VI Congreso Iberoamericano de Acústica - FIA 2008
Buenos Aires, 5, 6 y 7 de noviembre de 2008

FIA2008-A058

IDENTIFICACION BIOMETRICA DE LOCUTORES PARA EL AMBITO FORENSE: ESTADO DEL ARTE

Felipe Ochoa[(a)], César San Martin[(b)], Roberto Carrillo[(b)].

(a) Laboratorio de Criminalística Regional Temuco, Policía de Investigaciones de Chile. Arturo Prat N° 19, Temuco, Chile. E-mail: fochoae@investigaciones.cl

(b) Departamento de Ingeniería Eléctrica, Facultad de Ingeniería, Ciencias y Administración, Avenida Francisco Salazar N°01145, Temuco, Chile.

Abstract

This paper presents a review in biometrics speaker identification methodologies and its application in forensic science. Initially, the features extraction step to transform the voice in features vectors is presented and then, different pattern recognition algorithms are briefly treated, such that vectorial quantization, neural networks, hidden Markov models, Gaussian mixture models and support vector machines. Finally, by means of the Bayes decision rule, the adaptation of these methodologies in forensic science and results using real data are presented.

Resumen

En este trabajo se presenta una revisión bibliográfica actualizada sobre las metodologías de identificación biométrica de locutores y su aplicación en el ámbito forense. Inicialmente, se muestra la etapa de extracción de características para transformar la voz humana en vectores característicos, para luego revisar los brevemente los principales algoritmos clasificadores, entre los cuales se destacan Cuantización Vectorial, Redes Neuronales, Cadenas Ocultas de Markov, Modelos de Mezclas Gaussianas y Maquinas Vectoriales. Finalmente, por medio de la regla de decisión de Bayes, se presenta la adaptación de esas metodologías en el trabajo forense y su aplicación a datos reales.

1 Introducción

La biometría refiere al concepto de identificación de individuos a través de rasgos biológicos, personales y unívocos. El enfoque forense implica la aplicación de la biometría para identificar individuos mediante estos rasgos, los que son dejados en las evidencias que conforman un caso delictual. La tecnología actual ha desarrollado diversas técnicas para poder automatizar estos procesos desarrollando algoritmos para poder extraer, cuantificar y clasificar las características biométricas, con mayor o menor éxito según el atributo a analizar. En el caso de la identificación humana a través de la voz, ésta ha tenido un desarrollo sostenido en las últimas dos décadas, obteniendo avances significativos en la capacidad de discriminación a partir de la presentación del Modelo de Mezclas Gaussianas en 1995. La principal dificultad de esta medida biométrica radica en la intravariabilidad inherente al proceso de producción de la voz, situación que ha sido resuelta con razonable éxito con la aplicación de clasificadores adecuados. El documento comienza con el método convencional

para extracción de las características más adecuadas de la voz, primer elemento en el paradigma de reconocimiento de patrones. Luego se explican brevemente los principales clasificadores usados para separar las clases obtenidas, a saber, Cuantización Vectorial (VQ), Redes Neuronales Artificiales (NN), Cadenas Ocultas de Markov (HMM), Modelo de Mezclas Gaussianas (GMM) y Máquinas de Vectores de Soporte (SVM). Finalmente, el testing del sistema finaliza con un umbral de decisión sobre la discriminación de las clases, el cual es adecuado para aplicaciones de control de acceso, pero no para la inferencia forense. En ese sentido, se entrega un concepto bayesiano de interpretación de los valores de test, que permite migrar de un sistema de discriminación a uno de aproximación probabilística, susceptible de ser usado como evidencia en una corte.

2 Paradigma del Reconocimiento de patrones

El objetivo de un reconocedor de patrones es, dada una serie de atributos de una clase desconocida (entiéndase clase como el elemento a identificar por el reconocedor, formas, letras, caras, voces, etc.), identificarla dentro de las clases conocidas por el reconocedor. Así, esta metodología consta de dos etapas: un entrenamiento del reconocedor para caracterizar las distintas clases con sus peculiaridades (*training*) y el proceso de identificar características desconocidas en este universo entrenado (*testing)*

El reconocimiento automatizado de hablantes comienza con la definición si el sistema será un Identificador o Verificador. El primero corresponde a un reconocimiento dentro de N clases o individuos, siendo este conjunto *Open-Set* o *Closed-Set*. *Open Set* responde a la eventualidad que el individuo target no esté dentro del conjunto de individuos, mientras que *Closed Set* incluye el target dentro de las clases disponibles. Un verificador de locutor representa la aceptación o rechazo sobre una clase en particular. Es decir, en el proceso de *testing* se evalúan los atributos desconocidos y se discrimina si corresponde o no a la clase invocada. Finalmente, el reconocedor de locutor puede ser Dependiente o Independiente de texto. Un reconocedor dependiente de texto limita su accionar al uso de los mismos fonemas tanto en el entrenamiento como en las pruebas, mientras que un sistema independiente de texto libera esta condición.

