Comentarios sobre el contexto actual de la identificación forense de locutores.

Dr. Carlos Delgado Romero

Comisaría General de Policía Científica, Dirección General de la Policía, Madrid, España.
carlos.delgado@dgp.mir.es

Abstract.-

En consonancia a la rápida modernización científica y tecnológica de las últimas décadas, las diferentes disciplinas forenses en general, y la identificación de locutores en particular, tratan de adaptar sus herramientas, procedimientos y referencias de análisis a las necesidades procesales de las sociedades de vanguardia. Este compromiso de adecuación es un camino en el que confluyen múltiples elementos de dificultad. A lo largo del presente artículo realizaremos un repaso sobre dichos aspectos, incidiendo en tres marcos fundamentales: la ciencia forense ante los tribunales de justicia, las actuales aportaciones del reconocimiento automático y la expresión de conclusiones en la emisión de informes periciales para la Justicia.

PARTE I
Los tribunales de Justicia y el científico policial en España.

Introducción.-

Quienes no están familiarizados con las técnicas científicas de investigación forense y, concretamente, con aquellas desarrolladas por los expertos de las agencias policiales, probablemente desconocen los importantes inconvenientes a los que dichos expertos han de enfrentarse en muy diversos aspectos de su trabajo cotidiano. Para empezar, la propia obtención de una formación específica y la consiguiente cualificación en un área de experiencia determinada supone, ya desde el principio, una carrera de obstáculos. En muchos casos, la previa disposición de una base teórica afín a la técnica forense practicada -habitualmente formación universitaria- representa y procura un importante apoyo de autoridad. Así ocurre en el caso de la acústica, los análisis químicos y biológicos, los estudios antropológicos y entomológicos, etc . No obstante, el aprendizaje y ejecución de cada técnica en particular conlleva una etapa de educación adicional, a la vez que un importante compromiso de responsabilidad por parte del experto. En algunas circunstancias, cuando diferentes disciplinas teóricas confluyen en la vertebración científica de una misma técnica, el desarrollo de ciertos procesos de entrenamiento adquiere un alto grado de complejidad (así ocurre por ejemplo, en la identificación de locutores) .

Pero los problemas del científico policial no se reducen a los citados aspectos. En una u otra medida, su actividad profesional siempre va acompañada de otros elementos de dificultad: análisis sobre muestras degradadas, urgencia en el tiempo de respuesta, falta de recursos tecnológicos, imposibilidad de ampliar y vincular sus tareas habituales a otras más específicas de investigación y desarrollo, etc

Los análisis científicos como medio de prueba.-

Las unidades científicas de investigación policial juegan un papel fundamental en multitud de tareas de apoyo a otras desarrolladas por sus colegas operativos. No obstante, en muchas otras ocasiones su trabajo -materializado en informes periciales- está orientado al auxilio técnico de las diferentes autoridades judiciales. En la gran mayoría de estos casos, el experto es llamado a declarar en una vista oral para explicar en detalle las conclusiones y procedimientos relacionados con su estudio pericial. Y es, principalmente en este acto, donde sus análisis de laboratorio adquieren una trascendencia real de cara a la Justicia.

Desde la perspectiva del científico forense se observa con cierta extrañeza la notoriedad que los órganos jurisdiccionales conceden a la declaración testifical del experto. Nuestro ordenamiento jurídico -Constitución, Ley Orgánica del Poder Judicial, Ley de Enjuiciamiento Criminal- establece y ensalza, por encima del propio valor de prueba que en sí mismos representan los informes de expertos, la importancia del testimonio oral de los peritos. La forma en que estos expresan verbalmente sus conclusiones para una mejor apreciación por parte del tribunal o jurado, se presenta como algo determinante. A primera vista puede parecer evidente la pertinencia de este proceder, pues los dominios técnicos en los que discurren muchos medios probatorios documentales son inaccesibles al entendimiento de aquellos no relacionados con tales entornos. Resulta muy complicado para un juez o miembro de un jurado llegar a conformar un grado de convicción respecto de unos resultados de análisis, cuando éstos vienen expresados en una nomenclatura o unos términos deductivos que les son total o parcialmente desconocidos. Desgraciadamente, las personas en las que descansa la responsabilidad de impartir Justicia no pueden dominar todos los campos del conocimiento científico vinculados a la amplia gama de elementos evidenciarios que pone a su disposición la Policía Científica. Por esta razón, y asumiendo de antemano que la labor testifical del experto ha de argumentarse en la necesidad de trasladar al *román paladino* los aspectos técnicos recogidos en sus informes, tampoco ha de ignorarse la componente negativa que este acto puede originar cuando deriva en una incorrecta lectura por parte de la Autoridad Judicial.

El planteamiento pretende ser claro. Para un científico, la aportación al procedimiento judicial de unas conclusiones o resultados a través del correspondiente informe técnico, ya constituye en sí mismo un elemento de valoración. O dicho con mayor precisión: "el" elemento de valoración. Es cierto, que en la conciencia del perito forense siempre ha de residir un permanente ánimo de interpretación que posibilite la mejor comprensión de aquellos matices específicos propios de su especialidad. No obstante, el hecho de apartarse de lo estrictamente técnico también puede generar malas consecuencias.

Por una parte, pudiéramos correr el riesgo de adentrarnos en un terreno que es competencia de las autoridades judiciales. El hecho de "interpretar" resultados ha de entenderse como "traducir" a un lenguaje entendible, y no como "deducir" conclusiones, ajenas a lo que es el entorno puramente científico. Por otro lado -en algunas ocasiones ocurre- se propicia el éxito de quienes se expresan más elocuentemente en la vista oral, a veces, en detrimento de aquellos otros que por una u otra razón carecen de tal habilidad; todo ello, con independencia de la disposición o no

de un criterio de experto fundamentado en el sólido conocimiento de su área de experiencia. Es decir, pudiera acontecer –de hecho acontece- que durante su testimonio oral, un excelente científico por timidez u otro motivo no supiese trasladar al tribunal o jurado, en una forma suficientemente entendible, los resultados de su estudio. Y de la misma manera, que como consecuencia de su locuacidad, un mal perito obtuviese una inmerecida credibilidad.

