INFLUENCIA DE LOS PARÁMETROS DE UN ALGORITMO DE DECISIÓN BASADO EN DISTACIA GEOMÉTRICA NORMALIZADA EN LAS TASAS DE ERROR EN IDENTIFICACIÓN DE LOCUTORES CON FINES FORENSES

PACS: 43.72.Fx

Romá Romero, Miguel[1]; Ramón García, José Luis[2]; Bleda Pérez, Sergio[1]; Pueo Ortega, Basilio[1]
[1] Dpto. Física, Ingeniería de Sistemas y Teoría de la Señal. Escuela Politécnica Superior
Universidad de Alicante. Apartado de correos 99
03080 Alicante. España
Telf. 965 903 682
Fax. 965 903 682
E-Mail: mroma@disc.ua.es, sergio@disc.ua.es, basilio@disc.ua.es
[2] Cátedra de Física Médica. Facultad de Medicina
Universidad de Murcia. Avda. Teniente Flomesta, nº 5
30003 Murcia. España
E-Mail: relovo01@um.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. A method employed in forensic tasks must be able to work with only one or two samples of the voice in study. The proposed algorithm uses the Euclidean distance between parameter vectors normalised to a predefined range. This method employs distraction voices in the place of the statistical characterisation of the speakers. In the present work the effect of varying the parameters of the algorithm in the error rates is studied.

RESUMEN

Dentro del desarrollo de un sistema global de reconocimiento de locutores resulta esencial el disponer de una herramienta de decisión que, aplicada a los parámetros correspondientes, permita determinar si la muestra de voz dubitada corresponde o no con el locutor indubitado. En aplicaciones forenses no siempre se puede disponer de la información suficiente por lo que se busca un sistema de decisión que pueda funcionar con tan solo una o dos muestras de la voz en estudio. El método propuesto se basa en el empleo de la distancia euclídea normalizada aplicada al vector de parámetros que se considere, en un entorno similar a la rueda de reconocimiento de sospechosos, ofreciendo al algoritmo un número variable de voces “de distracción”. En el presente trabajo se estudia el efecto de los principales parámetros del algoritmo.

INTRODUCCIÓN

Este estudio está enmarcado en un proyecto global de desarrollo de un sistema de identificación de locutores por la voz para aplicaciones forenses, realizado en colaboración con la Unidad de Investigación y Criminalística del grupo de la Policía Judicial de la Guardia Civil. Las etapas principales que deben cubrirse para tal fin son, en primer lugar, la implementación de los algoritmos que permitan medir el grado de parecido entre las muestras de voz comparadas y decidir en función de tal medida si pertenecen o no a la misma persona; en segundo lugar debe determinarse el vector o vectores de parámetros a emplear para la caracterización de los diferentes locutores de forma que se consiga el mayor grado posible de

discriminación, o dicho de otro modo, los que muestren una mayor dependencia del locutor y por último debe procederse al desarrollo del sistema completo del sistema de identificación complementado con el diseño de las herramientas informáticas que se consideren necesarias para su uso. El trabajo presentado se encuentra enmarcado en el primero de los puntos, en el que se analiza la evolución de las prestaciones del sistema de comparación y decisión en función de sus distintos parámetros.

La aplicación del sistema en estudio a aplicaciones de identificación de locutores con carácter forense delimita las prestaciones que debe tener. Las muestras de voz dubitadas no siguen un patrón determinado, por lo que los algoritmos a emplear deben funcionar en modo independiente de texto, de forma que pueda realizarse la comparación de muestras sin dependencia del contenido textual de las mismas. Debe destacarse también que en las aplicaciones en cuestión no es necesaria la inmediatez en los resultados, lo que dota al método de un mayor grado de flexibilidad del que poseen los sistemas de verificación automáticos en los que la aceptación o el rechazo inmediatos son un requisito imprescindible.

