.

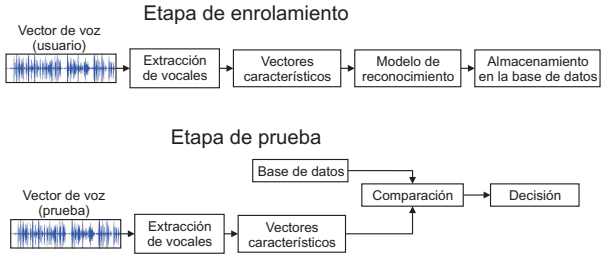

Figura 3. Diagrama general del sistema propuesto.

3.1 Base de datos

Actualmente se cuenta con una base de datos de 8 personas, con un total de 90 vocales por persona, 18 de cada una de las vocales. Para la creación de la base de datos, se realizaron grabaciones con un micrófono externo, a una frecuencia de 8 kHz, grabadas directamente en MATLAB©. La dinámica para las grabaciones consiste en decir la misma palabra con distintos tonos de voz, de modo que se puedan caracterizar los cambios en una persona.

Posterior a esto se hizo una detección de segmentos vocalizados de donde finalmente fueron extraídas las vocales de cada palabra. En la Figura 4, se muestra la comparación de la misma vocal dicha en distintos tonos, por la misma persona y se aprecia cómo varía su frecuencia y su amplitud, mientras que en la Figura 5, se compara la misma vocal en el mismo tono pero dicha por distintas personas.

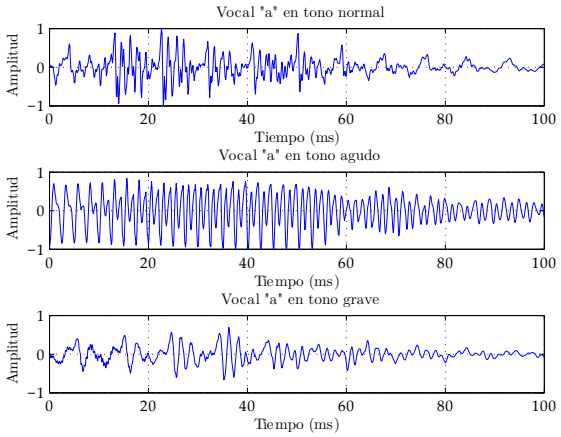

Figura 4. Comparación de la vocal “a” en distintos tonos, con la misma persona.

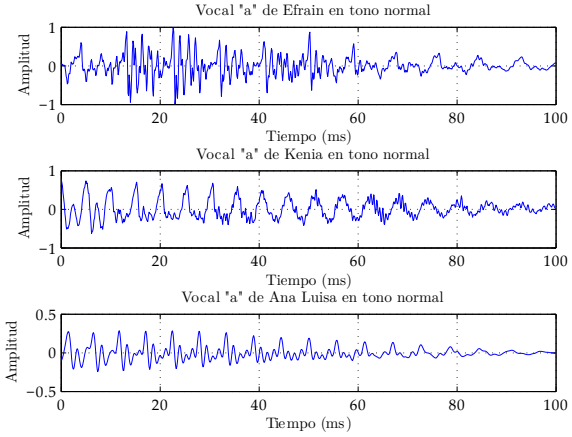

Figura 5. Comparación de la vocal “a” entre distintas personas en el mismo tono.

3.2 Extracción de vocales

Para la extracción de vocales se realizaron los siguientes pasos:

- Filtro pasa - altas de 100 kHz para eliminación de ruido.
- Normalización del vector.
- Ventaneo con la función de Hamming.
- Análisis de energía.
- Cantidad de cruces por cero.

En la Figura 6, se muestra cómo se lleva a cabo el proceso de la detección de segmentos vocalizados,

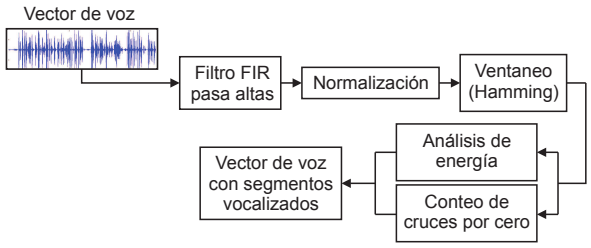

Figura 6. Proceso para la detección de segmentos vocalizados.

para posteriormente extraer las vocales del vector de voz presentado al sistema.