3 Extracción de Características

En esta primera etapa, se busca reconocer y adquirir los atributos que mejor representen los rasgos de interés, que en este caso corresponden a información del hablante. La señal de la voz se descompone en dos tipos de información: uno referente a los sonidos que se emiten y otro que caracteriza el locutor, que dice relación con las resonancias internas del tracto vocal.

3.1 Modelo del aparato fonador humano

El tracto vocal posee cavidades cuyas dimensionen varían de una persona a otra y donde el sonido generado en la cavidad glótica, impulsado por los pulmones u diafragma, modifica su contenido espectral conforme resuena en las cavidades laríngea, nasal y vocal (Fig. 1). Las sucesivas resonancias conforman un espectro característico del sonido emitido, que en su estructura fina es cuasi estacionario ante la variación de fonemas. La figura 2 muestra un modelo físico de filtros del aparato fonador.

Figura 1. Tracto vocal simplificado (Batvox Training, 2006)

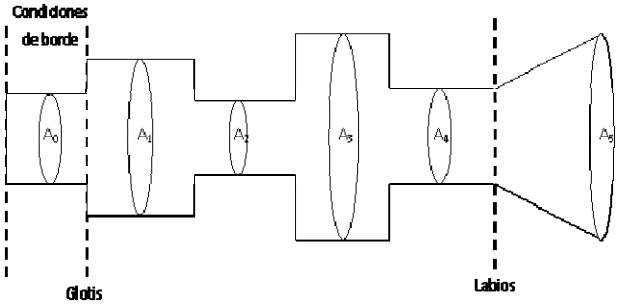

Figura 2. Sistemas de filtros resonadores que modelan el tracto vocal (Campbell, 1997)

Este modelo simplificado permite la obtención de ecuaciones que representen el comportamiento de la onda sonora que transita por estos filtros. Desde esta modelación derivan los coeficientes de predicción lineal o LPC que caracterizan el tracto. De manera alternativa, otra aproximación se enfoca también en parametrizar la envolvente espectral del sonido emitido, a través del uso de un banco de filtros, llamados filtros Mel. Esta última es más extensamente utilizada y ambas desembocan en la transformación a los coeficientes cepstrales.

3.2 Coeficientes Cepstrales y Mel Cepstrales

La parametrización de la señal del habla consiste en transformar esta señal en un set de vectores de características. Esta transformación permite obtener una representación más compacta, menos redundante y que sea capaz de ser comparada a través de un score o puntaje. Uno de los métodos más extendidos es el uso de la representación cepstral de la señal vocal. La figura 3 esquematiza la obtención de los coeficientes cepstrales para el algoritmo LPC, mientras que la figura 4 calcula los coeficientes mel cepstrales para un modelo basado en la envolvente espectral.

Figura 3. Diagrama de parametrización de la señal hablada basada en el algoritmo LPC

El ventanamiento corresponde al tiempo de apertura en la adquisición de la señal, siendo típicamente una ventana Hamming o Hanning de 20 milisegundos, con 10 milisegundos de solapamiento. El algoritmo LPC tiene su principio en estimar los parámetros de un filtro auto regresivo sobre una ventana. Cuando la ventana se mueve, unos nuevos coeficientes (llamados Coeficientes Predictivos) son estimados.

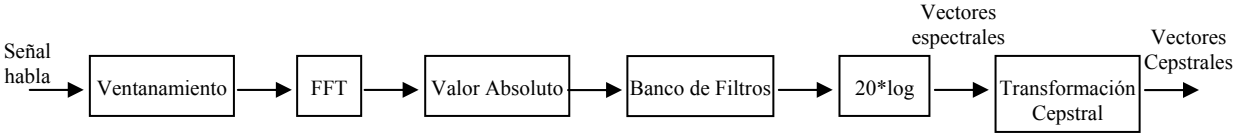

Figura 4. Diagrama de parametrización del habla basado en el uso de filtros Mel.

