Verificación de Hablantes a través de la Voz: Trabajo Final de Procesamiento Digital de Señales

Iván F. Schweikofski, Camila Saucedo y Darién J. Ramírez
Trabajo práctico final de "Procesamiento Digital de Señales", II-FICH-UNL.

***Resumen*— Este trabajo se centra en desarrollar la verificación del hablante, en la cual se hace uso de una máquina y algoritmos para verificar la identidad que reclama una persona a partir de su señal de voz. El enfoque adoptado en este caso es *texto-dependiente* y lo que se hace es extraer características de la señal de voz de entrada para compararlas posteriormente con características de señales de voz que ya se encuentran almacenadas en la base de datos del sistema, pertenecientes a personas que tienen acceso al mismo.**

***Palabras clave*— verificación, voz, F_0, MFCC, DTW**

I. Introducción

El reconocimiento automático del hablante es el uso de una máquina para reconocer a una persona desde una frase hablada, y puede funcionar en dos modos: para identificar una persona en particular o para verificar la reclamación de identidad de una persona. La verificación del hablante implica decidir si una persona es quien dice ser; la misma pronuncia una frase en el micrófono y el sistema analiza esa señal para aceptar o rechazar su ingreso. La identificación del hablante implica decidir si el mismo es una persona específica o está en un grupo de personas, es decir, el sistema busca cuál hablante de una población de hablantes coincide mejor con el desconocido o si no coincide con ninguno. A su vez, cada uno de estos sistemas puede ser *texto dependiente*, en el que se pronuncia una frase predefinida, o *texto independiente*, donde el hablante puede decir cualquier frase.

Para el sistema de verificación del hablante texto dependiente (Figura 1), se propone un algoritmo que aplique las técnicas del procesamiento digital del habla para procesar una señal de voz, extraer de allí características que identifican a la persona y luego compararlas con características de señales de voz almacenadas en la base de datos del sistema, correspondientes a personas que tienen acceso al mismo.

Existen muchos factores que pueden contribuir a los errores en la verificación, algunos de los cuales pueden ser:

- Frases mal dichas o mal interpretadas.
- Estados emocionales extremos como el estrés.
- Variaciones en la posición del micrófono.
- Acústica del lugar en el que se graba la voz (ruido).
- Utilización de diferentes micrófonos para grabar cada señal de voz.
- Enfermedad que afecta a la persona, por ejemplo, sus cuerdas vocales.
- Envejecimiento.
- Velocidad con la que habla la persona en distintas ocasiones (en algunos casos puede hablar más rápido o más lento).
- Variaciones de entonación que pueda tener la frase enunciada.

Si bien algunos de ellos, ambientales o humanos, pueden reducirse teniéndolos en cuenta en los algoritmos, no siempre pueden ser controlados. En este trabajo se tienen en cuenta las variaciones de velocidad de pronunciación a la hora de comparar las características mediante el algoritmo *Dynamic Time Warping* (DTW) y se analizarán los efectos del ruido en el ambiente.

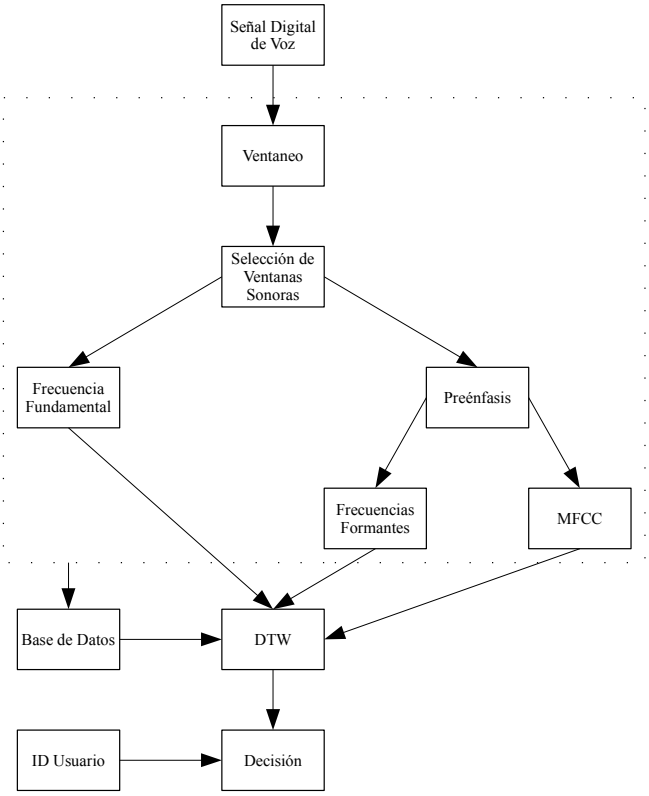

Fig. 1. Sistema de verificación del hablante genérico.