El filtrado de la señal de voz adquirida es una etapa fundamental en el acondicionamiento de la señal. El objetivo de este filtrado es la eliminación de ruido, principalmente el de 60 Hz. Se diseñó un filtro pasa altas de respuesta finita al impulso (FIR) de orden 100 con una frecuencia de corte de 100 Hz [8].

Una vez realizado el filtrado se procede con la normalización. Este paso tiene como finalidad que las señales formadas por componentes semejantes tengan a la salida del bloque de acondicionamiento la misma energía. Se realiza dividiendo el vector de datos entre el valor máximo de la señal.

Una práctica común en los sistemas de acondicionamiento es hacer pasar la señal de cada segmento por una función ventana. Este proceso, llamado ventaneo, permite eliminar errores debido a discontinuidades generadas en la segmentación. Para el ventaneo fueron usados segmentos de 10 ms con una ventana de Hamming de 20 ms, de modo que no haya pérdida de información en las transiciones de una secuencia a la siguiente. Esto resuelve el problema de las transiciones entre tramas sin necesidad de realizar traslapes [8].

Posteriormente se realizan de manera simultánea el análisis de energía y la cantidad de cruces por cero de cada segmento, de manera que al obtener la salida de ambos análisis se determina si cada uno de los segmentos es vocalizado o no vocalizado. En la Figura 7, se muestra el resultado de implementar el algoritmo de detección de segmentos vocalizados en la palabra "arquetipo".

3.3 Vectores característicos

La parte de extracción de características de la señal es una de las más importantes, debido a que de esta manera la máquina podrá modelar el modo de hablar de cada persona. Previamente, se habló de algunos extractores de características, como los MFCC, los LPC, entre otros, sin embargo, en el presente trabajo se usaron medidas estadísticas sobre las características

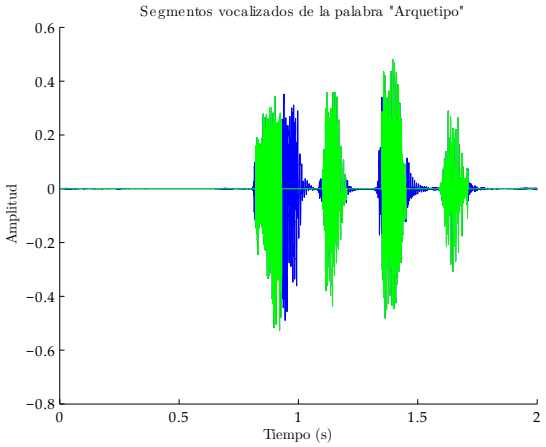

Figura 7. Gráfica de segmentos vocalizados de la palabra "arquetipo".

de voz antes mencionadas y MFCC para la elaboración de los vectores característicos.

Las medidas estadísticas utilizadas son:

- Promedio: Es la suma de los valores numéricos de cada posición comprendida dentro de un intervalo, dividido entre el total de posiciones.
- Desviación estándar: variación esperada con respecto a la media aritmética.
- Mediana: representa el valor de la variable de posición central en un conjunto de datos ordenados.
- Mínimo: variable con menor valor numérico dentro de un conjunto de datos.
- Máximo: representa a la variable con mayor valor numérico en un conjunto de datos.
- Tamaño de rango: diferencia entre el valor máximo y el valor mínimo de un intervalo.

Para las siguientes tres medidas es importante definir dos conceptos: cuantil y cuartil.

El cuantil de orden p de una distribución, con $0 < p < 1$ es el valor de la variable X_p que marca un corte de modo que una proporción p de valores de la población es menor o igual que X_p.

Por ejemplo: el cuantil de orden 0.36 dejaría un 36 % de valores por debajo de él y el cuantil de orden 0.50 corresponde con la mediana de la distribución. Los cuartiles son los tres valores que dividen a la distribución en cuatro partes porcentuales iguales, correspondiendo así a los cuantiles 0.25, 0.50 y 0.75.

- 1^{er} Cuartil: mediana de la primera mitad de los datos.

- 3er Cuartil: mediana de la segunda mitad de datos.
- Rango intercuartílico: diferencia entre el tercer y primer cuartil.