Ante supuestos de estas características, ¿hasta qué punto podría demandarse la correspondiente responsabilidad de nuestras instituciones de justicia?.

Carecería de sentido, exigir de los profesionales que integran la administración de justicia, un suficiente nivel de conocimiento sobre el conjunto de disciplinas y técnicas utilizadas por los distintos expertos en sus tareas de apoyo a la investigación judicial. Pero además, la coyuntura aquí planteada se prolonga más allá de lo que es la mera interpretación de los estudios periciales. Existen otra serie de interrogantes que encuentran perfecta cabida dentro del mismo escenario:

¿Ante qué presupuestos de análisis resultan fiables las técnicas utilizadas?

¿Qué cualificación y experiencia profesional poseen los expertos que las practican?

¿Las bases científicas que sustentan dichas técnicas están convenientemente validadas?.

Desde la óptica de un científico, toda esta problemática cuando menos se manifiesta desconcertante. Es complicado comprender cómo la potestad de conferir la calidad de perito experto en un área concreta del saber, puede residir en alguien que es completamente ajeno a tal ámbito de conocimiento. Tradicionalmente, las instituciones judiciales depositan su confianza en los especialistas policiales, que dicho sea de paso, tanto por el hecho de poseer una dilatada experiencia profesional, como por el carácter absolutamente aséptico, sistemático y cotidiano de sus evaluaciones, parecen contar -a priori- con una excelente carta de presentación. No debemos olvidar, que nos encontramos en un entorno donde cualquier actuación debe cimentarse en los pilares del rigor y la responsabilidad ya que, en ciertas ocasiones, los resultados de un estudio científico pueden llegar a constituir un elemento de prueba fundamental. Por este motivo, y porque en definitiva estamos hablando de los derechos fundamentales de las personas, los tribunales de justicia han de asegurarse de estar siempre asesorados por un equipo de científicos con una clara conciencia de cuáles son los límites y referentes que circunscriben sus técnicas.

A diferencia de lo que ocurre en otros países, en España no disponemos de unos estándares de admisibilidad para la evidencia científica. Por eso, cuando surge la controversia en torno al grado de fiabilidad de una nueva alternativa o práctica de análisis, tanto los magistrados como los expertos se encuentran desorientados. Los unos, porque desconocen las prestaciones reales de esa nueva técnica o método. Y los otros, porque no saben a qué tipo de referencias han de atenerse a la hora de legitimar o no la utilización de esa nueva opción. De todos son conocidos ciertos estándares ya definidos para la evaluación de nuevas aportaciones científicas dentro del marco pericial judicial. En los Estados Unidos, el denominado *"Frye test"* o, más recientemente -a raíz del

conocido caso *"Daubert vs. Merrell D.Ph."* [1993]- la regla federal para la evidencia FRE/702, son dos buenos ejemplos en este sentido. Sin embargo, en nuestro país todo parece estar confiado a la discrecionalidad y sentido común de cada tribunal.

Si bien es cierto que la razón nunca puede ser considerada una mala compañera de viaje, también lo es, que lo que unos estimen como válido o racional, otros no lo contemplen de la misma manera. Desde luego, no sería la primera ocasión en que, ante la apreciación de evidencias científicas de idéntica índole, tribunales distintos obtienen diferentes niveles de convicción.

Aunque suene a utópico, se antoja ya imprescindible la presencia de una institución, que de forma similar a lo que ocurre con la "National Academy of Sciencies" norteamericana, proporcione al ámbito judicial el apoyo necesario en cada momento. En tanto ello acontece, sería muy conveniente que los responsables de los órganos judiciales catalizaran esa imprescindible adaptación de sus procedimientos a los imperativos evolutivos de la ciencia. De igual forma, los institutos de investigación forense han de esmerarse en otorgar el máximo rigor científico a sus protocolos y prácticas de análisis pericial.

Si nos situamos en un plano más cercano a lo que es nuestra realidad actual, eso sí, con la esperanza de que en un plazo no demasiado lejano alcancemos el deseable nivel de competencia, podríamos sugerir unas líneas inmediatas de actuación que debieran materializarse a tres niveles:

- en el ámbito de los tribunales de justicia sería muy importante una actualización de la normativa relacionada con la prueba pericial y elevar el nivel de exigencia en cuanto a la utilización de todos los medios posibles para acercarse al conocimiento de los trabajos científico-periciales (cualificación y experiencia de sus expertos, prestaciones y fundamentos de sus técnicas, etc)

- los responsables policiales deben seleccionar y cultivar cuidadosamente los perfiles académicos de sus expertos procurando respondan en todo momento a las necesidades que demanda cada área de especialidad. Por otra parte, deben plantearse como objetivos de prioridad, tanto el fomento de las actividades de formación, investigación y desarrollo de los expertos, como la divulgación continuada de las posibilidades y novedades técnicas de cara a los organismos judiciales. La continuidad de los expertos en su puesto de trabajo ha de ser contemplada como un capítulo fundamental.

- los propios expertos han de adaptar su conocimiento y práctica profesional al ritmo de evolución marcado por la comunidad científica de su entorno. En este sentido, el intercambio científico y la normalización de alternativas o métodos de trabajo, se revelan como herramientas idóneas para la consecución de tal fin.

PARTE II
La aportación de los sistemas de reconocimiento automático.

La identificación de locutores, instrumento de investigación forense.

El trayecto que una nueva técnica o método ha de recorrer desde su introducción en el entorno científico forense hasta el momento de su consolidación, es un camino delicado, laborioso y lleno de obstáculos. Generalmente, los laboratorios policiales juegan un papel fundamental a la hora de explorar la viabilidad y eficacia de las nuevas opciones tecnológicas que los distintos campos del conocimiento ponen a disposición de la investigación judicial.

