Identificación forense de hablantes: un tutorial

Pedro Univaso

***Abstract*— This paper presents an overview of the methodologies used in the forensic field for the identification of speakers. First, an introduction shows the interest of the forensic community in speaker recognition and its limitations. Then the history of the evolution of speaker recognition systems -from Bell Laboratories in the fifties to current i-vector/PLDA approach-, and the much older history of forensic speaker identification that dates back to the seventeenth century at the court of Charles I of England -from recognition techniques by listening to semi-automatic systems nowadays-, are presented. To clarify concepts, some important concepts as speaker recognition tasks: verification and identification, and speaker recognition modalities: text-dependent/independent, are defined. Topics covered in this work range from particularities of forensic speaker identification tasks and complexities of collecting and processing speech samples, to final presentation of results to judge. Finally, a classification of forensic speaker identification approaches: auditory-spectrographic, auditory-perceptual, acoustic-phonetic, automatic, semi-automatic and combined, are analyzed in depth. A section presents an analysis of the institutions in Argentina working in this area.**

***Keywords*—Forensic speaker identification, i-vector, .**

data. The evaluation of a speaker verification system is then detailed, and the detection error trade-off (DET) curve is explained.
Several extensions of speaker verification are then enumerated, including speaker tracking and segmentation by speakers. Then, some applications of speaker verification are proposed, including on-site applications, remote applications, applications relative to structuring audio information, and games. Issues concerning the forensic area are then recalled, as we believe it is very important to inform people about the actual performance and limitations of speaker verification systems. This paper concludes by giving a few research trends in speaker verification for the next couple of years.

Keywords and phrases: speaker verification, text-independent, cepstral analysis, Gaussian mixture modeling.

I. INTRODUCCIÓN

EL reconocimiento de hablantes (RH), también conocido como reconocimiento de locutor o de las personas por la voz, es una técnica biométrica de comparación que incluye la verificación o autenticación, identificación y clasificación de una persona por su voz, y por extensión, la segmentación, detección y rastreo de hablantes.

El interés particular de esta técnica biométrica con respecto a otras, como pueden ser las del ácido desoxirribonucleico (ADN), huellas dactilares, iris, córnea, etc., es su naturaleza no-intrusiva y la posibilidad de procesamiento remoto, especialmente ante el creciente uso de la telefonía e Internet en la cultura actual. Por otra parte las desventajas son la variabilidad intra-hablante, los efectos del canal de transmisión y la susceptibilidad a los ruidos de fondo.

El RH puede referirse al reconocimiento automático realizado por medios informáticos en contraposición del reconocimiento basado en humanos que puede ser por escucha y/o por análisis visual de representaciones gráficas del habla.

El reconocimiento de hablantes se basa en la modelación estadística o matemática de las características del tracto vocal de una persona. La modelación representa la fisiología de la persona que produce el habla humana, expresada en una señal acústica. Una vez que un modelo es asociado a una persona, se calcula la verosimilitud de la emisión acústica incógnita como emitida por dicho modelo en contraposición con la de otros modelos de diferentes hablantes. Esta metodología base es el sustento de los diferentes tipos de sistemas de reconocimiento de hablantes.

Para poder incorporar a estas características fisiológicas otras particulares del modo de hablar del hablante, los sistemas actuales adicionan a este modelo rasgos distintivos del habla de largo plazo (tono, duración, acento y calidad de voz).

Existe una diferencia entre el reconocimiento de hablantes y el reconocimiento de habla. Ambos términos puede confundirse y el reconocimiento de voz puede emplearse para expresar a ambos. Mientras que el reconocimiento de habla busca determinar qué es lo que una persona está diciendo el reconocimiento de hablantes intenta determinar quién está hablando.

II. HISTORIA DEL RECONOCIMIENTO DE HABLANTES

La historia de los primeros trabajos sobre RH se remonta a los años 50 [1, 2], habiendo sido Pruzansky [3] en los 60s, de los Laboratorios Bell, el que desarrolló uno de los primeros sistemas, que empleaba bancos de filtros para la comparación de espectrogramas. Posteriormente Doddington en Texas Instruments reemplazó los bancos de filtro por el análisis de formantes [4].

En la década de los 70s las variaciones intra-hablantes fueron investigadas por Endres *et al*. [5] y Furui [6].