Para conformar el vector primero se usan las medidas estadísticas de las características de voz, y posteriormente los MFCC. Cada característica aporta la siguiente cantidad de coeficientes:

- Periodo de Pitch (9)
- Contorno de energía en decibelios (8)
- Jitter (5)
- Shimmer (5)
- PMR (1)
- Coeficientes cepstrales en la frecuencia de Mel (380)

En la Figura 8, se muestra gráficamente como se lleva a cabo el proceso de la elaboración de los vectores característicos. La señal es introducida a cada uno de los módulos de análisis de características y se obtiene la cantidad de coeficientes antes mencionada para conformar el template.

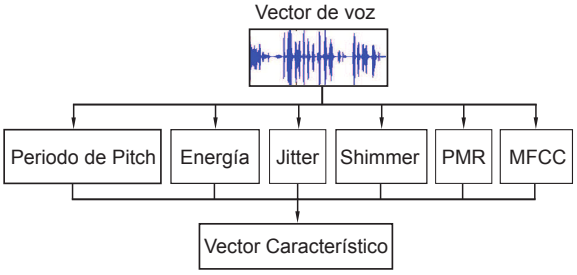

Figura 8. Proceso de elaboración de vectores característicos.

De esta manera se conforman los vectores característicos. En la Figura 9, se observa la similitud de entre los vectores de la misma persona, por colores, y la diferencia entre personas, distintos colores. Mientras que en la Figura 10 se aprecia la diferencia del estilo del vector característico de cada vocal.

4. Reconocimiento con SVM y ANN

La prueba de reconocimiento se realizó con una SVM y una ANN programada para MATLAB©. Una SVM es esencialmente un clasificador binario no lineal, capaz de determinar si un vector de entrada "x" pertenece a

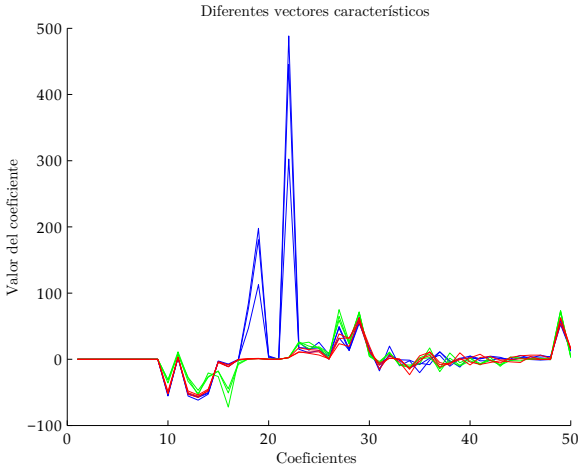

Figura 9. Primeros coeficientes de los vectores característicos de la vocal "a" de tres personas distintas.

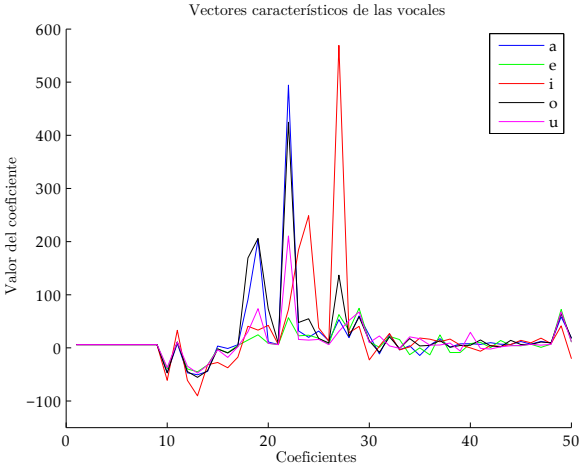

Figura 10. Comparación de los vectores característicos de cada una de las vocales.

una clase 1 (la salida deseada sería entonces $y = 1$) o a una clase 2 ($y = -1$).

Este algoritmo fue propuesto por primera vez en [10] en 1992, y es una versión no lineal de un algoritmo lineal mucho más antiguo, la regla de decisión del hiperplano óptimo (también conocido como el algoritmo de retrato generalizado) que fue introducido en los años sesenta [11]. En la Ecuación 6, se muestra la fórmula general de la SVM y en [11] se encuentran los *kernels* más usados.

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \qquad (6)$$

Para la realización de estas pruebas, además, se utilizó una ANN con un algoritmo de retropropagación, la cual es un tipo de red con aprendizaje supervisado, el cual emplea un ciclo propagación-adaptación de dos

fases.

