DESARROLLO DE UN MÉTODO PARA LA EVALUACIÓN DE UN ALGORITMO DE COMPARACIÓN Y DECISIÓN EN IDENTIFICACIÓN DE LOCUTORES CON FINES FORENSES

Miguel Romá Romero

Departamento de Física, Ingeniería de Sistemas y Teoría de la Señal
Universidad de Alicante
mroma@disc.ua.es

José Luis Ramón García

Cátedra de Física Médica. Facultad de Medicina

Universidad de Murcia
relovo01@um.es

Sergio Bleda Pérez

Departamento de Física, Ingeniería de Sistemas y Teoría de la Señal
Universidad de Alicante
sergio@disc.ua.es

Basilio Pueo Ortega

Departamento de Física, Ingeniería de Sistemas y Teoría de la Señal
Universidad de Alicante
basilio@disc.ua.es

ABSTRACT

For the design of a comprehensive speaker recognition system, the first step is to perform a decision algorithm with the role of concluding if two voice samples belong or not to the same person. The measuring method for real forensic purposes must work employing a reduced number of samples of each speaker, so the ones based in statistical characterisation can't be used. Even though the parameter vector must be chosen to reach the minimum error rates, in a first stage the decision algorithm and its parameters will be tested. In order to make the greatest number of combinations, a computer code have been developed to make evident the advantages and disadvantages of the method proposed.

1. INTRODUCCIÓN

Los procesos de identificación de personas por la voz se basan en la parametrización de diferentes muestras de voz y en un sistema que, comparando los parámetros de las muestras empleadas, decida sobre la igualdad o no de tales muestras. Así pues es necesario elegir los parámetros a emplear y el algoritmo de comparación que permita cuantificar el grado de parecido entre las muestras.

Los algoritmos habituales de comparación en los que se basan los sistemas de reconocimiento automáticos emplean una caracterización estadística de los locutores para realizar la comparación, basada en criterios de máxima verosimilitud (o su versión simplificada por medio de la distancia de Mahalanobis), o modelos ocultos de Markov (HMM), que necesitan de un número suficiente de muestras de cada locutor para la fase de entrenamiento. En situaciones de identificación en condiciones forenses, sin embargo, las muestras empleadas son un parámetro no controlable, reducido en número y habitualmente también en longitud, de forma clara en el caso de las muestras dubitadas (la grabación de una amenaza telefónica, por ejemplo), pero también en las indubitadas por problemas de falta de colaboración de la persona implicada. Es necesario, pues, disponer de una medida de distancia que funcione a partir de un número reducido de muestras. Es igualmente necesario emplear un conjunto de parámetros con los que trabajar que permitan un funcionamiento independiente de texto, lo que supone una clara ventaja en aplicaciones forenses frente a los sistemas dependientes de texto.

1.1. Distancia euclídea normalizada

Una forma de paliar el déficit de muestras disponibles para caracterizar al locutor estudiado se basa en comparar las muestras con un conjunto de muestras de voz pertenecientes a otros locutores empleados para "distraer al sistema" [1], algo similar a lo que se hace en las ruedas de reconocimiento de sospechosos. La hipótesis de partida para emplear este método es que un locutor siempre se parecerá más a sí mismo que al resto de los locutores con los que se compare (variabilidad intra-locutor<variabilidad inter-locutor), para lo que habrá que determinar el vector de parámetros más significativo. Para poder comparar la distancia obtenida independientemente del tipo de parámetro empleado es necesario normalizar el resultado de las distintas comparaciones en un rango común. Puesto que el sistema va ha ser empleado para aplicaciones forenses, en las que la inmediatez en obtener el resultado no es una premisa, si el resultado de la medida no es concluyente puede repetirse el proceso con un vector de parámetros diferente. Como prueba del test, la muestra indubitada se compara, además, con una segunda muestra dubitada, de forma que puedan descartarse resultados en aquellos locutores en los que el vector empleado no resulte significativo. La notación empleada es D_n (n=1,..,15) para los distractores, *DB* para la muestra dubitada y ID_n (n=1,2) para las dos muestras indubitadas. Las medidas realizadas son, pues:

$$\begin{aligned} &d(ID_1, D_n), n = 1,..,15 \\ &d(ID_1, DB), d(ID_1, ID_2) \end{aligned} \qquad (1)$$

Aunque esta distancia ha sido utilizada en trabajos previos [2], no se ha estudiado la influencia en el resultado del número de distractores. Los elelmentos que deben determinarse son, por un lado, el número de voces de distracción e emplear, y por otro, el criterio que debe seguirse para decidir si el resultado de la comparación es positivo o negativo. Las alternativas que se estudiarán son dar resultado positivo si *DB* es el vecino más próximo de ID_1, si $d(ID_1,DB)$ está por debajo de un determinado umbral, si $d(ID_1,D_n)>>d(IB_1,DB)$ para cualquier valor de *n*, o por medio de combinaciones de las anteriores. Poder emplear distintos criterios permite adaptar el método a trabajo en grupo cerrado o grupo abierto, permitiendo considerar o no la respuesta nula (el locutor dubitado no es ninguno de los indubitados), lo que diferencia ambas situaciones. En cualquiera de los casos, el valor de $d(ID_1,ID_2)$ será empleado como indicador de validez de la medida.