Los algoritmos habituales de comparación en los que se basan los sistemas de reconocimiento automáticos, que son en los que más profundamente se está trabajando en la actualidad, 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), entre otros. Para la puesta en funcionamiento de tales sistemas es necesario disponer de un número suficiente de muestras de cada locutor para la fase de entrenamiento. En situaciones forenses, sin embargo, las muestras empleadas son un parámetro no controlable, reducido en número y habitualmente también en longitud, de forma clara en 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 un método que funcione a partir de un número reducido de muestras. El método en estudio, basado en un proceso propuesto por Hollien-Jiang [1], sustituye la caracterización estadística de los locutores por la comparación de las muestras en cuestión justo con un conjunto de tamaño indeterminado de muestras correspondientes a voces de distracción. El objetivo del presente trabajo es determinar de qué manera afectan a las tasas de error los parámetros involucrados en el algoritmo de comparación y decisión empleado. Puesto que el método tan solo tiene sentido si las muestras de distracción son de la misma naturaleza que las muestras en estudio, resulta imprescindible, por cuestiones de viabilidad del sistema, determinar el número mínimo de integrantes del grupo de distracción que asegure unas tasas de error por debajo de los límites buscados.

DISTANCIA EUCLÍDEA NORMALIZADA

El algoritmo para comparar muestras y decidir sobre su identidad basado en la distancia euclídea normalizada presenta la ventaja de funcionar sin necesidad de realizar una caracterización estadística de los diferentes locutores, por lo que puede ser empleado con un número reducido de muestras de voz. Como queda mostrado en [1] parte de su éxito radica en haber sido propuesto por un grupo de trabajo compuesto tanto por expertos en fonética forense como por ingenieros eléctricos y electrónicos. El método destaca por su simplicidad, tanto de cálculo como de implementación. Para paliar el déficit de muestras disponibles, lo que se hace es comparar las muestras en estudio con un conjunto de muestras de voz pertenecientes a otros locutores cuyo papel es de distraer al sistema, en una estructura similar a la que se encuentra en las ruedas de reconocimiento de sospechosos. La hipótesis de partida es que, si el vector de parámetros es suficientemente dependiente del locutor, éste se parecerá más a sí mismo que al resto, es decir, la variabilidad intra-locutor del vector debe ser menor que la variabilidad inter-locutor.

La medida del parecido entre muestras se realiza por medio del cálculo de su distancia euclídea, midiéndose tal distancia entre la muestra dubitada y la indubitada así como con cada una de las muestras de distracción. Puesto que las componentes del vector de parámetros pueden ser de diferente orden de magnitud, cuanto mayor valor presente una componente tanto más peso va a tener en la medida de la distancia. Para evitar un efecto de baremo no

deseado entre las diferentes componentes, la distancia se mide normalizando los valores con un vector de normalización que iguale el orden de magnitud de cada una de las componentes. Aunque el método puede aplicarse a vectores de cualquier magnitud, en principio se trabajará con dimensión cuatro. De este modo, si $V=(v_1, v_2, v_3, v_4)$ y $V'=(v_1', v_2', v_3', v_4')$ son los vectores a comparar y $U=(u_1, u_2, u_3, u_4)$ es el vector de normalización, la medida de distancia empleada resulta:

$$d(V,V') = \sqrt{\left(\frac{v_1 - v_1'}{u_1}\right)^2 + \left(\frac{v_2 - v_2'}{u_2}\right)^2 + \left(\frac{v_3 - v_3'}{u_3}\right)^2 + \left(\frac{v_4 - v_4'}{u_4}\right)^2}$$

Si las componentes del vector de parámetros presentan un diferente grado de discriminación, el vector de normalización puede emplearse también para otorgar un mayor peso a las componentes más discriminantes. Con el fin de que el mismo proceso pueda aplicarse independientemente a cualquier vector, y para que los resultados puedan compararse directamente, el resultado de la medida de la distancia entre las diferentes muestras se normaliza a su vez en un rango común, de modo que la distancia menor valga 1 y la mayor 10. De este modo, si el resultado de una medida no es concluyente acerca de la identidad del locutor dubitado, la prueba puede complementarse repitiendo la medida con un vector de parámetros distinto.