Como ya es conocido, en el caso de la identificación forense de locutores, (I.F.L.) los primeros pasos fueron especialmente dificultosos. Sirvan como referencia los antecedentes históricos acontecidos en Estados Unidos, los cuales pueden considerarse pioneros, a la vez que un válido exponente de la controversia que caracterizó el desarrollo de esta técnica en sus primeros pasos. Dentro de este contexto concreto, la falta de rigor por parte de algunos expertos, junto a la existencia de distintos enfoques de estudio, podrían citarse como principales substratos desencadenantes de tal situación. De hecho, todavía en la actualidad persisten ciertas reticencias entre expertos como consecuencia de esta problemática inicial. [Delgado, 1991]

Sin embargo, al margen de determinadas circunstancias puntuales, la I.F.L. ha de considerarse como una técnica plenamente consolidada. Su práctica sistemática está extendida por todos los laboratorios forenses de vanguardia y la discusión científica en torno a la misma se centra ahora en alcanzar un consenso sobre qué protocolos metodológicos se adecuan mejor a cada una de las distintas alternativas de análisis existentes.

Hoy en día, un investigador forense que se precie de conocer el estado de la cuestión, no puede plantearse si es, o no es posible, identificar a una persona a través de su voz. Sin necesidad de ser un experto, cualquiera de nosotros es capaz de reconocer la voz de un familiar o de una persona conocida, incluso a través del teléfono. Por otra parte, también es indiscutible que el habla, referencia biométrica de comportamiento sujeta a diferentes factores de variabilidad (producción articulatoria y fonatoria, componentes emocionales, expresivos, retóricos, etc) se revela como uno de los retos de investigación forense de mayor complejidad. A ello, no sólo contribuye el carácter multidisciplinar de las distintas perspectivas de análisis que se proyectan sobre nuestro objeto de estudio: ingeniería y física acústica, fonética, lingüística, patologías del habla, percepción, etc., sino también, las condiciones degradadas que habitualmente caracterizan las muestras de análisis utilizadas en nuestro entorno: grabaciones de transferencia telefónica con diferentes tipos de ruido, distorsión...

En la actualidad, las metodologías forenses más practicadas por los laboratorios policiales son las denominadas "combinadas". Bajo este concepto general se agrupan aquellas técnicas que de una u otra forma sustentan sus fundamentos de estudio en tres perspectivas: acústica, fonético-lingüística y auditivo-perceptiva. Dichas técnicas, pueden complementarse o desarrollarse a través de sistemas semiautomáticos de cálculo o análisis. En los últimos años, la eficacia de ciertas aplicaciones de reconocimiento automático hace vislumbrar un futuro esperanzador en cuanto a su utilización con carácter exclusivo. [Delgado, 1991].

La problemática del reconocimiento automático.

Algunos laboratorios forenses están incorporando sistemas automáticos de reconocimiento de locutores (SARL) para desarrollar tareas de identificación (un candidato vs una población) y verificación (un candidato vs un sospechoso) . En ambos casos, el sistema necesita contar con una población de referencia o UBM (Universal Background Model) para establecer las distancias de similitud entre los modelos de voz contenidos en dicha población y las muestras de los candidatos que se le presentan. Es decir, aun en el caso de una tarea de verificación, el ratio de similitud entre la muestra "dubitada" y la "indubitada" siempre se calcula en referencia al resto de modelos de voz existentes en la base de datos poblacional. Precisamente, la necesidad de disponer de una base de datos suficientemente representativa, es uno de los inconvenientes a considerar, especialmente a la hora de interpretar los resultados comparativos obtenidos por el sistema. De ello hablaremos más adelante.

Por otro lado, en lo relativo a la mera construcción y funcionalidad de la aplicación, hemos de advertir que la generación de un modelo de voz que caracterice fielmente los distintos actos de habla de un locutor, es una labor complicada. No sólo en cuanto al hecho de alcanzar una heterogeneidad en el plano lingüístico, emocional, expresivo, articulatorio, etc sino también en lo que afecta a las propias características técnicas de las grabaciones utilizadas. La casi totalidad de voces dubitadas manejadas en el entorno forense provienen de interceptaciones de telefonía móvil o de línea terrestre, que a su vez son registradas en diferentes equipos y soportes de grabación. La unión de esta circunstancia, a la de la frecuente presencia de otros factores de degradación de la señal (ruidos, distorsiones, solapamientos de voz, etc) supone un serio obstáculo en el rendimiento óptimo de los SARL.

Los informes NIST

Desde 1996 el "Speech Group" del Instituto Nacional de Estándares y Tecnologías de los Estados Unidos (NIST), realiza evaluaciones anuales sobre los progresos de los SARL a nivel internacional [Przybocki, M. y Martín A. 1998]; [Martín, A. y Przybocki, M., 2002] . Para ello, diseña una serie de tests que tratan de verificar el rendimiento de dichos sistemas, tomando como punto de partida cuatro ejes de referencia: el tipo de entrenamiento, la duración de los segmentos-muestra, edad/sexo de los locutores y la influencia del "factor canal".

Es destacable la evaluación NIST-1998 donde se describen y analizan las características y resultados de un test de reconocimiento automático, independiente de texto [Doddington, G. et al, 2000]. La estructura canónica del mismo define tres marcos de actuación. El primero se refiere a las tareas de procesado de señal relacionadas con la extracción de parámetros y las técnicas de normalización de canal utilizadas. La información espectral procesada debe limitarse al rango de frecuencia comprendido entre 300Hz y 3.400Hz (banda telefónica) . En la fase de modelado o entrenamiento se establece una dicotomía general entre modelos de representación acústica supervisada y no supervisada, (caso de los GMMs ó Gaussian Mixture Models). Por último, se exponen las técnicas de normalización de "scores" para compensación de resultados ante la influencia de determinados factores críticos.

En líneas generales, salvando las buenas prestaciones de algunos de los sistemas de fusión que integran distintas opciones o procedimientos-base de las diferentes aplicaciones participantes en la evaluación, los SARL basados en modelado por mezclas de gaussianas son considerados los competidores más funcionales, debido a su consistencia y reducido coste computacional. Además, como principales capítulos que perturban la eficacia de los sistemas de reconocimiento, se relacionan los siguientes:

- influencia de la utilización de distinto canal de transmisión telefónico, especialmente vinculada al tipo de micrófono incorporado a cada terminal.
- la duración temporal de los segmentos test (voces dubitadas)
- el número de sesiones de entrenamiento utilizadas para obtener los modelos de la UBM (Factor multisesión).
- sensibles fluctuaciones entre modelos y segmentos test de parámetros no espectrales (caso del pitch).