Hasta la década de los 80s los primeros sistemas de reconocimiento de habla utilizaban técnicas de programación dinámica (*DTW*) para alinear las emisiones de habla con las secuencias fonémicas a comparar [7], siendo posteriormente suplantadas por el uso de los modelos ocultos de Markov

Pedro Univaso, BlackVOX & Laboratorio de Investigaciones Sensoriales (INIGEM-CONICET, UBA), Buenos Aires, Argentina, punivaso@blackvox.com.ar

(*HMM*), metodología también empleada en los primeros sistemas de reconocimiento de hablantes [8]. Al comienzo de la década de los 90s Rose y Reynolds [9] propusieron el uso de un *HMM* de un único estado, denominado Modelo de Mezclas Gaussianas (*GMM*) y fue empleado en la mayoría de los sistemas de RH. En esa década se profundizó el RH robusto que ataca las diferencias intra-hablante, el ruido y las diferencias de canal. También se introdujo el uso de modelos de referencia universal y cohortes de hablantes para el cálculo de relaciones de verosimilitud [10] y el modelado empleando redes neuronales [11]. A finales de los 90s el Instituto Nacional de Estándares y Tecnología (*NIST*) de los Estados Unidos de Norteamérica inició las evaluaciones de sistemas de reconocimiento de hablantes (*SRE*), que continúan hasta la actualidad, para determinar el estado del arte.

En los 2000s se incorporaron diversas técnicas de normalización de modelos de hablantes basadas en la normalización cero (*Z-norm*) desarrollada por Li y Porter a fines de los 90s [12], la introducción de los primeros rasgos distintivos de largo plazo -en este caso rasgos idiolectales [13]-, y se comenzaron a emplear las máquinas de soporte vectorial (*SVM*) como clasificadores [14]. Uno de los principales sistemas de RH, desarrollado en el laboratorio *SRI International*, incorpora rasgos suprasegmentales o prosódicos para mejorar la performance de los RH [15].

En la década del 2010 están siendo investigados los métodos de representación del habla en subespacios vectoriales, como herramientas para remover o atenuar las características que no son propias del hablante (e.g. efectos relacionados con el hablante y el canal), con técnicas de transformación como el análisis factorial conjunto (*JFA - Joint Factor Analysis*) [16] y las aproximaciones por vectores de factor total (*i-vectors*) [17] normalizados según el análisis discriminativo lineal probabilístico (*PLDA – Probabilistic Linear Discriminant Analysis*) [18]. Técnicas que, empleando exclusivamente parámetros segmentales, son consideradas el estado del arte en la verificación de hablantes en la actualidad, de acuerdo a las evaluaciones de *NIST*.

La historia del reconocimiento de hablantes en el ámbito forense es mucho más antigua y se remonta al siglo XVII en la corte de Carlos I de Inglaterra [19]. La primera investigación científica fue realizada en 1937 en el caso Lindbergh [20], donde la voz del incriminado, Bruno R. Hauptmann, fue positivamente identificado por parte del médico John F. Condon, intermediario en la negociación del rescate.

La identificación de hablantes comenzó empleando técnicas de reconocimiento por escucha, haciendo uso de la discriminación de voces por parte de los humanos. La identificación de hablantes comienza a ser practicada de forma sistemática por científicos forenses a finales de los 40s en la Unión de Repúblicas Socialistas Soviéticas y en los 50s en los Estados Unidos de Norteamérica.

Durante los años 60, y a la luz de la confiabilidad de los sistemas de huellas dactilares, desarrollados a fines del siglo XIX, Kersta propuso un sistema similar denominado: huella de voz [21], basado en el análisis visual de espectrogramas, el cual no alcanzó los resultados esperados dada la variabilidad intrínseca del habla. La controversia sobre la conveniencia de su uso duró más de 10 años y concluyó con el informe de Stevens *et al.* [22] que determinó que el método auditivo era más preciso que la inspección visual.

A partir de los 70s, los métodos basados en humanos, contaron con el aporte de los sistemas automáticos de reconocimiento de hablantes desarrollados para aplicaciones biométricas. A pesar de que dichos sistemas evolucionaron en aplicaciones comerciales, no se pudieron crear sistemas confiables a ser empleados en el ámbito forense, existiendo una permanente controversia en su uso por parte de los expertos forenses y los científicos del RH.