MATERIAL Y METODOLOGÍA

Las muestras de voz empleadas en el estudio pertenecen a la base de datos Ahumada [2], que consta de un corpus de 103 locutores de los que se han realizado grabaciones microfónicas y telefónicas empleando diferentes micrófonos y aparatos telefónicos en tres sesiones espaciadas en el tiempo. De las distintas realizaciones de cada uno de los locutores se han empleado aquellas relativas al habla espontánea pertenecientes a todas las sesiones de grabación realizadas empleando el micrófono SONY ECM-66B, con un total de tres realizaciones por cada uno de los locutores empleados. Las grabaciones se encuentran digitalizadas empleando una frecuencia de muestreo de 16 kHz. Para homogeneizar los valores de los vectores obtenidos para la caracterización de los locutores se ha realizado un procesado de normalización de la amplitud de las grabaciones empleadas, ajustando el valor de mayor amplitud a 0dB en la escala digital en todos los casos. Igualmente se ha procedido a eliminar los periodos de silencio de las grabaciones puesto que su presencia tan solo aumenta el nivel de ruido en los resultados. Se ha considerado silencio a los segmentos con un nivel inferior a –38 dB durante un tiempo superior a 5 ms, medidos en los ficheros con la amplitud normalizada. Tras la eliminación de los intervalos de silencio, los ficheros originales con una duración de aproximadamente 1 min:30 s, quedan finalmente reducidos a una longitud de entre 50 y 55 s.

Puesto que no es objeto de este estudio validar un determinado vector de parámetros para su uso en sistemas de reconocimiento de locutores, las pruebas se han llevado a cabo empleando el espectro promediado a largo plazo (LTA, *Long Term Average Spectrum*), por su carácter independiente del texto y por disponerse de información relativa a su capacidad de discriminación sobre los locutores de la base de datos empleada [3]. El LTA se ha calculado empleando secuencias de 512 puntos, con ventana de Hamming y un pre-énfasis de 0.980. El vector se forma con las cuatro primeras frecuencias resonantes de LTA, considerando un vector de normalización $U=(500, 1500, 2500, 3500)$ ya que son los valores aproximados de frecuencia en los que se producen las resonancias. De la totalidad de locutores de la base de datos se han elegido 18 con la condición de que en las tres realizaciones parametrizadas estuvieran presentes, como mínimo, tres de los cuatro máximos. De los 18, 14 presentan los cuatro máximos en cada una de las realizaciones, mientras que de los cuatro restantes, tres presentan tres máximos en una realización y cuatro en las otras dos y el cuarto presenta tres máximos en dos realizaciones y cuatro en la tercera. Cuando se ha empleado un locutor cuyo LTA carece de un máximo, el factor correspondiente en el cálculo de la distancia se ha supuesto igual a cero, de forma que no influya en el resultado. En cada caso estudiado los locutores pueden ser designados como dubitado (*DB*), indubitado (*ID*) o distractor (D_n).

Para proceder a la aceptación o rechazo de la muestra dubitada como la indubitada se realiza la medida de la distancia euclídea normalizada entre las diferentes muestras, obteniéndose los valores de:

$$d(ID, DB)$$
$$d(ID, D_n), \forall n$$

Se estudian tres criterios diferentes para decidir sobre la identidad del locutor. En primer lugar se considera aceptación si el locutor DB es el vecino más próximo de ID (si es el que más se parece), en segundo lugar si $d(ID,DB)$ está por debajo de un determinado umbral U (si DB es suficientemente parecido a ID) y en tercero si se cumple el primer criterio y, además, $d(ID,D_n)$ está por encima del umbral para todos los distractores. Para el sistema definitivo se considera también una prueba de bondad del método en la que se comparan dos muestras del locutor indubitado para corroborar que el vector empleado es en realidad poco variable para éste ($d(ID_1,ID_2)$). En el presente trabajo se estudia el efecto del número de locutores de distracción así como del valor del umbral empleado en la decisión en las tasas de error alcanzadas empleando cada uno de los criterios.