Una vez aplicado un patrón de entrenamiento a la entrada de la red, este se propaga desde la primera capa a través de las capas subsecuentes de la red, hasta generar una salida, la cual es comparada con la salida deseada y se calcula una señal de error para cada una de las salidas, a su vez esta es propagada hacia atrás, empezando de la capa de salida, hacia todas las capas de la red hasta llegar a la capa de entrada, con la finalidad de actualizar los pesos de conexión de cada neurona, para hacer que la red converja a un estado que le permita clasificar correctamente todos los patrones de entrenamiento [12]. La estructura general se muestra en la Figura 11.

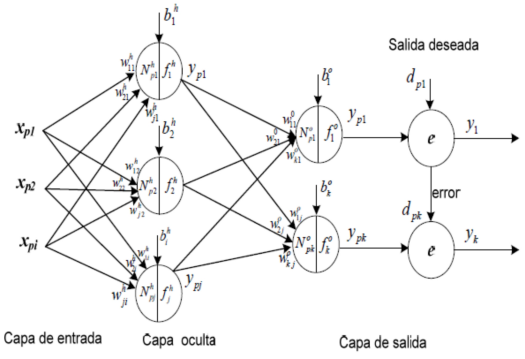

Figura 11. Modelo de la ANN de retropropagación.

La ANN utilizada en el modelo cuenta con tres capas de neuronas. El número de neuronas en la capa oculta depende del tamaño del vector de entrada, mientras que a la salida cuenta con tres neuronas, las cuales dan una salida binaria para ubicar a la persona detectada dentro de la base de datos.

La dinámica del experimento consistió en el proceso indicado a continuación. Una vez determinado el segmento vocalizado, escuchar el mismo e introducir la vocal al sistema para que el mismo determinara de un conjunto de modelos de dicha vocal, quién era la persona que estaba hablando.

Entiéndase por modelo la estructura representativa de cada persona elaborada para la SVM y la matriz de pesos que se obtiene con la ANN.

Cada persona cuenta con 18 vectores característicos por vocal, de este total se usaron 12 vectores para elaborar su modelo por vocal y se dejaron 6 para prueba. Finalmente, se tienen 8 personas cada una con 5 modelos (correspondientes a cada vocal). En la Figura 12 se muestra la pantalla acoplada para cada uno de los sistemas de modo que solo muestra el resultado final del análisis.

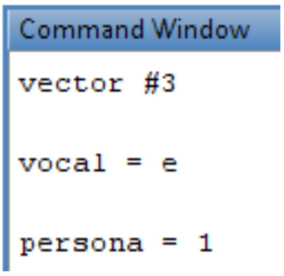

Figura 12. Pantalla de identificación persona dada la vocal correcta.

5. Resultados

Una vez que fueron realizadas las pruebas, los resultados se clasificaron de tres maneras: aciertos, errores e incertidumbres. El primero identifica correctamente la persona a quien pertenece el vector, el error nos da una sola respuesta pero no es la persona correcta, y la incertidumbre se presenta cuando se obtienen como resultado 2 o más personas para un mismo vector, es decir, no hay un resultado concreto para este último. Esto es en el caso de la SVM.

Para el caso de la ANN, como su salida es binaria, no se presenta el caso de la incertidumbre, por lo tanto se definen de otra manera los resultados: acierto, falso rechazo y falsa aceptación.

El primero es un resultado positivo a la persona correcta, el segundo una respuesta negativa cuando se trata de la persona correcta y el tercero es una respuesta positiva cuando se trata de la persona incorrecta.

En la Tabla 1 se muestra cual fue el porcentaje de reconocimiento final de persona por vocal y el porcentaje total de identificaciones usando la SVM y en la Tabla 2 se muestran los resultados de la misma prueba usando la ANN

Tabla 1. Porcentajes de acierto, errores e incertidumbres de la prueba de reconocimiento de la SVM.

Vocal	Aciertos	Errores	Incertidumbres
A	87,50 %	10,41 %	2,08 %
E	91,60 %	8,30 %	0 %
I	97,91 %	0 %	2,08 %
O	89,58 %	8,33 %	2,08 %
U	85,41 %	14,58 %	0 %
Total	90,41 %	8,33 %	1,25 %

Actualmente, la investigación de sistemas de RAH robusta a cambios en la voz no es amplia. El tema presentado es bastante variante debido a factores como el rango de timbres presentado en las muestras, la

Tabla 2. Porcentajes de aciertos, errores e incertidumbres de la prueba de reconocimiento usando una ANN.