2. MATERIAL Y METODOLOGÍA

Las muestras empleadas para las pruebas proceden, casi en su totalidad, de la base de datos Ahumada [3]. Los datos que no proceden de la base de datos citada han sido obtenidos a partir de grabaciones realizadas, ex profeso, por los autores del presente trabajo. En la primera versión del programa se trabaja con vectores de dimensión cuatro, aunque este valor puede ser ampliado fácilmente a cualquier valor. Para evaluar el sistema de obtención de distancia se ha empleado como vector de parámetros el formado por las cuatro primeras resonancias del espectro promediado a largo plazo (LTA)[4] por su carácter independiente de texto y por disponerse de datos sobre su capacidad de discriminación en locutores de la base Ahumada [5]. Se ha elaborado una base de datos con tres valores del vector LTA de sendas realizaciones de cada locutor. Como paso previo a la obtención de los parámetros del LTA, en un proceso de conformación de datos, los diferentes archivos son normalizados en amplitud para uniformizar el nivel de los resultados, y los intervalos de silencio en cada grabación son eliminados puesto que su influencia en el análisis tan solo eleva el nivel de ruido.

2.1. Entradas del programa

Los datos de partida son la elección de los locutores dubitado e indubitado así como cuál de las realizaciones de cada uno debe emplearse. Una vez introducidos tales datos, la elección de los distractores se realiza a través de una lista en la que los dos anteriores no pueden ser seleccionados. Respecto a las realizaciones a usar de cada uno de estos, se ofrecen tres alternativas: el empleo en todos la correspondiente a la misma sesión, selección aleatoria por parte del programa o una selección manual por el usuario independiente para cada locutor. Para que el método pueda funcionar con cualquier vector de parámetros, se ofrece la posibilidad de introducir valores de normalización para compensar el desigual peso en la distancia euclídea que pueden presentar componentes de muy distinto orden de magnitud en un mismo vector. En el caso del LTA los valores de normalización empleados son, para cada uno de los cuatro máximos, los valores aproximados de frecuencia en los que se producen las resonancias en el tracto bucal. La distancia entre dos vectores de LTA se mide, finalmente, según la expresión (2).

$$d=\sqrt{\left(\frac{f_1-f_1'}{500}\right)^2+\left(\frac{f_2-f_2'}{1500}\right)^2+\left(\frac{f_3-f_3'}{2500}\right)^2+\left(\frac{f_4-f_4'}{3500}\right)^2} \qquad (2)$$

2.2. Medidas realizadas

El programa realiza los cálculos correspondientes a las distancias con las 15 voces distractoras. A continuación se realiza el mismo proceso pero empleando grupos de 12, 10, 8, 5, 3 y 2 locutores de distracción. La selección de los integrantes de cada uno de los grupos puede realizarse de forma manual por el usuario, en grupos predefinidos, o un número determinado de grupos con componentes elegidos de forma aleatoria. Se consigue estudiar tanto el efecto de la variación del número de voces de distracción como deficiencias derivadas de características particulares de alguno de los integrantes del grupo de distracción. Con cada medida se realiza la decisión teniendo en cuenta cada uno de los criterios expuestos. Los resultados se presentan en forma gráfica y numérica.

3. CONCLUSIONES

El programa descrito permite evaluar las características de un algoritmo de comparación y decisión para identificación de locutores en aplicaciones forenses, basado en la distancia euclídea normalizada en una escala predeterminada. Una vez optimizado el algoritmo y elegidos los parámetros de la voz a emplear para caracterizar a los locutores, con una ligera modificación en el código fuente, se elaborará un programa similar que permita obtener las prestaciones del sistema en función de tales parámetros así como del proceso global de identificación, trabajando de forma automática con los datos obtenidos a partir de las grabaciones disponibles, realizando un elevado número de combinaciones y calculando las tasas de error de falso rechazo y falsa aceptación alcanzadas.

4. REFERENCIAS

[1] Hollien, H. and Jiang, M. “The challenge of effective speaker identification”, Speaker Recognition and its Commercial and Forensic Applications (RLA2C), Avignon 1998.

[2] Jiang, M. “Fundamental frequency vector for a speaker identification system”, Forensic Linguistics 23, 1996.

[3] García, R. y Díaz J.J. “Base de datos de voz para identificación y verificación de locutores”, UPM, 1998.

[4] Furui, S. “Digital speech processing, synthesis and recognition”, M. Decker, N.Y., 1989.

[5] Ramón, J.L., Sánchez, J.A., Canteras, M. y Garcerán, V., “Identificación semiautomática de locutores mediante parámetros extraídos del promedio de espectros suavizados en locutores de larga duración (LTA) y el valor medio de la frecuencia fundamental (F0)”, 1er congreso SEAF, Madrid 2000.