Para el análisis de los resultados se ha empleado un programa diseñado al efecto [4] cuyas entradas son el locutor dubitado, el indubitado y los distractores, cuál de las realizaciones se emplea en cada caso, el umbral de decisión y el vector de normalización. A partir de estos datos el programa calcula las distancias entre los diferentes locutores y presenta resultados gráficos, numéricos y estadísticos analizando todos los casos posibles con grupos de 15, 12, 10, 8, 5, 3 y 2 distractores (figura 1).

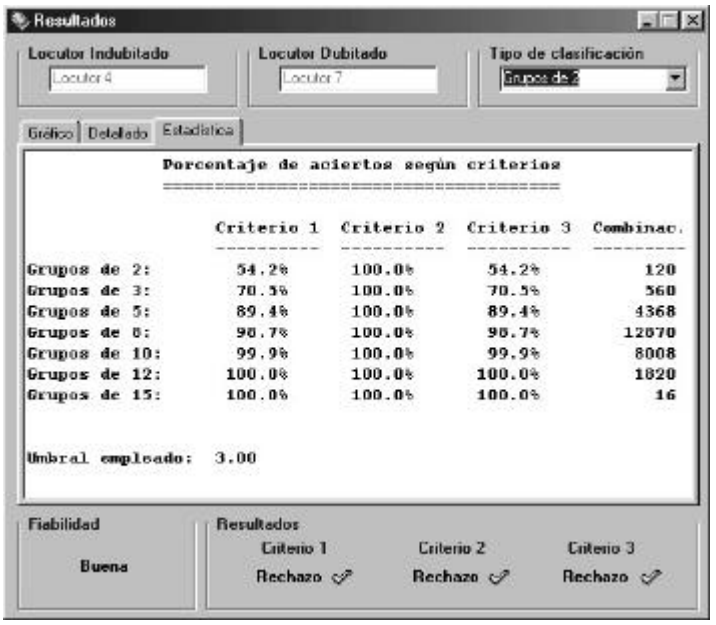

Figura 1. Presentación de datos estadísticos del programa empleado para el estudio de los parámetros.

En el caso en el que todos los locutores sean designados como distractores, exceptuando aquellos empleados como DB y ID, el número total de casos analizados resulta de casi 28000 combinaciones con los diferentes grupos de distractores, para cada pareja de DB-ID. Se han realizado dos series de medidas, considerando, por un lado, casos en los que el locutor indubitado y el dubitado son el mismo (cuyo resultado debería ser de aceptación), y por otro, los casos en los que son personas distintas (que deberían ser rechazados). Puesto que no es objeto del trabajo determinar el o los vectores de parámetros con un mayor índice de discriminación, para las pruebas en las que $ID=DB$ (cierta aceptación o falso rechazo) se han elegido como locutores indubitados aquellos que presentan al menos dos de las realizaciones con valores de LTA razonablemente parecidos. En las pruebas de cierto rechazo o falsa

aceptación no se ha considerado tal restricción puesto que los locutores dubitado e indubitado son distintos.

RESULTADOS

En la figura 2.a se muestran los resultados correspondientes a los porcentajes de aceptación positiva al variar el número N de locutores que integran el grupo de distracción, en el caso en el que $ID=DB$, correspondiendo cada serie al locutor indicado en la leyenda. Los datos presentados pertenecen a la evaluación según el segundo criterio (identificación positiva si $d(ID,DB)<U$), realizada para un valor umbral U=3. Como puede verse, la tendencia generalizada, que coincide con la observada en todos los casos estudiados, es de un crecimiento en el índice de aciertos al aumentar el tamaño del grupo de distracción. El crecimiento, en la mayoría de los casos, se estabiliza a partir de 8 locutores distractores. En la figura 2.b se puede ver la tendencia que sigue el índice de aciertos en cada criterio al variar el número de distractores. En el segundo criterio aumenta el porcentaje de aciertos con el tamaño del grupo de distracción, mientras que en los criterios 1 y 3 la tendencia es la contraria.