Vocal	Aciertos	Falso rechazo	Falsa aceptación
A	31,25 %	68,75 %	0 %
E	33,33 %	66,66 %	0 %
I	14,58 %	85,41 %	0 %
O	31,25 %	66,66 %	2,08 %
U	22,91 %	77,08 %	0 %
Total	26,66 %	72,91 %	0,41 %

duración de las mismas, la detección de los segmentos vocalizados, entre otros.

Algunas tecnologías que han sido desarrolladas, similares a la presentada en su variabilidad, son el reconocimiento de emociones mediante voz y la detección de enfermedades mediante la misma.

Debido a la insuficiencia de trabajos bajo esta área no se puede realizar una comparación objetiva de los resultados del trabajo presente; sin embargo, al analizar las ventajas del sistema propuesto, se pueden visualizar una nueva gama de aplicaciones de seguridad informática y personal, desde nuevas técnicas de protección de archivos y equipos hasta una correcta identificación de personas en situaciones de riesgo personal.

Al usar en conjunto las características de voz con los MFCC se obtuvo un mejor resultado que al entrenar con cada una de estas técnicas por separado. Además, se estudian nuevos análisis que brindarán un mejor resultado a este tipo de reconocimiento.

6. Conclusiones

En este proyecto se propuso el trabajo conjunto de vectores característicos, con propiedades de voz y MFCC, usando una SVM y una ANN como clasificadores.

Es importante señalar que los cambios en la voz son extremadamente variantes, tanto en su frecuencia, su amplitud como en su duración, siendo esta última un factor determinante para la correcta elaboración de un vector característico y para un reconocimiento eficiente.

Con base a los resultados obtenidos (Tabla 1 y Tabla 2) se concluye que el funcionamiento de la SVM fue mucho mejor que el de la ANN para esta aplicación. También es importante señalar que el apoyar al sistema dándole la vocal correcta da un alto resultado de reconocimiento.

Existen diversas combinaciones posibles entre métodos de análisis y clasificadores: la extracción de características, donde se encuentran LPC, las ventanas atómicas, entre otros. Mientras que en los clasificadores existe GMM, HMM, los vectores de cuantización (VQ), por citar algunos.

El próximo trabajo tendrá el compromiso de profundizar en la experimentación con nuevos análisis como el LPC para valorar su funcionalidad en la detección de segmentos vocalizados [11] o sobre la robustez en cambios de voz.

7. 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] Y. Hong-wu, L. Ya-li, and H. De-zhi, "Speaker recognition based on weighted mel-cepstrum," in *Fourth International Conference on Computer Sciences and Convergence Information Technology. ICCIT'09.* IEEE, 2009, pp. 200–203.

[2] J. Padrell-Sendra, D. Martın-Iglesias, and F. Dıaz-de Marıa, "Support vector machines for continuous speech recognition," in *Proceedings of the 14th European Signal Processing Conference, Florence, Italy*, vol. 160, 2006.

[3] M. Kesarkar, "Feature extraction for speech recognition," *Electronic Systems, EE. Dept., IIT Bombay*, 2003.

[4] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," in *Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, China.*, vol. 1000. Citeseer, October, 2000, pp. 676–679.

[5] ——, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. 333–336.

[6] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31*, 2007, pp. 778–781.

[7] P. Del Pino, I. Granadillo, M. Miranda, C. Jiménez, and J. Díaz, "Diseño de un sistema de medición de parámetros característicos y de calidad de señales de voz," *Revista Ingeniería UC*, vol. 15, no. 2, pp. 13–20, 2008.

[8] A. V. Mantilla C, "Análisis, reconocimiento y síntesis de voz esofágica," Ph.D. dissertation, Sección de Estudios de Posgrado e Investigación, Escuela Superior de Ingeniería Mecánica y Eléctrica, Instituto Politécnico Nacional, Agosto, 2007.

[9] L. Siegel and A. Bessey, "Voiced / unvoiced / mixed excitation classification of speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, no. 3, pp. 451–460, 1982.

[10] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th annual workshop on Computational Learning Theory*. ACM, 1992, pp. 144–152.

[11] 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.

[12] L. Cruz-Beltrán and M. Acevedo-Mosqueda, "Reconocimiento de voz usando redes neuronales artificiales backpropagation y coeficientes lpc," in *6to Congreso Internacional de Cómputo en Optimización y Software*. CiCos, 2008, pp. 89–99.