La proporción de casos en los que se admite la evidencia empleando técnicas de identificación de hablantes no ha aumentado durante los últimos años a pesar de la mejora en la tecnología empleada. De veinte cortes en los Estados Unidos de Norteamérica del año 1977 al 1999 que consideraron emplear estas técnicas, doce admitieron su uso y ocho las excluyeron [23].

La recomendación del comité de expertos convocado por la Academia Nacional de Ciencias de los Estados Unidos de América durante el año 1976, a pedido del FBI [24], y que incluyó expertos en acústica, ciencias del habla, patología de la voz, ingenieros eléctricos y electrónicos, en grabación de audio y en leyes y evidencias criminales, fue la siguiente:

a. Se puede obtener alguna información de la identidad de una persona a través de la percepción acústica y del análisis de espectrogramas.
b. Los espectrogramas difieren de las huellas digitales en que una misma palabra puede variar acústicamente cada vez que es dicha por una misma persona (variaciones intra-hablante).
c. No ha sido suficientemente demostrado científicamente que las variaciones acústicas intra-hablante sean menores que las variaciones entre-hablantes.
d. Los errores de identificación son muy dependientes de las propiedades de la voz, de las condiciones de las muestras, del equipamiento empleado y de las habilidades del experto examinador.

El comité concluyó que para lograr un sistema de identificación completo basado en la percepción auditiva-visual y con el empleo de métodos automáticos se deben continuar con las investigaciones y desarrollos en el área. Y sugirió que se expliciten las limitaciones de las técnicas empleadas ante cada presentación de una evidencia de identificación. Después de la publicación de este reporte el FBI dejó de brindar servicios de identificación de hablantes con el propósito de ser empleados como testimonio en las cortes judiciales norteamericanas.

En el 2003 Bonastre *et al.* [25] presentaron un informe alertando sobre la imposibilidad de identificar unívocamente a una persona por su voz con los avances científicos alcanzados hasta el momento, especialmente en el ámbito forense donde el entorno y los factores que afectan la performance pueden variar tremendamente con respecto al ámbito comercial.

En 2005 Saks y Koehler [26] propusieron un cambio de paradigma en la ciencia forense de identificación en referencia a la discernibilidad y unicidad de las muestras, en base a las evidencias de error en pruebas realizadas y casos reales. Dicho postulado concluye con la necesidad de lograr niveles de confiabilidad similares a los obtenidos en la identificación con ADN por medio de ensayos estandarizados con base empírica y probabilística. En esa época se comenzó a emplear una nueva metodología de evaluación de los sistemas de reconocimiento de hablantes en el ámbito forense, denominada función de costo logarítmica (C_{llr}) [27], la cual introduce medidas de validez y confiabilidad.

Durante la primera década del 2000, un grupo de investigadores del Reino Unido [28, 29, 30] comenzaron una discusión sobre las características que debe cumplir la comparación de hablantes en el ámbito forense, concluyendo en la importancia de la presentación de resultados cuantitativos en la forma de relaciones de verosimilitud (*LR*) en lugar de resultados binarios, siendo el juez quien determine la identidad del sospechoso en base a las comparaciones realizadas por los peritos.

Para profundizar en la problemática particular del ámbito forense, el Instituto Nacional de Estándares y Tecnología (*NIST*) de los Estados Unidos de Norteamérica, en la evaluación de sistemas de reconocimiento de hablantes de 2010 (*SRE*), incorporó una nueva evaluación diseñada para comparar sistemas con intervención humana denominada *HASR* (*Human Assisted Speaker Recognition*) [33]. Los Estados Unidos de Norteamérica, a través del *NIST*, han formado recientemente una organización científica (*OSAC*) (http://www.nist.gov/forensics/osac.cfm) para formalizar el proceso de mejores prácticas para el área forense en general.

III. RECONOCIMIENTO DE HABLANTES: DEFINICIÓN

El área del reconocimiento de hablantes tiene como objetivo la determinación de una persona a partir de su voz, pudiéndosela dividir en dos sub-áreas: la verificación y la identificación de hablantes.

En la verificación se pretende determinar si un hablante es quien dice ser mediante su voz, o detectarlo en una conversación, estableciendo si un segmento de habla fue emitido por él. En la verificación la respuesta del sistema es binaria: acepta o rechaza la identidad del hablante haciendo una sola comparación y utilizando un umbral que pesa el costo de aceptar un impostor o rechazar un hablante verdadero.