El banco de filtros que usualmente se utilizan corresponde a la escala Bark/Mel, la cual se basa en un comportamiento no lineal del oído humano dado por

$$fmel = 1000 \times \frac{log(1+f\text{linear} \div 1000)}{\log 2}, \quad (1)$$

donde *f linear* es la frecuencia lineal en Hz y *f mel* es el nuevo valor de frecuencia en Hz, calculado por (1). Para este caso, el cálculo de los coeficientes cepstrales se obtiene a través de la transformada discreta del coseno:

$$c_n = \sum_{k=1}^{K} s_k \times cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right],\ n = 1, 2, \ldots, L, \quad (2)$$

donde K es el número de coeficientes espectrales calculados, L es el número de coeficientes cepstrales a calcular y s_k es el k-ésimo coeficiente espectral. Los vectores obtenidos serán un grupo de valores para cada ventana de adquisición. Típicamente, se utilizan los 19 primeros coeficientes, pero además se incorpora información dinámica de la variación de estos en el tiempo. Así, se agregan los coeficientes Δ (velocidad) y ΔΔ (aceleración)

4 Algoritmos de Clasificación

Una vez que se extraen las características de cada individuo, la siguiente etapa es clasificarlos, agrupándolos de tal forma que su separación pueda ser medida y comparada. Si se considera el cálculo de al menos 38 coeficientes cepstrales por cada 20 milisegundos de señal de la voz, entonces se tiene 1900 coeficientes por cada segundo, que pueden ser agrupados usando diversas técnicas, siendo las principales las que se definen a continuación.

4.1 Cuantización Vectorial (VQ)

Este método ha sido tempranamente utilizado como algoritmo de clasificación de las características del hablante. Se basa en aprendizaje no supervisado, donde un algoritmo (K-means) agrupa automáticamente cada clase conocida (set de hablantes). Existen muchos algoritmos de agrupamiento, siendo los más conocidos Linde-Buzo-Gray (LBG) y learning vector quantization (LVQ).

En esencia, las características de cada uno de los distintos individuos se dividen en un set de regiones convexas mutuamente exclusivas, computándose el centroide de cada región. La colección de centroides se denomina *codebook*, siendo típico en reconocimiento del hablante el cálculo entre 32 y 64 centroides por *codebook*. Para el proceso de test, el vector de características de una voz desconocida se compara con las distancias a los distintos *codebook*, siendo el matching la distancia acumulada mínima.

4.2 Redes Neuronales (NN)

Otra categoría de algoritmos clasificadores son los clasificadores supervisados, siendo el más popular las redes neuronales, en particular del perceptrón multicapa (Fig. 5). El vector de coeficientes cepstrales se multiplica por los respectivos pesos y los resultados se suman y se evalúa en la función de activación, del tipo sigmoidal, escalón u otra. En este caso, la salida puede ser 1 o -1, simbolizando la pertenencia o no a alguna clase. La esencia principal del aprendizaje supervisado responde a la existencia de una serie de ejemplares de entrenamiento, los cuales pasan sucesivamente por esta red neuronal, en este caso forzando uno de los dos valores en la salida, lo que trae consigo que el error sea asumido por la modificación de los valores de los pesos (retropropagación del error). Múltiples perceptrones son utilizados para separar más clases, siendo usual que la salida de un perceptrón sea la entrada a otro. Gráficamente, el perceptrón multicapa modela una línea de decisión de las clases entrenadas, lo que trae consigo que para incorporar una nueva clase, se deba entrenar nuevamente toda la red. En el caso del reconocimiento del hablante, el entrenamiento se realiza con las voces de todos los individuos de interés. En el test, el vector de características de la voz desconocida es sometido al paso de esta red ya entrenada (con todos los pesos determinados) y su salida se asociará con alguna de las clases.

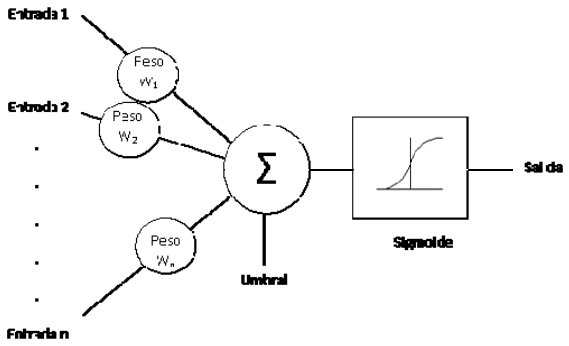

Figura 5. Esquema del perceptrón, neurona unitaria de una red.