De igual forma puede deducirse, que el rendimiento de un SARL es superior :

- a mayor número de sesiones de entrenamiento.
- a mayor duración de los segmentos test. Si bien no existe una relación lineal a este respecto, pues alcanzada una duración determinada la eficacia del sistema no evoluciona.
- utilizando el mismo canal y terminal telefónico.
- utilizando en los terminales micrófonos tipo "electrect"
- reconociendo voces de varones con F_0 grave.
- reconociendo voces de mujeres con F_0 aguda.

A pesar de ser admitido y bien conocido el negativo efecto que el factor ruido ejerce sobre la "robustez" de un SARL, el test evaluado en NIST-1998 no incide en detalle sobre el citado aspecto. Sí es cierto, que se etiquetan subjetivamente algunas de las muestras en tres niveles de calidad, en orden a la mayor o menor presencia de ruido (buena, mala y muy mala) aunque no se hace una mención expresa de los tipos de ruido ni del nivel de los mismos, en valores

SNR. No obstante, sí se subraya la necesidad de ampliar los objetivos de investigación en esta línea de trabajo.

Aunque en el ámbito forense factores como el ruido o la distorsión representan el pan de cada día, hemos de admitir que resulta muy complicado conjugar todos los elementos de dificultad que en una u otra forma afectan el buen funcionamiento de los SARL, sobre todo, teniendo en cuenta que muchos de ellos dependen directamente del comportamiento y características fono-articulatorias del hablante (emociones, patologías, ratios de intensidad y velocidad de elocución, etc).

Evaluaciones más recientes -NIST 2000 y 2001- incorporan como principal novedad nuevas bases de datos que incluyen habla conversacional por teléfonos móviles. Si bien los progresos informados no han sido relevantes, sí se intuye una nueva vía de trabajo, que combinada con los prototipos de SARL más competitivos, puede ofrecer una sensible mejora de su rendimiento. Nos estamos refiriendo a los recientes estudios desarrollados por G. Doddington, en los que se ha detectado la gran importancia de ciertas informaciones de caracterización temporal de la señal. Tradicionalmente, los esfuerzos de investigación y desarrollo de las tecnologías de reconocimiento automático de locutores, se han centrado en el análisis de la información espectral de bajo nivel. Tomando como base este tipo de referencias de análisis, los últimos resultados proporcionados por el NIST ponen de manifiesto un estancamiento de las mejoras sensibles de rendimiento. Sin embargo, Doddington reflexiona sobre el notable peso identificativo que por sí mismas, y como complemento a los parámetros clásicos de caracterización automática representan, las que denomina características idiolectales. Partiendo del análisis de simples transcripciones, propone la utilización de tramos a largo plazo (palabras o frases) y estructuras suprasegmentales asociadas a dichos tramos: rasgos prosódicos, énfasis, ratio elocutivo, etc. La eficacia de esta nueva, aunque simple y lógica perspectiva, ha sido ya experimentada [G. Doddington, 2000] y se revela como una herramienta de modelado prioritaria en las próximas evaluaciones del NIST.

No deja de ser evidente, pero a la vez curioso, el hecho de que “a estas alturas” haya que acudir a los objetos y mecanismos de destreza propios de los procesos perceptivos para la discriminación de voces familiares.

Nuevos proyectos-test. El FASR del F.B.I.

Recientemente, en la misma línea de investigación referida, aunque con una finalidad específicamente forense, el Instituto Forense de Holanda (N.F.I.) y el T.N.O. (organización para la investigación científica aplicada de Holanda) han efectuado un plan de evaluación conjunto con el que pretenden explorar la aplicación de sistemas de reconocimiento automático en nuestro entorno de trabajo [Leeuwen D. and Bouten, J., 2003] . La principal novedad que aporta esta nueva propuesta, es la utilización de registros de interceptaciones telefónicas policiales reales como material de test. Esperan presentar sus primeros resultados en la próxima reunión del Speaker Odyssey, 2004. [1].

Como complemento ilustrativo a esta panorámica general sobre los SARL, haremos una última incursión en un interesante sistema de reconocimiento automático, específicamente diseñado y testado para su aplicación forense. El denominado FASR, (Forensic Automatic Speaker Recognition program) es la aplicación elegida por el F.B.I. para explorar nuevas alternativas de análisis en sus tareas de identificación de registros de habla. Hasta el momento presente, la agencia federal norteamericana viene utilizando el método "auditivo-espectrográfico" a partir de muestras dependientes de texto y con similares características de registro. No obstante, consideran que el desarrollo de los SARL ha alcanzado la suficiente madurez como para ser tenidos en cuenta de cara a su posible utilización en el apoyo a la investigación de sus unidades operativas.

El FASR fue desarrollado entre 1998 y 1999 tras ser sometidos a test doce sistemas-candidatos seleccionados por el departamento federal. Algunos de estos sistemas, participaron en el concurso NIST-1998, anteriormente comentado. Básicamente, el sistema se soporta en una estación de que posibilita la ejecución de diversas funciones: conversiones A/D; D/A, distintas representaciones gráficas de la señal (incluidos sonogramas), así como segmentación y etiquetado manual o automático de la misma. También puede detectar y filtrar tonos de interferencia, o determinar y seleccionar -mediante valores SNR o de ancho de banda- niveles cualitativos o cuantitativos de la señal. El programa puede efectuar tanto tareas de identificación como de verificación, apoyándose en tres bases de datos que contienen los archivos test, modelos y poblaciones de referencia. El algoritmo de reconocimiento se sustenta en un robusto clasificador GMM que, esencialmente, considera parámetros psico-acústicos MFCC y compensa el efecto canal mediante normalizaciones CMN ó RASTA. [Nakasone, H. y Beck, S. , 2001].