II. Implementación

El sistema de verificación del hablante requiere inicialmente de una base de datos con señales de voz de las personas que tienen acceso y sus respectivas características, las cuales servirán para realizar la comparación cuando una persona solicita el ingreso. Para ello se grabaron señales de voz en diferentes condiciones, para cinco personas distintas, con un mismo micrófono y con una frecuencia de muestreo de 16 KHz. Éstas señales se utilizan en las pruebas experimentales, como parte de la base de datos o como entradas para la verificación de una persona. Sus características son calculadas del mismo modo que para la señal de entrada, lo cual se explicará a continuación.

Las características, tanto de la base de datos como de la señal de entrada al sistema, deben estar normalizadas para ser comparadas y para ello se calcula un vector normalizador, formado por los coeficientes normalizadores correspondientes de cada una de ellas. La función que

calcula este vector, toma todas las características de todas las ventanas de cada una de las señales almacenadas en el sistema y guarda el valor máximo de cada una de ellas.

Para comenzar con la verificación se toma la señal de voz de entrada y una vez cargados sus valores, se procede a ventanearla.

La función que ventanea la señal de entrada utiliza ventanas de Hamming de 20 ms solapadas un 50 % para considerarlas invariantes en el tiempo, además calcula la energía y cantidad de cruces por cero en cada una de las ventanas. De este modo, al recorrerlas todas, retorna en una matriz sólo las ventanas que contienen fonemas sonoros (aquellas con energía mayor a la media de energías de todas las ventanas y con cantidad de cruces por cero menor a la media de cruces por cero de todas las ventanas).

Una vez obtenidas las ventanas sonoras se calcula la evolución de la frecuencia fundamental a lo largo de la señal, es decir, la frecuencia fundamental de cada ventana.

Previo a calcular otras características como las formantes o los coeficientes cepstrales en escala de Mel, se realiza un pre-énfasis en cada ventana de la señal:

$$y[n] = x[n] - ax[n-1] \qquad 0{,}9 \leq a \leq 0{,}97 \qquad (1)$$

Esta función tiene por objetivo enfatizar o incrementar la amplitud de las altas frecuencias para reducir el efecto del ruido y poder identificar mejor las características.

Entonces, utilizando las ventanas con pre-énfasis (Figura 2), se procede a calcular las formantes y los coeficientes cepstrales en escala de Mel de cada una de ellas.

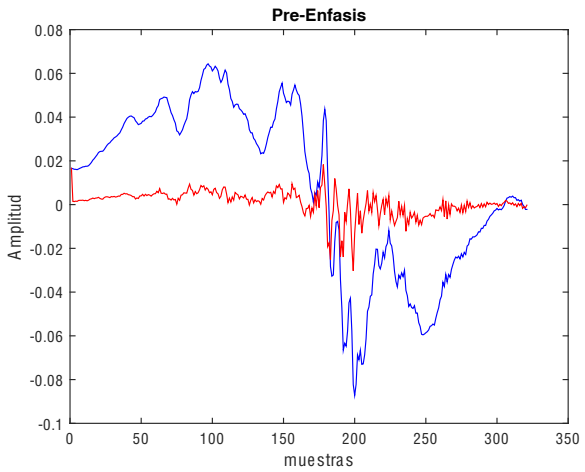

Fig. 2. Señales antes y después del proceso de pre-énfasis.

Todas estas características calculadas antes: *frecuencia fundamental*, *formantes* y *coeficientes cepstrales en escala de Mel* (de los cuales se toman 15) son normalizados por sus respectivos coeficientes, que se encuentran en el vector normalizador calculado inicialmente.

Si bien se calculan todas las características, las utilizadas en la comparación son los coeficientes cepstrales en escala de mel. La matriz de estos coeficientes de todas las ventanas sonoras de la señal de entrada se compara con la de cada señal de voz almacenada en el sistema para luego proceder a decidir si la persona puede o no ingresar. Esta comparación se realiza mediante el método Dynamic Time Warping (DTW) y permite obtener una serie de *puntos de coincidencia* con la base de datos, teniendo en cuenta los desfases temporales que pueden existir entre las señales.

Una vez aplicado el algoritmo DTW para cada señal de la base de datos, se obtiene la mínima de todas las distancias que devolvió. Ese valor mínimo, sumado a un código que ha ingresado el usuario previo a su entrada de voz, son las entradas del módulo que toma la decisión de aceptar o rechazar la solicitud de ingreso. Si ese valor mínimo se encuentra por debajo del umbral definido experimentalmente en 2×10^{-3} y el código ingresado coincide con el correspondiente almacenado en la base de datos, entonces el usuario es aceptado. De lo contrario, se rechaza la solicitud.