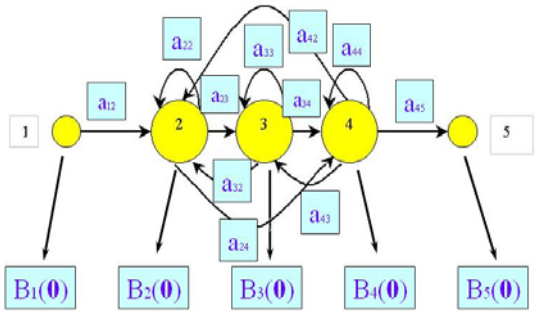

Figura 6. Esquema de operación de un HMM de 5 estados.

4.3 Modelos Ocultos de Markov (HMM)

Los modelos ocultos de Markov (HMM) corresponden a modelos estocásticos que han sido usados con éxito en el ámbito de reconocimiento de hablante dependiente de texto. Cada palabra de un locutor determinado es generada por un modelo de Markov, el que consiste en una serie finita de estados interconectados por probabilidades de transición. Cada uno de los vectores de características tiene cierta probabilidad de mantenerse en el estado actual o avanzar al siguiente. Por su parte, cada uno de los estados tiene una probabilidad (o densidad de probabilidad) de presenciar una cierta observación, es decir, observar un vector de características. En este caso, solamente son visibles las observaciones, desconociéndose la secuencia de estados.

En la fase de entrenamiento, se genera el modelo de cada hablante que corresponde a (3), tal que se maximice la probabilidad de la observación dado el modelo

$$\lambda_j = (A, B, \pi), \tag{3}$$

donde $A = \{a_{ij}\}$ es la matriz de probabilidades de transición, $B = b_j(k)$ es la matriz de probabilidades de la observación y π es la probabilidad que cada estado sea el primero (Fig. 6). El cálculo de las matrices A, B y π se realiza a través del algoritmo Baum-Welch o métodos de gradiente, pero no se obtendrá el óptimo, sino máximos locales.

En el test, el problema de matching o similitud entre una voz desconocida y una clase conocida puede ser formulada por la distancia entre una observación y un modelo de hablante conocido. La observación corresponde a un vector de características, es decir, un vector de coeficientes cepstrales y el modelo del hablante conocido es el modelo oculto de Markov. La distancia antes citada corresponde a una densidad de probabilidad dada por

$$P(S_i = S_j/O, \lambda_j), \tag{4}$$

donde S_i es el hablante desconocido, S_j el hablante conocido, O es el vector de observaciones, siendo cada una de ellas una serie de coeficientes cepstrales y λ es el modelo del hablante conocido.

4.4 Modelo de Mezclas Gaussianas (GMM)

El modelo de mezclas gaussianas (GMM) fue introducido por primera vez para reconocimiento de hablante por Reynolds en "Speaker identification and verification using Gaussian mixture speaker models" en 1995 y es el algoritmo estadístico de clasificación más exitoso en la actualidad. El conjunto de vectores de coeficientes cepstrales de un hablante

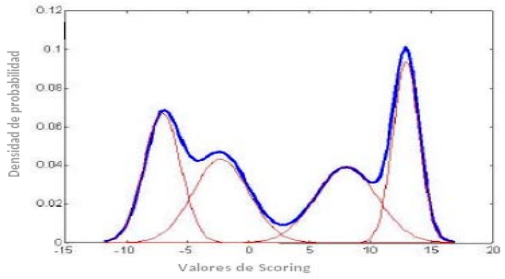

Figura 7. Modelamiento de 4 gaussianas sobre una distribución (Batvox training)

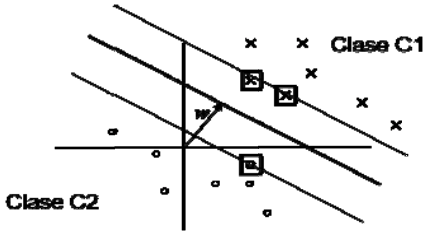

Figura 8. Hiperplano separador óptimo. Los vectores que se observan en rectángulo son los de soporte.