Los sistemas-candidatos fueron evaluados contra la base de datos FV1, desarrollada como parte del proyecto CAVIS durante el periodo 1985-89. La FV1, es una base de datos integrada por tres colecciones de registros de voz, de veinticuatro, veintisiete y cincuenta locutores distintos, respectivamente. Contempla cuatro variables fundamentales e imprescindibles para caracterizar un entorno de comunicación forense:

- tipo de emisión hablada (espontánea, lectura, repetición)
- tipo de canal de transmisión (micrófono, teléfonos, transmisores de RF)
- diferentes duraciones de las muestras
- factor multisesión (diferentes tomas a lo largo del tiempo)

Los registros utilizados están referenciados en sus correspondientes duraciones, formatos de muestreo y valores SNR. Los diferentes tests a los que se sometieron los sistemas -de identificación cerrada y verificación abierta- combinaban las mencionadas variables, estableciendo cuatro criterios generales de dificultad:

- NIVEL I : Independencia de texto + independencia de canal
- NIVEL II : Dependencia de texto + independencia de canal

- NIVEL III : Independencia de texto + dependencia de canal
- NIVEL IV : Dependencia de texto + dependencia de canal

Lógicamente, el nivel I era el de mayor dificultad y el IV el de menor. En cada uno de estos cuatro niveles se ubicaron doce pruebas, por lo que al final se generaron cuarenta y ocho tests independientes. Los resultados de los ensayos de verificación abierta fueron ploteados mediante curvas DET (detección error trade-off) y se tabularon mediante valores de la tasa de error EER (equal error rate) y coeficientes Neyman-Pearson del ratio de falsa aceptación sobre una tasa fija del 10% de falso rechazo, y del ratio de falso rechazo sobre una tasa fija del 10% de falsa aceptación .

En cuanto a las pruebas de identificación sobre conjuntos cerrados, los resultados de rendimiento del sistema fueron evaluados sobre dos modalidades. Una de ellas (B), presentaba categorizados los tres candidatos que más puntuaban. La otra (A), ofrecía únicamente el mejor candidato. Como es lógico, los porcentajes de acierto eran superiores cuando los sistemas ofrecían un ranking de tres candidatos. Para el nivel de dificultad III casi todos los sistemas mostraban una alta eficacia (90-100%) cuando las muestras de entrenamiento y test eran de 30sg y habían sido registradas en similares condiciones de canal. Sin embargo, cuando se utilizaban muestras test de 3sg, el rendimiento de los sistemas decrecía de forma crítica, situándose en torno al 53% (tipo A) para el mejor reconocedor. En el nivel de dificultad I, modalidad (A), los porcentajes de acierto no superaron en el mejor de los casos el 95.3% de acierto, aunque el porcentaje medio para muestras test de duraciones iguales o superiores a los 30sg, oscilaba entre el 65 y el 85 %. Al igual que ocurría en el nivel III, ante fragmentos test de 3sg la eficacia media del mejor competidor descendía de forma notable (40%) .

Por lo tanto, podemos afirmar que, en términos generales, el rendimiento de los competidores en tareas de identificación se vio afectado negativamente ante factores de variación de canal, duración y lapso temporal inter muestras. De igual forma, los registros test de corta duración y la ausencia de técnicas de normalización de canal en el proceso, contribuyen a un sensible descenso de la eficacia en los sistemas. Los mismos factores y circunstancias adversas acontecieron en los ensayos de verificación.

El informe de Nakasone y Beck concluye afirmando que, en la actualidad, la tecnología de reconocimiento automático no proporciona los resultados que serían deseables, especialmente cuando se enfrenta a las denominadas condiciones forenses. Califican como muy improbable el hecho de que algún día puedan llegar a alcanzarse decisiones de absoluta certeza a través del uso exclusivo de una aplicación de reconocimiento automático, si bien, apuntan algunos aspectos de investigación como objetivos prioritarios para la mejora del rendimiento de los SRAL: mejora de las técnicas de normalización de canal, incorporación de filtros de evaluación cualitativa o cuantitativa de las muestras, uso de diferentes modelos de UBM para cada supuesto de trabajo, integración de información sobre parámetros de alto nivel (suprasegmentos...) etc

PARTE III

Los criterios de decisión en la I.F.L.

La propuesta de los entornos Bayesianos.

Un habitual tema de discusión e inquietud científica en nuestro contexto forense lo constituyen la diversidad de criterios existentes a la hora de materializar las conclusiones de estudio en un informe pericial. Con ello, no pretendemos referirnos al simple hecho de la expresión de unos cálculos o resultados de análisis, sino más bien al de vislumbrar que protocolo es el más idóneo para plasmar tales resultados de una manera objetiva y entendible.

A pesar de las peculiaridades asociadas al proceso de individualizar una voz, la irrupción en escena del reconocimiento automático plantea la conveniencia de establecer reglas de decisión a través de un entorno Bayesiano. Dicho ámbito, nos introduce en el cálculo de la probabilidad de la ocurrencia de un suceso, condicionado por la existencia de otro(s) suceso(s) conocido(s) que, sin ninguna duda, acontecen, acontecerán o ya han acontecido.

Para el marco forense (sobre todo referido a áreas como la identificación por DNA) C. Aitken [1995] propone una interpretación del teorema de Bayes utilizando una relación de probabilidades (apuestas) sobre dos hipótesis competitivas que, a su vez, resultan excluyentes entre sí:

1.- el sospechoso ha realizado la voz dubitada (hipótesis A ó HA) y
2.- la voz dubitada no ha sido realizada por el sospechoso(hip. B ó HB)

A primera vista, esta interpretación del teorema resulta interesante pues, como argumentan sus defensores, permite diferenciar las tareas del científico forense y las del resto de miembros del proceso judicial (jueces, jurado, fiscal, abogados...) . La siguiente igualdad expresa tal propuesta de interpretación:

$$\frac{p(HA/E,i)}{p(HB/E,i)} = \frac{p(E/HA,i)}{p(E/HB,i)} \times \frac{p(HA/i)}{p(HB/i)}$$