El proceso de verificación consiste en que el hablante a verificar debe identificarse previamente, generalmente por métodos no-orales (Nombre de usuario, Número de Identificación, Clave de usuario, etc.), generando una identificación personal (ID). Dicha ID selecciona de una base de datos el modelo acústico correspondiente a dicho hablante. Este modelo, denominado modelo de referencia, debe haber sido previamente generado por dicho hablante y almacenado en la base. Entonces la emisión acústica del hablante es comparada con el modelo de referencia para verificar la identidad del hablante de prueba. Dicha comparación no es suficiente para tal alcance, dado que es necesario contrastarla con respecto a algún patrón que nos determine el parecido o diferencia entre ambas muestras.

Los métodos empleados actualmente para contrastar al hablante de referencia con el hablante de prueba consisten en introducir uno o varios modelos competitivos. El primer método emplea un modelo de referencia universal (*UBM*) conformado por datos pertenecientes a una gran población de hablantes. La idea subyacente es que si la emisión del hablante de prueba es más parecida al promedio de la población que al hablante de referencia, entonces es más probable que dicho hablante no sea el hablante de referencia, y viceversa.

El segundo método emplea un modelo de cohorte, siendo los miembros de esta cohorte aquellos que poseen una voz similar a la del hablante de prueba. La metodología de contraste es similar a la anterior, sólo que en este caso no es necesario involucrar a toda la población sino a un grupo selecto de la misma. La comparación en este caso es entre el hablante de prueba y su cohorte.

Como puede verse se requieren sólo dos comparaciones para la verificación de cada hablante, una con respecto al modelo de referencia y otra con respecto al modelo competitivo. Con lo cual, y a pesar de que la cantidad de hablantes a reconocer crezca, la capacidad de cómputo requerida queda constante. Esta simplicidad de cálculo, aunque no es que sea un problema intrínsecamente simple, hace que esta categoría de reconocedor de hablantes sea la de mayor popularidad y la que primero ha llegado al mercado comercial.

Uno de los principales problemas es la determinación de la cantidad y características de los hablantes a incorporar en los modelos competitivos. Dado que es imposible abarcar todos los casos posibles requeridos se debe llegar a una situación de compromiso que permita hacer práctico el empleo de esta metodología para la verificación de hablantes.

La identificación, en cambio, realiza la comparación del audio del hablante de prueba con un número determinado de modelos de hablantes. La probabilidad de error de la identificación tiende a uno en la medida que el número de hablantes con que debe compararse aumenta. Al aumentar el número de hablantes crecen las probabilidades de que dos o más hablantes tengan distribuciones muy cercanas unas a otras. En esas circunstancias la identificación es una tarea difícil de resolver.

A pesar de las diferencias enumeradas es común encontrar en la literatura indistintamente la denominación de identificación, verificación o simplemente reconocimiento a estos diferentes tipos de reconocimiento de hablantes.

IV. MODALIDADES DEL RECONOCIMIENTO DE HABLANTES

Modalidad dependiente del texto

Un sistema de reconocimiento de hablantes es dependiente del texto o de modo limitado si emplea una cantidad limitada de fonemas, palabras o frases al momento de requerirle información al hablante de prueba. Por ejemplo: "*por favor, diga 11-22-43*" o "*diga su contraseña de 4 letras*".

El empleo de restricciones en el texto permite lograr mejoras en la performance del sistema pero lo hace menos flexible. Por otra parte es necesario que el usuario coopere con el sistema y conozca su funcionamiento *a priori*.

Modalidad independiente del texto

Un sistema de reconocimiento de hablantes es independiente del texto o de modo ilimitado si no restringe el texto a ser empleado por el hablante de prueba al momento de requerirle información.

Existen diversos grados de libertad en la independencia del texto, pudiendo existir restricciones parciales de texto, lenguaje o idioma.

Un sistema puro independiente del texto y del idioma analiza únicamente las características del tracto vocal sin considerar el resto del contexto del habla. Uno de los mayores inconvenientes de los sistemas independientes del texto es la imposibilidad de cubrir todas las variantes del habla. Durante el entrenamiento del modelo de referencia se trata de reducir el tiempo de la sesión y la extensión de las frases. Esto hace que se logre una cobertura fonética limitada y que parte del habla de prueba nunca haya sido entrenada y por lo tanto no se encuentre incorporada al modelo de referencia, pudiendo producir errores en el reconocimiento. Para disminuir estos errores las sesiones de entrenamiento de este tipo de sistemas son más extensas que los dependientes del texto, pero poseen como ventaja la posibilidad de poder emplear con libertad cualquier texto para el entrenamiento.