distribución hiper dimensional, la cual es modelada por múltiples gaussianas, tal y como se muestra a manera de ejemplo en la Figura 7. Típicamente en reconocimiento de hablante, se utilizan entre 512 y 1024 mezclas para una óptima caracterización. Diversos algoritmos se utilizan para generar estas gaussianas, destacándose Expectation-Maximitation (EM) y Maximun a posteriori probality (MAP). En ellos, iterativamente se refina la ubicación, anchura y máximo de las gaussianas hasta un umbral de convergencia. Entonces, en la fase de entrenamiento de una voz conocida se genera un modelo basado en sus propias mezclas de gaussianas. Luego en la fase de test, se evalúa cada vector de características en la distribución del modelo. En suma, el scoring final será la suma de todas estas evaluaciones. A mayor cercanía entre los vectores cepstrales de la voz desconocida y el modelo, la evaluación en la distribución caerá en los valores más altos de las gaussianas respectivas, por lo tanto su scoring será más alto. Algunas sesiones de entrenamiento bastará para determinar el óptimo valor de scoring umbral de la discriminación.

4.5 Vectores de Soporte (SVM)

El método de vectores de soporte es un método general para la resolución de problemas de clasificación, regresión y estimación, se basa en la teoría estadística del aprendizaje. El principal objetivo es transformar los vectores de entrada, en este caso los vectores de características n-dimensionales en vectores de dimensión más alta en los que el problema puede solucionarse linealmente. Esta frontera de decisión lineal corresponde a un hiper plano óptimo que separa solamente dos clases, por lo que es una decisión binaria. Por lo anterior, en el reconocimiento de hablantes este método se utiliza en verificación de locutor. El hiper plano elegido es el que mejor separa las clases, esto es, que maximiza la distancia euclídea a los vectores de características más cercanos de cada clase, por eso se denominan clasificadores de margen máximo. Estos vectores corresponden a los vectores de soporte. Para ello se construyen 2 planos paralelos, que pasan por los vectores de soporte, según se observa en la figura 8. Cuando las clases no son linealmente separables, se busca el hiper plano que minimice los errores a través de una función de costo. La extensión a límites no lineales se realiza a través de un núcleo o kernel que satisfaga las condiciones de Mercer. En esencia, cada vector de características es mapeado a un espacio de alta dimensionalidad en que los datos son separables linealmente. El producto escalar de los vectores transformados, se puede escribir como el núcleo, por lo que no es necesario trabajar en el espacio extendido. Al igual que las redes neuronales, SVM trabaja en bloque, por lo que un cambio en los datos supone la obtención de una nueva máquina. La etapa de test verifica a qué lado de la línea de separación están los vectores de características de la voz desconocida, siendo un proceso de discriminación. A un lado están los vectores que caracterizan al individuo y al otro el resto de la población.

5 Interpretación de Resultados para Análisis Forense: Regla Bayesiana

La etapa final dentro del reconocimiento de patrones es el test, donde un ejemplar desconocido es clasificado dentro de las clases entrenadas. En el caso de reconocimiento de locutor, dado un segmento de voz desconocido, el objetivo final es evaluar si corresponde a las voces previamente entrenadas. Esto se realiza a través de una puntuación o scoring de forma discriminativa entre las clases conocidas, es decir, dado un umbral pre definido, si la evaluación de la voz desconocida supera ese umbral, entonces ésta corresponde a la clase comparada. Este proceso se utiliza, por ejemplo, en sistemas de verificación de locutor para control de acceso. Sin embargo, este proceso discriminativo no es adecuado para una inferencia forense, puesto que el resultado no puede estar determinado por un umbral de identificación, más bien por una probabilidad de reconocimiento. En los siguientes puntos se explicitará la manera de obtener un valor de scoring y cómo transformar ese valor en una probabilidad, a través de una interpretación bayesiana

5.1 Scoring

La puntuación o scoring se define como el valor que entrega la evaluación de características vocales de una voz desconocida dentro de las clases conocidas y que han sido entrenadas. Como ejemplo, la figura 9 muestra el test entre características de una clase desconocida y dos clases conocidas entrenadas con el método de Cuantización Vectorial o VQ.