Apuestas a posteriori **LR** **Apuestas a priori**

Apliquemos este planteamiento a un supuesto de investigación representativo de nuestro campo. Supongamos que una voz de varón anuncia con una llamada telefónica a una centralita de la Policía, la colocación de un artefacto explosivo. La Policía graba dicha llamada (E) y logra determinar el número del terminal desde el que se ha efectuado la misma. La fracción correspondiente a las **apuestas *"a priori"*** relaciona la probabilidad de la hipótesis del fiscal ó HA -en base a unos datos de investigación (i) existe un sospechoso que ha podido realizar la llamada- con la del abogado defensor del

sospechoso, quien defiende la inocencia de su cliente (HB). La Policía ha investigado la relación de llamadas efectuadas en los últimos meses desde el terminal en cuestión, y comprueban que existen multitud de ellas realizadas a números de centralitas policiales. Además, se logra conocer la identidad del propietario del terminal, el cual, tiene antecedentes policiales por delitos de daños y amenazas. Puestos en contacto los investigadores con el titular del teléfono, comprueban que su voz se percibe bastante similar a la de la llamada maliciosa (E). El individuo es señalado como sospechoso y se solicita la autorización del juez para registrar una muestra de voz del sujeto y efectuar el correspondiente estudio comparativo.

Las muestras de voz dubitada e indubitada son remitidas a los expertos forenses para que determinen si estas pueden, o no, pertenecer a la misma persona. Para ello, habrán de conocer la relación existente entre la probabilidad de que la muestra dubitada presente unas características concretas si ha sido producida por el sospechoso y la probabilidad de que la dubitada posea dichas características si no ha sido emitida por el locutor señalado como sospechoso. Lógicamente, se presupone *a priori* que la voz del sospechoso presenta tales características. El valor obtenido entre esta relación de probabilidades se conoce como **LR (Likelihood Ratio)** ó ratio de verosimilitud, y es la parcela propuesta para enmarcar el campo de actuación del científico forense. La labor del juez o jurado queda circunscrita a la resolución del cociente que definen las denominadas ***"apuestas a posteriori"***. O lo que es lo mismo, la relación entre las probabilidades a favor y en contra de la hipótesis del fiscal, en función de la existencia de unos indicios de investigación y las evidencias científicas.

Aunque los factores que componen la relación del LR pudieran mostrarse no muy entendibles, en definitiva, no representan otra cosa que la esencia formulada en el teorema de Bayes. Trasladado a nuestro entorno, podríamos decir: que para poder conocer la relación entre las probabilidades de si un individuo ha sido o no el autor de un hecho delictivo, en función de una evidencia concreta (voz dubitada) y unos indicios de investigación conocidos, necesitamos calcular las probabilidades de ambas hipótesis y la relación existente entre ambas, condicionando los sucesos de las mismas en sentido inverso. Es decir, a pesar de que las características del mensaje malicioso (E) son conocidas, tratar de inferirlas a partir de los rasgos que caracterizan la voz del sospechoso.

En el caso de la identificación por ADN, el valor de la probabilidad que se refleja en el numerador de la fracción del LR sólo puede ser 1 ó 0, puesto que la comparación del genotipo hallado en la evidencia (por ejemplo, una mancha de sangre en la escena del crimen que no pertenece a la víctima) con el genotipo del ADN de la persona sospechosa, únicamente puede deparar un resultado afirmativo o negativo (exceptuemos casos como las mezclas de sangres). El valor de la probabilidad obtenido en el denominador del LR se deducirá de la frecuencia de aparición del genotipo que caracteriza el ADN de la evidencia en el ADN de los individuos que integren la población de referencia utilizada.

Pero regresemos a nuestro caso del mensaje-amenaza. No es muy difícil darse cuenta de la complicación que comporta el cálculo del LR cuando la

referencia biométrica objeto de análisis ya no es un elemento de carácter invariable (ADN) sino un output de comportamiento, caso de la voz. El gran problema no sólo se refiere a la dificultad de establecer cual es la talla y características idóneas de la población-control, en la que hay que considerar la múltiple variedad de elementos sociolectales, dialectales e idiolectales del habla, sino también, a la diversidad de factores de variabilidad relacionados con otros aspectos de tipo patológico o emocional y otros muchos vinculados a los propios procesos de registro, transmisión, reproducción, conversión, compresión, etc de las emisiones habladas. Es decir, en cierta forma, cada locutor es en sí mismo una población. Además, todo proceso de registro, transmisión o codificación de su voz supondrá una mayor o menor modificación de su cualidad original y, en muchas ocasiones, la incorporación de importantes factores de degradación que dificultarán su evaluación por parte de los expertos forenses.

Como ya hemos comentado, los partidarios de la incorporación de SRAL al campo forense, señalan el entorno Bayesiano como el más idóneo para el cálculo de los resultados de estudio alcanzados por el sistema. Después, exponen distintas alternativas complementarias para lograr una óptima representación de tales resultados, contribuyendo a su mejor interpretación. Una de las más frecuentemente utilizadas son las Tippet plots [Tippett, C. Et al., 1968] especialmente recomendadas para el análisis forense del ADN [Evett I. Et al. 1996].

La funcionalidad de las "escalas de probabilidad verbal".

El ámbito Bayesiano es una herramienta de indudable utilidad para la formulación de resultados de análisis y no sólo en el caso del reconocimiento automático. Su carácter integrador permite incorporar y conjugar los diferentes valores paramétricos, aunque provengan de aproximaciones de análisis independientes. No obstante, en la actualidad la mayoría de los laboratorios que desarrollan técnicas de I.F.L. utilizan las denominadas "escalas de probabilidad verbal" para materializar sus conclusiones de estudio. En síntesis, y dado que existen distintos protocolos periciales en cada laboratorio, la escala de probabilidad verbal suele incluir diversos niveles certeza en los que se enmarca el grado de similitud global obtenido tras completar los distintos análisis de cotejo de muestras.