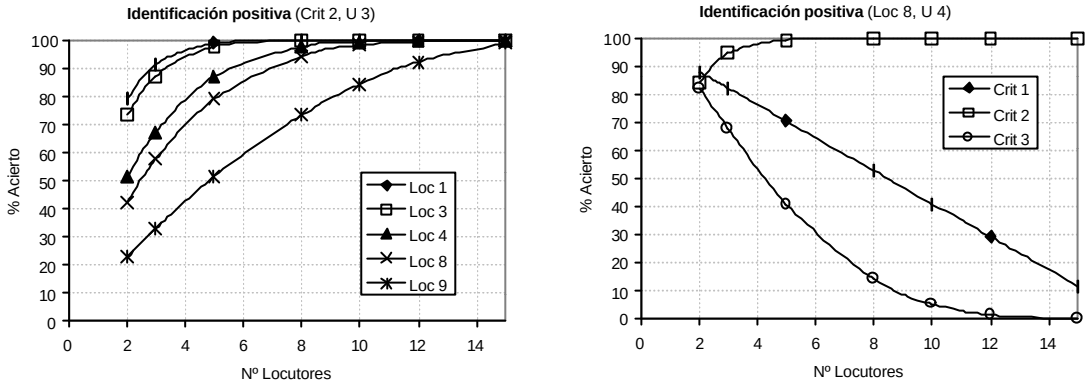

Figura 2. Aciertos (%) si $ID=DB$ (A) para distinto locutor y el criterio 2, (B) tendencia para un locutor y los tres criterios.

Para estudiar los aciertos en el caso en que $ID \neq DB$ (figura 3), se han empleado los mismos locutores como indubitados que en el caso anterior, a cada uno de lo cuales se han presentado dos locutores dubitados. Para el análisis de resultados se ha elegido el peor de los casos obtenido con cada locutor.

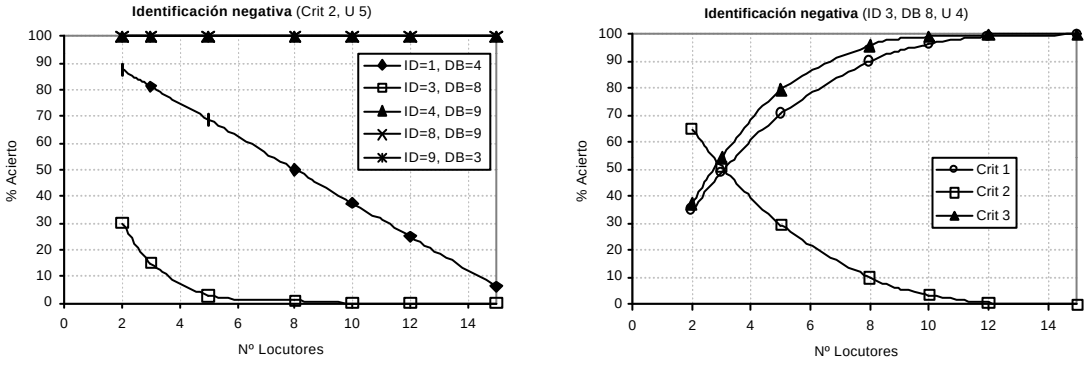

Figura 3. Aciertos (%) si $ID \neq DB$ (A) para distinto locutor y el criterio 2, (B) tendencia para un locutor y los tres criterios.

Para representar la figura 3.a se han empleado los correspondientes a un valor de U=5, porque para valores menores el porcentaje de aciertos prácticamente no baja del 100%, por lo que no puede verse la tendencia. En cualquier caso, la tendencia general mostrada por los tres criterios es la contraria a la encontrada en los casos en los que $ID=DB$ (figura 3.b).