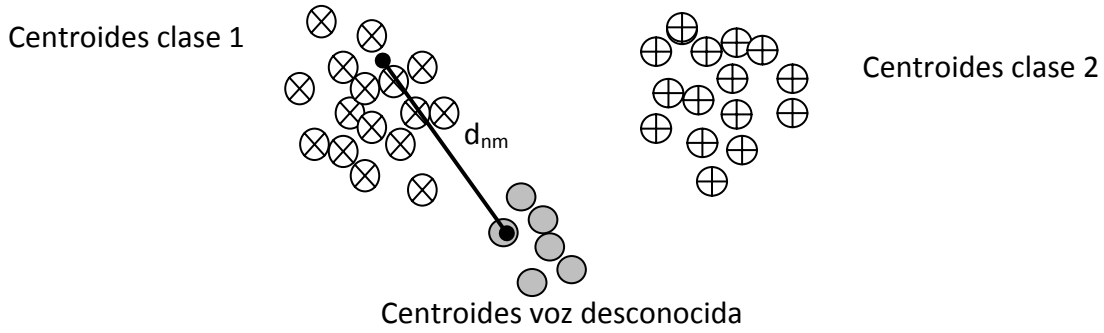

Figura 9. Esquema de cálculo de puntuación para dos clases entrenadas y una clase desconocida. Las figuras corresponden a los codebook de cada clase obtenidos con la técnica de cuantización vectorial o VQ, mientras que d_{nm} corresponde a la distancia euclídea entre el centroide desconocido n y el centroide N°m de la clase 1.

Siguiendo el ejemplo de la figura 9, en la etapa test se calculan todas las distancias desde los centroides desconocidos a los centroides de las respectivas clases, almacenando la suma de estas distancias para cada una de ellas. Este valor será el scoring. En el caso closed set, la clasificación de la clase desconocida será aquella cuyo scoring sea menor, es decir, la suma de todas sus distancias es más pequeña, lo que se traduce en una cercanía entre ambas clases. Sin embargo, en el caso open set, se debe establecer un umbral de decisión para asociar las características vocales desconocidas con alguna de las clases entrenadas. Si la distancia acumulada para todas las clases es mayor que ese umbral, significa que la voz desconocida no está lo suficientemente cerca de alguna de las clases, por lo que no hay identificación positiva. Para establecer ese umbral, se utilizan muchos ejemplares conocidos que pertenezcan a las clases entrenadas, que progresivamente van refinando el valor umbral.

5.2 Regla Bayesiana

El análisis de voz para fines forenses ha ido en aumento conforme la acreditación de muchos delitos es a través de interceptaciones telefónicas o grabaciones de audio en general. En este contexto, la voz desconocida es la grabación *dubitada*, mientras que las clases conocidas son los sospechosos de ser la voz registrada en la grabación. A estas voces se le denominan *voces indubitadas*. Los valores de scoring obtenidos por alguna técnica, ya sea dentro de la estructura de verificación o identificación de locutor, se basan en umbrales subjetivos. Llevado a esto a un enfoque forense, este valor fuerza al experto a tomar decisiones binarias, es decir, *es* o *no es* la voz desconocida. Lo anterior es una usurpación de la labor que debe desarrollar la corte, puesto que son ellos y solamente ellos los llamados a establecer esta decisión más allá de toda duda razonable. Entonces, diversos autores han propuesto el teorema de Bayes y la metodología de manejo de los datos disponibles, para llevar a cabo una inferencia probabilística útil para ser usado en los juicios. Cabe consignar que esta técnica tiene un trabajo de desarrollo extensivo en otros campos forenses, siendo muy útil para interpretar un análisis de perfil de ADN, por ejemplo.

Las probabilidades expresados en la forma del teorema de Bayes permite relacionar cómo nuevos datos, en este caso la grabación dubitada, puede ser combinado con información previa (probabilidades a priori) para entregar un probabilidad posterior. En otras palabras, el valor de la evidencia estudiada será en cuánto modifica, a favor o en contra, la hipótesis de culpabilidad o inocencia, dado un conocimiento anterior es

$$\frac{P(H_p/E,I)}{P(H_d/E,I)} = \frac{P(E/H_p,I)}{P(E/H_d,I)} \times \frac{P(H_p/I)}{P(H_d/I)}, \tag{5}$$

donde P(x/y) es la probabilidad condicional, H_p es la hipótesis del prosecutor, es decir, "el sospechoso es quien aparece en la grabación dubitada", H_d es la hipótesis de la defensa, "otra persona, dentro de una población relevante de individuos, es quien aparece en la grabación dubitada, I representa la información adicional no relativa a la evidencia E, por ejemplo, el número de teléfono de la interceptación, la relación del sospechoso con el teléfono de la víctima, etc. En (5), se observa cómo la valoración de la evidencia que provee el experto forense modifica las probabilidades a priori para entregar una probabilidad modificada.