La casi totalidad de expertos forenses que utilizan estas escalas practican una **metodología combinada** en la que se interrelacionan estudios sobre parámetros acústicos, fonéticos y lingüísticos. Hasta esta última década no puede hablarse de una verdadera actividad de intercambio científico entre dichos laboratorios y, sin embargo, es curioso observar cómo una práctica autónoma de la técnica ha desembocado en la adopción de una fórmula similar para la expresión de conclusiones de análisis. Las escalas de probabilidad verbal han sido tachadas de poseer un carácter subjetivo aunque parecen ser la solución natural para aquellos expertos que han desarrollado sistemáticamente la I.F.L. . El argumento fundamental en torno a este matiz de subjetividad se centra en el proceso de cómo el experto traslada un determinado nivel de similitud entre las muestras comparadas, a un rango de certeza dentro la escala. En general, puede

decirse que cada laboratorio dispone de sus propias características paramétricas y adjudica a las mismas diferentes pesos identificativos en función de la frecuencia de aparición de los mismos en la población de referencia. Por este motivo, y a diferencia de lo parece ocurrir con el reconocimiento automático, las metodologías combinadas presentan limitaciones en cuanto a su ámbito de aplicación lingüística. De igual forma, el establecimiento de unos estándares de uso común para los laboratorios forenses de diferentes países se plantea como una tarea de la máxima complejidad.

Con independencia de estas consideraciones, no hemos de olvidar que salvo en aquellos casos donde la práctica de la técnica se restringe al auxilio de las investigaciones internas de ciertas agencias policiales, la norma habitual es dirigir las conclusiones de trabajo a los tribunales de Justicia. Y, no nos quepa duda de que los tribunales siempre demandarán, además de un conocimiento exacto de las prestaciones de cada técnica, el mayor grado de claridad para poder interpretar correctamente las conclusiones de estudio. En este sentido, las escalas verbales procuran un entendimiento satisfactorio pues pueden ser diseñadas en consonancia a la semántica procesal que más se adecua a cada ámbito jurisdiccional. Pero, ¿ocurre lo mismo cuando enunciamos nuestras reglas de decisión mediante valores LR o formulismos matemáticos? .

Evett y Weir [1998] en su libro sobre interpretación de la evidencia de ADN proponen una equivalencia entre ratios de verosimilitud y conceptos de probabilidad verbal:

Likelihood ratio	Verbal equivalent
1 to 10	Limited support
10 to 100	Moderate support
100 to 1000	Strong support
more than 1000	Very strong support

Ellos mismos reconocen que sobre esta propuesta hay mucho que debatir y que no debe ser tomada al pie de la letra, aunque implícitamente plantean como necesario el hecho de establecer una escala de probabilidad verbal *"como ayuda a un mejor entendimiento de los valores LR"*. Gustan de utilizar el término "support" (apoyo) porque en su entorno lingüístico se les manifiesta como más nítido y aséptico para reflejar el rol que el científico debe desempeñar en su trabajos periciales para la Justicia. Consideran que un valor de verosimilitud superior a 1000 no debe de representar algo más que *"un muy fuerte apoyo"* a la hipótesis barajada.

Regresando a nuestro ejemplo de la llamada maliciosa, supongamos que tras calcular el valor de la relación de verosimilitud sobre las hipótesis y evidencias planteadas, obtenemos un LR de 1000. Entonces, estaríamos afirmando que es 1000 veces más probable la ocurrencia de las características de la evidencia bajo la hipótesis del fiscal, que la de la ocurrencia de las características de la evidencia bajo el planteamiento que propugna el abogado defensor. Referido a una población significaría que de cada 1000 locutores, 1 podría presentar los rasgos de habla estimados en las muestras procedentes del sospechoso. Esta conclusión, que según la propuesta de Evett otorgaría un nivel

de máxima certeza a la hipótesis de culpabilidad del sospechoso, podría ser extrapolada a la globalidad de la población de España (unos 41 millones de personas) y entonces habríamos de admitir que el número esperado de españoles cuyo habla podría presentar unos rasgos similares a los de las muestras registradas del sospechoso, sería de unos 41.000.

Así expuestos, los criterios de correspondencia de Evett pudieran producir, cuando menos, escalofríos. Sin embargo, en el caso de la identificación por "huella genética" (contexto para el que se propone la mencionada tabla de equivalencia) la situación no se revela tan alarmante. Cada genotipo suele comprender 15 marcadores genéticos y a cada uno de ellos se le otorga un valor de verosimilitud. Dado que el LR de cada marcador es absolutamente independiente del de los otros 14, el valor final del LR para un individuo (producto de las probabilidades que definen el LR de cada uno de los 15 marcadores) puede llegar a expresarse en términos de billones y trillones. En general, para un país de etnia homogénea, las coordenadas poblacionales suelen establecerse en función de aquellos perfiles genéticos afines y representativos del número total de habitantes.

Desafortunadamente, en el caso de la I.F.L., la determinación de los ejes que dimensionen la población de referencia, constituye un serio inconveniente. El cálculo del ratio de verosimilitud está claramente influenciado por la talla y características de las grabaciones de voz incluidas en dicha población. Como ya hemos comentado, el problema es a su vez extensivo a la modelación de los registros de voz de un único individuo, pues existen multitud de variables a combinar y considerar. Además, a diferencia de lo que acontece con el ADN, los parámetros y rasgos del habla fluctúan en el transcurso del tiempo, tanto en su plano sociolectal como a nivel individual.

Desde un punto de vista teórico y general, la utilización de estimaciones de verosimilitud para materializar los resultados de análisis forenses, parece una buena solución. Sin embargo, y con independencia de las particularidades ya referidas para la identificación de locutores, la propuesta de interpretación del teorema de Bayes, entendida en su globalidad, no acaba de ofrecer el grado de satisfacción que sería deseable. Si bien los papeles del científico, juez, fiscal, abogado, aparecen diferenciados, no se entiende muy bien qué sentido tiene la relación de sus roles y conclusiones en una igualdad matemática que conjuga supuestas valoraciones objetivas con apreciaciones de carácter subjetivo.