Para estudiar el efecto del umbral se ha realizado la media de los resultados obtenidos para cada criterio en todas las pruebas realizadas para diferente tamaño de la muestra de distracción. En la figura 4.a se muestran las tasas de error de falso rechazo (EFR) y falsa

aceptación (*EFA*) para N=8 y para N=5, obteniéndose un umbral óptimo respectivo $U_{N=8}$=3 y $U_{N=5}$=3.5, con unos valores de la tasa de equierror $ERR_{N=8}$=1.4% y $ERR_{N=5}$=7.5%. En la figura 4.b se muestra la tendencia general de los criterios 2 y 3 (los que dependen del valor de *U*) con el umbral. Como es de esperar, los errores de falsa aceptación disminuyen al aumentar U, lo contrario que pasa con los errores de falso rechazo.

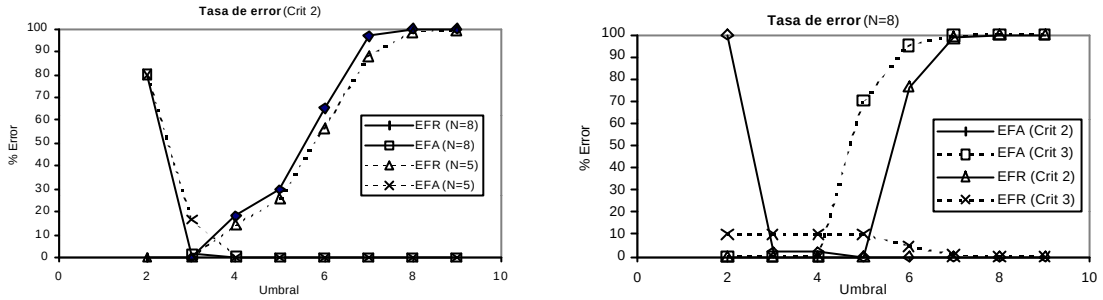

Figura 4. (A) Tasa de EFR y EFA para el criterio 2, con 8 y 5 distractores. (B) Evolución de los valores de EFA y EFR según los criterios 2 y 3 al variar el umbral de decisión.

CONCLUSIONES

Aunque el trabajo realizado es sólo un primer paso para desarrollar un sistema global, los resultados obtenidos en las primeras pruebas realizadas son alentadoras para el fin que se persigue, destacando la sencillez del método. Debe remarcarse que es necesario estudiar y elegir el vector de parámetros a emplear para la caracterización de los locutores, con la premisa de que funcione de forma independiente de texto.

Los resultados obtenidos se estabilizan a partir de un número de locutores del grupo de distracción igual a ocho, aumentando rápidamente las tasas de error para valores menores. Disponer de ocho muestras de la misma naturaleza de la voz en estudio resulta un valor razonable lo que hace pensar en la viabilidad del sistema con el método propuesto.

De los tres criterios estudiados el segundo es el que entrega unos resultados más coherentes, aunque no se descarta que las tasas de error mejoren empleando una combinación de criterios para la decisión final. En cualquier caso, los criterios 1 y 3 por sí mismos no se muestran suficientemente robustos.

AGRADECIMIENTOS

Los autores expresan su agradecimiento a Manuel Canteras Jordana, catedrático de bioestadística de la Facultad de Medicina de la Universidad de Murcia, por su asesoramiento en materia de estadística.

REFERENCIAS

[1] Hollien, H. and Jiang, J., “The challenge of effective speaker identification”, Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pp 2-9, Avignon 1998.
[2] García Jiménez, R. y Díaz Gómez, J.J., “Base de datos de voz para identificación y verificación de locutores”, EUITT-UPM, 1998.
[3] Ramón, J.L., Sánchez Molero, 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, pp. 163-168, Madrid 2000.
[4] Romá Romero, M., Ramón García, J.L, Bleda Pérez, S. y Pueo Ortega, B., “Desarrollo de un método para la evaluación de un algoritmo de comparación y decisión en identificación de locutores con fines forenses”, XVI Simposium Nacional de la Unión Científica Internacional de Radio (URSI), Madrid 2001.