Así, basados en el teorema de Bayes, se puede evaluar la razón de similitud de la evidencia E es

$$LR = \frac{P(E/H_p,I)}{P(E/H_d,I)}. \tag{6}$$

En el caso de una grabación dubitada analizada a través de un reconocimiento automático de locutor, estas probabilidades son funciones de densidad de probabilidad, asumidas comúnmente como gaussianas. La importancia de la ecuación (6) radica en el hecho que no solamente se evalúa la probabilidad que la voz desconocida sea del individuo (numerador), sino que además se incorpora la similitud con una población relevante de individuos (denominador). Así, por ejemplo, en el caso que una voz dubitada comparta características vocales de un individuo, se evalúa también si comparte dichas características con una población de otros individuos relevantes. Si es así, entonces la voz dubitada se parece tanto al individuo como a la población representativa, siendo entonces su valor de LR cercano a 1, es decir, no se modifica ni a favor ni en contra la información a priori.

Para implementar de manera práctica el framework bayesiano se requieren los siguientes set de datos: *Población de referencia* usado para modelar la voz y da cuenta de la variabilidad acústica de las características vocales; *Modelo de voz de interés* de mezclas

gaussianas obtenido de la voz conocida o indubitada; *Voz dubitada* o desconocida; *Tramos de control* de la voz conocida que se comparará con el modelo de voz de la misma. Con la población de referencia se generan modelos que se comparan con la voz desconocida obteniéndose una función de densidad de scoring llamada *intervariabilidad*. A su vez, los tramos de control de la voz indubitada se comparan con el modelo de la misma voz, para encontrar la función de densidad denominada *intravariabilidad*. Entonces, la razón de similitud LR se obtendrá como el cociente entre el valor de intravariabilidad y el de intervariabilidad, evaluado en el scoring que entrega la voz dubitada con el modelo conocido. Varios estudios han refinado la estimación de las funciones de densidad y mejorado la excesiva coherencia de la intravariabilidad, puesto que los tramos de control se extraen del tipo monosesión.

Diversas técnicas se han implementado para refinar y optimizar el cálculo del factor LR, haciendo que el análisis forense sea más transparente y testeable, elementos claves para la rigurosa cuantificación del peso de la evidencia. Los esfuerzos en este sentido se han enfocado en estimar la calibración del framework bayesiano, garantizando la presunción de inocencia con la incorporación de la técnica TDLRA (*Target Dependent Likelihood Ratio Alignment*), como un esquema de normalización que minimice el número de sospechosos no autores de la grabación dubitada cuyo valor de LR > 1, es decir, minimizando las falsas aceptaciones. Además, se ha trabajado en la mejor modelación de las distribuciones de probabilidad, en especial la relacionada con la intravariabilidad, incorporando correcciones para compensar el efecto monosesión en su cálculo.

6 Caso de éxito: la herramienta Batvox

El año 2006 la Policía de Investigaciones de Chile, en su Sección Sonido y Audiovisual del Laboratorio de Criminalística, incorporó en sus peritajes el reconocimiento de hablantes con la herramienta biométrica Batvox, desarrollado tempranamente por la Universidad Politécnica de Madrid y que en la actualidad es desarrollada y comercializada por la firma de reconocido prestigio Agnitio. Los desarrolladores de Agnitio han sabido plasmar en una herramienta usable las principales tendencias y mejoras existentes en el reconocimiento de locutor a nivel mundial, implementando la metodología bayesiana para proporcionar los resultados en la forma que demanda el análisis pericial, en forma probabilística, con pasos repetibles, transparentes y testeables, es decir, con un soporte científico. La experiencia particular en el uso intensivo de esta herramienta presenta una razonable performance, dada la gran variedad de situaciones presentes en el ámbito forense, siendo este enfoque, por lejos, el que presenta las mayores dificultades para la capacidad de discriminación de un reconocedor biométrico o automático.

Por cierto que existen elementos a subsanar en orden a elevar aún más la objetividad de los resultados, en especial la dependencia de los LR sobre la extensión y calidad de la grabación, mejor separación de las clases, estudios más profundos sobre el desempeño frente a señales degradadas, entre otros, pero se ha llegado a un nivel de robustez suficiente como para realizar estudios seguros y confiables con esta herramienta. En la actualidad se llevan a cabo estudios para incorporar nuevas técnicas que maximicen la discriminación y disminuyan el tiempo de cálculo. Se destaca la problemática del ruido aditivo que afecta la señal producto del canal de transmisión, que inicialmente se modelaba a priori de manera discreta, separándose para cada tipo de canal y que ahora se estudia el uso de un espacio continuo (*channel space*). La figura 10 entrega el resultado de un análisis real realizado con el software Batvox, que ilustra el framework bayesiano presentado.