El mismo Aitken [1995] refleja hasta qué punto resulta determinante el valor subjetivo de las apuestas "a priori" sobre el cálculo de las apuestas "a posteriori" partiendo de un valor fijo de LR :

Prior Odds	L.R.	Posterior Odds
1/10.000	1000	1/10
1/100	1000	10
1	1000	1000
100	1000	100.000

Centrándonos en la I.F.L., podríamos formular muchas más cuestiones al respecto. Por ejemplo, en el caso de la aplicación del entorno de Bayes para el análisis combinado clásico, ¿dónde ubicaríamos las tareas de selección o adecuación de las muestras que los laboratorios efectúan como paso previo a los estudios comparativos?. Es decir, ante un supuesto de tramos de voz simulada o afectada en alguna de las muestras, el experto puede decidir despreciar la información derivada de tales tramos. O en el caso del reconocimiento automático, ante muestras degradas por determinados tipos de ruido, puede requerirse de la realización de una labor de procesado para adecuar convenientemente la señal. La influencia de la ejecución de estas tareas, o la de la omisión de fragmentos que en definitiva forman parte de la evidencia, ¿dónde deben quedar reflejadas en la fórmula? ¿en la relación de las apuestas a "a priori"? ¿en el cociente de verosimilitud?. Al fin y al cabo, este tipo de decisiones dependen de la discrecionalidad de cada instituto forense y, por lo tanto seguimos encontrándonos con parcelas de subjetividad que intentamos integrar en un supuesto contexto de objetividad.

En cualquier caso, al final, el científico forense habrá de trasladar sus conclusiones a los miembros del tribunal o jurado haciendo uso de un lenguaje entendible. Por esta razón, hoy por hoy, el uso de las denominadas escalas de probabilidad verbal ha de contemplarse como una solución válida. Eso sí, siempre habrán de ser matizadas en referencia a las propias limitaciones de cada técnica y a las circunstancias particulares de cada caso objeto de estudio.

Conclusiones

Para alcanzar un mejor funcionamiento de las instituciones policiales y judiciales, en aquellos aspectos relacionados con la aportación y valoración de evidencias científicas, resultaría imprescindible el diseño de un plan de actuación que, por un lado posibilitase la articulación de una normativa acorde a la realidad de cada momento y, por otro, proveyese a dichos organismos de los expertos y métodos de trabajo más adecuados.

Las nuevas aplicaciones de reconocimiento automático de locutores, se perfilan como una alternativa complementaria a las aproximaciones de estudio combinadas. Su utilización reporta claras ventajas (agilidad comparativa ante grandes cantidades de información, reconocimiento independiente de texto) aunque todavía han de superar diversos inconvenientes. Probablemente, nos encontremos en una fase de transición en la que los programas automáticos van cobrando un mayor protagonismo. Por el momento, resulta complicado predecir si algún día llegará a materializarse una completa automatización. En este sentido, como capítulos prioritarios de resolución, cabría citar:

- el establecimiento de unos criterios que permitan definir con claridad los márgenes de admisibilidad para las muestras test y los mínimos de caracterización que deben reunir los modelos que integran la UBM de referencia. De esta forma, será posible conocer ante qué presupuestos de análisis los resultados de estudio aportan un rango suficiente de fiabilidad.

- la definición de las tallas y características de las bases de datos poblacionales. Argumento fundamental para una correcta evaluación de los resultados de estudio y la formulación de conclusiones.

- la incorporación a las aplicaciones de estimaciones de caracterización y modelado a nivel prosódico y lingüístico.

En el caso de la identificación forense de locutores, la interpretación de conclusiones de estudio a través de un entorno Bayesiano, resulta difícil de concretar. Existen aspectos, no sólo referidos a la población de referencia utilizada, sino también a la naturaleza variable del habla en un mismo individuo, que aparecen como serios obstáculos al carácter objetivo que dicho ámbito pretende conferir a las tareas asignadas a los científicos. Por otra parte, y haciéndolo ya extensivo a otras disciplinas forenses, es complicado evitar el uso de equivalencias verbales para expresar de una forma más comprensible los niveles de certeza alcanzados por los expertos. La intervención humana, tanto en los procesos de selección y análisis de muestras, como en el de traducción de resultados, aunque subjetiva, se antoja ineludible en la actualidad.

Referencias

[1] w.w.w. odyssey04.org, 2004.

[Aitken, C.G.G., 1995] *Statistics and the evaluation of evidence for forensic scientists*, Cheicester, Reino Unido, 1995.

[Daubert *vs* MerellD. Ph., 1993] 509 U.S. 579, 113 S. Ct. 2786, 125L. Ed. 2d 469, 1993.

[Delgado, C., 2001] *La identificación de locutores en el ámbito forense*, Tesis doctoral, Facultad de Ciencias de la Información, Universidad Complutense, Madrid.

[Doddington, G. et al., 2000] The NIST speaker recognition evaluation: overview, methodology, systems,results, prespective, *Speech Communication, 31, 2000*, pp. 225-254.

[Doddington, G., 2000] Some experiments on idiolectal differences among speakers, *(w.w.w.nist.gov./speech/tests/spk/2000/doc/N-gram_experiments-V06.pdf)*

[Evett, I., et al., 1996] "Statistical analysis of STR data" in : *"Advances in Forensic Haemogenetics",* Vol. 6. Springer-Verlag, Berlin. Pp. 79-86 in Carracedo, A., Brinkman and W. Bar (Eds.)

[Evett, I. y Weir, B., 1998] *Interpreting DNA evidence. Statistical Genetics for Forensic Scientists*. Sunderland, Massachusetts, 1998.

[Leeuwen, D. y Bouten, J., 2003] The NFI/TNO Forensic Speaker Recognition Evaluation Plan. Revision 2.0

[Martin, A. y Przybocki, M., 2002] The NIST Speaker Recognition Evaluations : 1996-2001 *(w.w.w.nist.gov./speech)*

[Nakasone, H. y Beck, S., 2001] Forensic Automatic Speaker Recognition. *Odyssey 2001 Speaker Recognition Workshop, Creta , Grecia, 18-22 de junio de 2001.*

[Przybocki, M. y Martin, A., 1998] The NIST Speaker Recognition Evaluation 1997, *RLA2C , Avignon, April 1998, pp. 120-123.*

[Tippet et al., 1968] "The evidential value of the comparison of Paint Flakes from sources other than vehicles. *Journal Forensic Sci. Soc., pp. 61-65, 1968.*