III. Sonidos Sonoros y Sordos

Los sonidos sonoros son los que al formarse hacen vibrar las cuerdas vocales (como pronunciar la *a*) y se caracterizan por tener baja cantidad de cruces por ceros y alta energía (Ecuación 2). Por contraparte, los sonidos sordos son los que no producen vibración en las cuerdas vocales (como sucede al pronunciar la letra *s* o la *f*) y tienen mayor densidad de cruces por cero y menor energía.

$$E(\mathbf{x}) = \|\mathbf{x}\|_2^2 = \sum_{n=1}^{N} |x_n|^2 \qquad (2)$$

Los sonidos sonoros poseen cierta periodicidad que permite identificar más claramente características de la señal, como la frecuencia fundamental, y esa es la razón por la cual se utilizan sólo ventanas sonoras para la extracción de características (Figura 3).

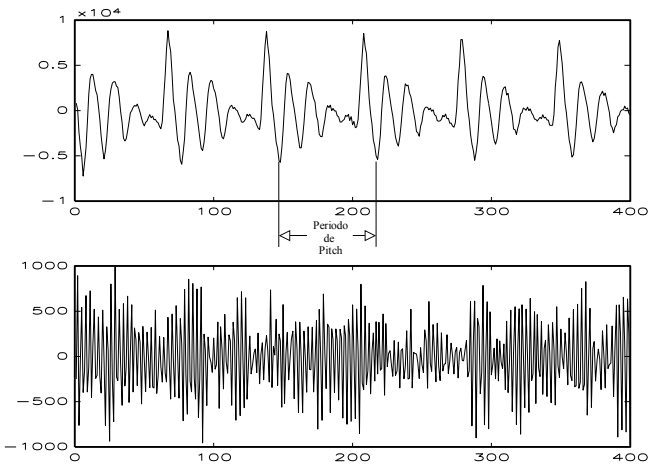

Fig. 3. Señales temporales de la palabra /dos/.

IV. Frecuencia Fundamental (F_0)

Para determinar esta característica se utiliza la autocorrelación de la señal, de la cual se puede extraer el *periodo fundamental* T_0 y, haciendo su inversa, la F_0.

Como normalmente la frecuencia fundamental se encuentra entre 50-300 Hz para el general y representativo de la población, los límites de búsqueda para el período fundamental es

$$\frac{1}{fmax} = \frac{1}{300} \leq T_0 \leq \frac{1}{50} = \frac{1}{f_{min}} \qquad (3)$$

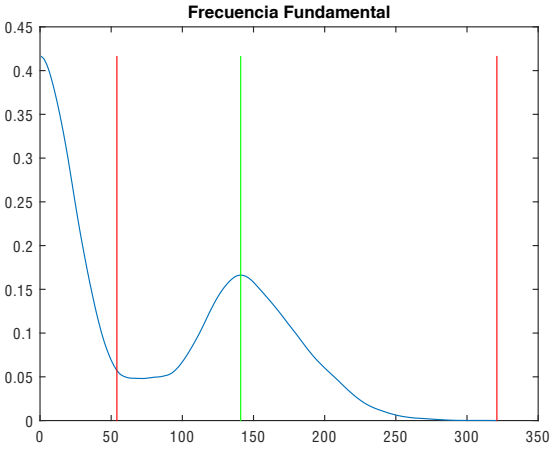

Fig. 4. Frecuencia fundamental y rango de búsqueda.

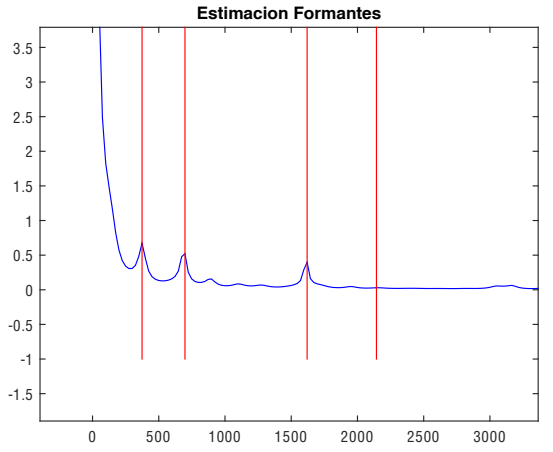

Fig. 5. Reconocimiento de formantes.