La modalidad independiente del texto es la más versátil y puede emplearse en cualquiera de las categorías del reconocimiento de hablantes.

V. IDENTIFICACIÓN FORENSE DE HABLANTES

La identificación de hablantes empleada en el ámbito forense parte de la grabación de una voz relacionada con un hecho delictivo (grabación dubitada, prueba o evidencia) la cual es comparada con otros registros atribuidos a una persona, normalmente conocida (grabación indubitada o plana de voz del imputado).

En el caso de delitos de los secuestros extorsivos la grabación dubitada generalmente se obtiene de registros telefónicos (teléfonos fijos o celulares), mientras que la indubitada se realiza durante la toma de declaración del imputado. En este último caso resulta aconsejable cumplir con las formalidades del código procesal penal de cada país, donde algunos prevén que para la rueda de voces se utilice un cuerpo o plana de voz correspondiente al imputado, y otras grabadas por terceros, en lo que se les hace repetir generalmente las mismas frases, a fin de cumplir con el requisito de condiciones semejantes exigidos por algunas legislaciones en materia de reconocimientos.

Cuando es factible esta última condición los peritos forenses aplican la identificación del hablante en la modalidad dependiente del texto. Dada la posibilidad, prevista ya en la ley, de que el imputado se rehúse a repetir las frases solicitadas, se realizan planas de voz indubitadas cuyo texto difiere de las dubitadas, aplicando en este caso la modalidad independiente del texto.

Una aplicación de la identificación del hablante es la requerida por la justicia en casos civiles o comerciales. En estos casos la grabación dubitada puede provenir de un registro telefónico, pero también de grabaciones en vivo autorizadas por el juzgado.

La identificación de hablantes empleada en los casos forenses y legales puede emplear métodos de reconocimiento basados en humanos o apoyados por la tecnología. La identificación basada en humanos puede ser realizada por los mismos testigos que deben identificar la voz del imputado de entre las presentes en la rueda de voces, o por expertos fonetistas, lingüistas o fonoaudiólogos. En este último caso, algunos emplean herramientas informáticas y sistemas de reconocimiento automático de hablantes que los ayudan en la toma de decisión.

Otra aplicación forense es la segmentación o clasificación de la grabación dubitada. La segmentación se utiliza para elegir los tramos de voz del imputado de entre varias fuentes de habla presentes en la grabación. La clasificación puede ayudar a categorizar el segmento de audio, así como identificar anomalías de interés como ser el disparo de un arma de fuego, una explosión o el sonido de una carretera o un tren en las cercanías.

La investigación forense de audio es un área más amplia que el reconocimiento de hablantes e incluye la autenticación y optimización de los registros de audio. En estos casos se emplean otras técnicas que ayudan a los peritos en su investigación: filtrado de ruidos, análisis de autenticidad de la grabación, filtrado de señales, recuperación de datos, etc.

Podemos enmarcar la identificación forense de hablantes dentro de la fonética forense [31], y más ampliamente, dentro de la lingüística forense [32]. Las áreas relacionadas con la fonética forense, según Rose [31], son: la identificación de hablantes – que es la abarcaremos en este trabajo -; la determinación de perfiles de hablantes (ante la falta de un sospechoso, dar información sobre su acento socioeconómico); la construcción de ruedas de voces (para el reconocimiento de voces por parte de testigos o víctimas); la identificación de contenido (determinando lo que fue dicho cuando la grabación es de mala calidad, o cuando la voz es patológica o tiene un acento extranjero); y la autenticación de los registros de audio (determinando si una grabación ha sido manipulada). Jessen [38] propone una clasificación que ha probado ser de gran utilidad en la práctica forense y que puede verse en la Tabla 1.