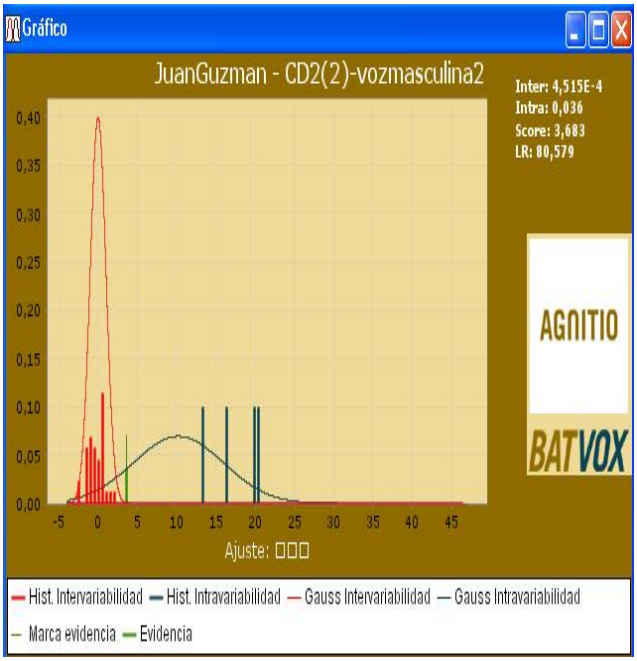

Figura 10. Gráfico de resultados del sistema Batvox para un caso real.

La grabación dubitada corresponde a un registro telefónico interceptado de forma digital, formato MP3 a 8000 Hz de muestreo. Por su parte, la voz del sospechoso se tomó con un sistema digital, microfónico, grabando a formato PCM WAV con 8000 Hz de muestreo. La población de referencia posee 98 individuos registrados con los mismos atributos que la grabación del sospechoso. La similar característica de grabación entre el sospechoso y la población de referencia es fundamental para el correcto cálculo de LR. En la Figura 10 se observan claramente la construcción de las distribuciones de probabilidad gaussianas desprendidas de los set de datos y la valoración de la evidencia dado por un scoring y una razón de similitud LR evaluado en las respectivas curvas.

Progresivamente los laboratorios forense de análisis de voz comienzan a incorporar este enfoque en su desarrollo pericial, que inicialmente estaba reservado para los métodos fonéticos o lingüísticos. A diferencia de ellos, en la Policía de Investigaciones de Chile se incorporó inmediatamente como método exclusivo de reconocimiento de hablantes, situación que debe ser complementada con las aproximaciones tradicionales, en orden a generar una aproximación más holística de esta problemática.

7 Agradecimientos

Este trabajo ha sido financiado parcialmente por la Universidad de La Frontera proyecto DIUFRO DI08-0015.

Referencias

D. Reynold, T. Quatieri, R. Dunn (2000). “Sepaker Verification Using Adapted Gaussian Mixture Models”. Digital Signal Processing, (10), N. 1-3, 19-41.

F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau,S: Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, D. Reynold, (2004). “A Tutorial on Text Independient Speaker Verification”. EURASIP Journal on Applied Signal Processing, (4), 430-451.

K. Farrel, R. Mammone, A. Richard; K. Assalch (1994). “Speaker Recognition Using Neural Network and Conventional Classifiers”. IEEE Tran. on Speech and Audio Processing, (2), No 1, Parte II.

J. Campbell (1997). “Speaker Recognition: A Tutorial”. Proc. of the IEEE, (85), No 9, 1437-1462.

C. Burges (1998). “A Tutorial on Support Vector Machinas for Pattern Recognition”. DM(2) 121-167.

B. Kenny (2006). “Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms”.

J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. Toledano, J. Ortega-García (2007). “Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition”. IEEE Tran. on Audio, Speech and Language processing, (15), No 7, 2104-2114.

J. Gonzalez-Rodriguez, S. Drygajlo, D. Ramos-Castro, Daniel, M. Garcia-Gomar, J. Ortega-García, (2005). “Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition”. Computer Speech and Language, (20), 331-355.

L. Rabiner, B. H. Juang (1997). “Fundamentals of Speech Recognition”. Prentice Hall, NY, USA.