V. Formantes

Para obtener estas cuatro características, en un principio se estima la respuesta en frecuencia del sistema que modela el tracto vocal. Para ello se aplica el método de predicción lineal: partiendo de la ecuación de Wiener-Hopf y mediante el algoritmo de Levinson-Durbin se obtienen los parámetros del sistema y el factor de ganancia, con lo cual se modela la función de transferencia $H(z)$. En dicha función se buscan los picos o máximos que corresponden a las frecuencias formantes F_1, F_2, F_3 y F_4 (Figura 5).

VI. MFCC

Los coeficientes cepstrales en escala de Mel son coeficientes para la representación del habla basados en la percepción auditiva humana y se calculan de la siguiente manera:

1. A cada tramo se le aplica la Transformada de Fourier discreta y se obtiene la potencia espectral de la señal.
2. Al espectro obtenido antes, se le aplica el banco de filtros correspondientes a la Escala Mel y se suman las energías en cada uno de ellos.
3. Se toma el logaritmo de todas las energías de cada frecuencia en escala de Mel.
4. Se le aplica la anti transformada de Fourier.

$$F_{mel} = 1000 \log_2 \left(1 + \frac{F_{Hz}}{1000}\right) \tag{4}$$

VII. DTW

El algoritmo DTW es un método para medir la similitud entre dos secuencias temporales que pueden presentar variaciones (como las pequeñas variaciones que existen al pronunciar una misma palabra en diferentes ocasiones).

Esta técnica calcula el grado de coincidencia como la distancia euclídea mínima entre las características de todas las ventanas de dos señales (Figura 7). Para ello encuentra la alineación óptima entre las dos series de ventanas, pues una serie temporal puede ser *deformada* de manera no lineal por estiramiento o contracción a lo largo del tiempo.

Se construye una grilla en la cual se calcula una métrica entre las i-ésimas muestras de las señal de entrada y la j-ésima muestra de la señal de la base de datos. Si ambas secuencias de datos son las mismas, se generará una matriz con ceros en la diagonal. Una vez calculadas todas las distancias, el parecido de ambas señales se determina con la longitud del camino más corto entre la posición $1, 1$ y la N, N. (Figura 6).

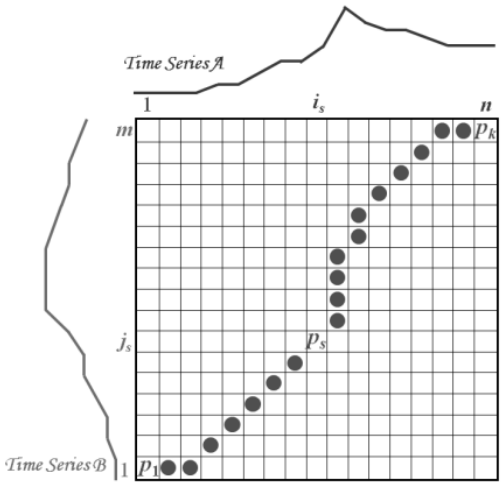

Fig. 6. Dinamic Time Warping.

VIII. Resultados

Todas las señales fueron obtenidas con el mismo dispositivo, con una frecuencia de muestreo de 16 KHz, en condiciones con el menor ruido ambiente posible.

En la siguiente tabla se muestra cómo los diferentes métodos verifican la identidad de la persona. La indicación de ERROR puede deberse a los dos tipos de errores nombrados anteriormente: Falsos Positivos o Falsos Negativos.

En la matriz de confusión (II) se observan los resultados de distintas pruebas realizadas, en las cuales el sistema ha otorgado o denegado el acceso de manera correcta o incorrecta.

Sensibilidad:

$$\frac{VP}{(VP + FN)} = \frac{8}{8+2} = 0{,}8 \tag{5}$$

Especificidad:

$$\frac{VN}{(VN + FP)} = \frac{9}{9+1} = 0{,}9 \tag{6}$$

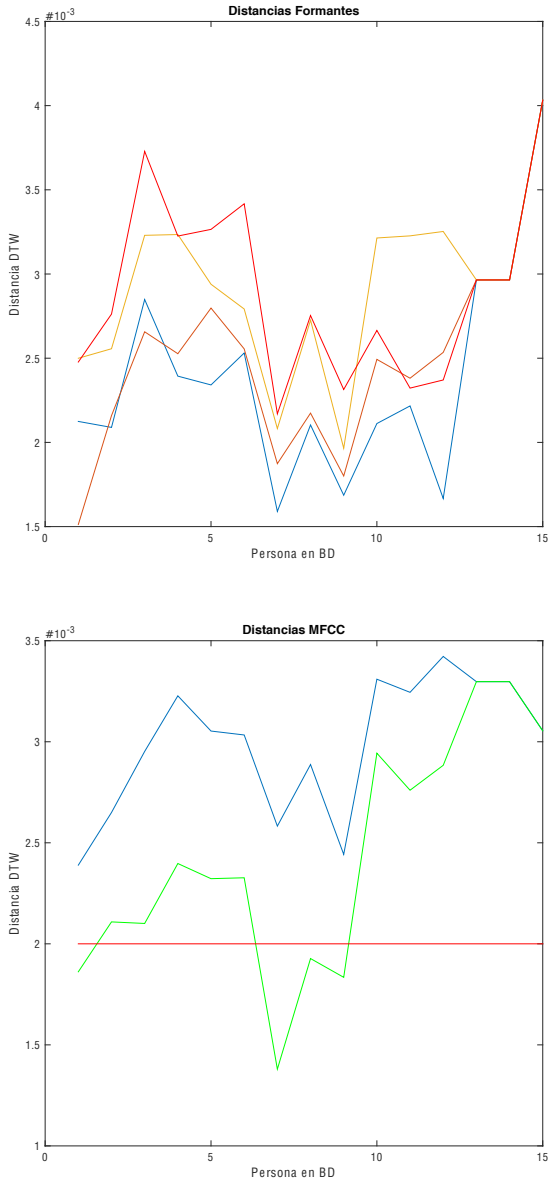

Fig. 7. Distancias DTW.

Intento / Persona	F_0	Formantes	MFCC
1	OK	OK	OK
2	OK	ERROR	OK
3	OK	OK	OK
4	OK	ERROR	OK
5	ERROR	OK	OK
6	OK	OK	OK
7	OK	OK	OK
8	OK	ERROR	OK
9	ERROR	ERROR	ERROR
10	ERROR	OK	OK

TABLA I
PRUEBAS REALIZADAS.

Precisión:

$$\frac{VP}{(VP+FP)} = \frac{8}{8+1} = 0{,}88 \qquad (7)$$

		Valor real	
		Verdaderos	Falsos
Valor predicto	Verdadero	8	2
	Falso	1	9

TABLA II
MATRIZ DE CONFUSIÓN CON MFCC.

Valor predictivo de negativos:

$$\frac{VN}{(VN+FN)} = \frac{9}{2+9} = 0{,}81 \qquad (8)$$

VP: Verdaderos Positivos.
VN: Verdaderos Negativos.
FP: Falso Positivo.
FN: Falso Negativo.

Resultados ante la presencia de ruido artificial agregado aditivamente:

Ruido blanco con media 0 y varianza 0.5. Para SNR = 45db el sistema identifica correctamente 9/10 pruebas.

Ruido ambiente:

Para SNR 30 db el sistema Identifica bien 10/10.
Para SNR 25 db el sistema Identifica bien 7/10.

IX. CONCLUSIONES

La F_0 brinda información sobre la entonación de la frase y es una característica cuyo valor puede ser el mismo para distintos hablantes, de modo que por si sola no es un buen método para verificar la identidad de la persona.

En cuanto a las *formantes* se comprobó experimentalmente que en la comparación realizada con DTW no se puede establecer un umbral debajo del cual las distancias mínimas corresponden a personas que pueden ingresar al sistema (Figura 7). De hecho, en algunos casos ocurrió que la distancia mínima de una persona que no puede ingresar al sistema estuvo por debajo de la de una persona que sí puede ingresar al sistema.

Los MFCC arrojaron muy buenos resultados, con una *sensibilidad* de 8/10 que indica que de 10 pruebas realizadas para acceder al sistema en 2 ha fallado negándole el acceso al mismo. Además posee una *especificidad* de 9/10 que muestra que de 10 intentos de ingreso no autorizado solo 1 lo ha logrado.

Todos los métodos son poco robustos al ruido, para que la verificación sea correcta, se necesita una relación señal/ruido de al menos 30 db.

Ante la distorsión de la voz de una persona que se encontraba pregrabada en la base de datos, la verificación se muestra inestable

REFERENCIAS

[1] JOSEPH P. CAMPBELL, JR. *Speaker Recognition: A Tutorial*. PROCEEDINGS OF THE IEEE, VOL. 85, NO. 9, 1997.

[2] Diego H. Milone, Hugo L. Rufiner, Rubén C. Acevedo, Leandro E. Di Persia, Humberto M. Torres. *Introducción a las Señales y los Sistemas Discretos*. sinc(i), Research Institute for Signals, Systems and Computational Intelligence, Santa Fe, Argentina, 2009.