TABLA 1. TAREAS DE IDENTIFICACIÓN FORENSE DE HABLANTES

<table>
<tr><th></th><th>Existe una grabación del hablante desconocido</th><th>No existe grabación, pero sí un testigo</th></tr>
<tr><td rowspan="2">Existe un sospechoso</td><td rowspan="2">Identificación del hablante desconocido</td><td>Declaración del testigo (si el testigo conoce al sospechoso)</td></tr>
<tr><td>Rueda de voces (si el testigo no conoce al sospechoso)</td></tr>
</table>

No existe un sospechoso	Determinación del perfil del hablante desconocido	Raramente se solicita la intervención de expertos

La tarea de identificación del hablante desconocido (e.g. secuestrador, traficante de drogas, acosador, etc.), si el sospechoso es cooperativo, puede realizarse obteniendo grabaciones de la voz del mismo. Se pueden realizar sesiones de grabación de textos leídos, similares o iguales a los presentes en la grabación del hablante desconocido. Como se vio, la metodología de reconocimiento de hablantes dependiente del texto permite lograr mejoras en la performance del sistema. Sin embargo, la dependencia del texto no es un requerimiento para la identificación del hablante, y la grabación que se obtiene en estos casos puede resultar en una prosodia poco natural, lo cual crea un problema adicional. Por lo tanto la grabación del sospechoso debería incluir también habla espontánea, similar a la del hablante desconocido. En el caso de que el sospechoso no coopere, y dependiendo de la legislación vigente del país, pueden utilizarse grabaciones previas del mismo (toma de declaraciones grabadas en sede policial, grabaciones telefónicas, etc.), las cuales pueden incluir graves interferencias si no se tienen en cuenta al realizar las grabaciones (e.g. ruido del teclado durante la toma de declaración, distancia entre el micrófono y el sospechoso, etc.). Otra forma de comportamiento no-cooperativo es cuando el sospechoso enmascara su voz de una manera aparente o sutil durante la grabación.

La determinación del perfil del hablante desconocido es útil cuando no se tiene un sospechoso, y la policía requiere concentrar la búsqueda en un número menor de posibles sospechosos. El perfil puede incluir el tamaño del cuerpo (altura y peso) del mismo -correlacionado con la dimensión del tracto vocal [39]-; las condiciones médicas en el ámbito del lenguaje, el habla y las patologías de la voz; la edad; el sexo o género; el sociolecto; el idiolecto (o regiolecto); el etnolecto o acento extranjero; y el tipo de lenguaje en casos de regiones de diferentes lenguas o multilingüismo.

En el caso de que exista un testigo que haya escuchado la voz del hablante desconocido y que conozca su identidad (e.g. pariente, vecino, etc.), la tarea forense se reduce a la toma de declaración. Pero si no lo conoce, se puede realizar una rueda de voces, en la que el testigo deba reconocer la voz del hablante desconocido entre varias voces similares de otros hablantes.

En la actualidad no se requiere ninguna tarea forense si no se tiene una grabación del hablante desconocido. En el futuro es de esperar que, existiendo un testigo, se puedan emplear técnicas de síntesis de habla para la generación de identikits acústicos.

Otra clasificación generalmente empleada considera los conocimientos de lingüística forense que posee la persona que realiza la identificación. De esta manera, las tareas presentadas en la Tabla 1 en las cuales se posee una grabación del hablante desconocido que deben ser realizadas por un profesional del habla se consideran dentro de la categoría de identificación por expertos, mientras que el resto pertenecerían a la clase de identificación naïve o realizada por no-expertos (testigo).

El tema que más ha interesado a los investigadores forenses en los últimos años es la necesidad de distinguir entre enfoques científicos y pseudo-científicos de la identificación de hablantes realizada por expertos. Hollien [40] incluye dentro del grupo de legítimos investigadores a fonetistas, lingüistas, ingenieros y computadores científicos que poseen conocimientos de la ciencia de la voz. Mientras que alerta sobre la presencia de "*un número de charlatanes que han invadido este campo*", entre los que se encontrarían algunos detectives privados, agentes de la ley, técnicos con algún conocimiento sobre el procesamiento de señales de audio, entre otros.

En el contexto de la identificación de hablantes en el ámbito forense, el investigador debe poder evaluar la verosimilitud de que el habla del delincuente sea la del sospechoso (similitud), con respecto a la verosimilitud de que el habla del delincuente sea la de otra persona de la población relevante (tipicidad), empleando para ello resultados cuantitativos (*LR*). Otros investigadores resaltan la importancia del uso de bases de datos representativas de la población relevante que reflejen las del caso bajo investigación, así como el empleo de modelos estadísticos [41].

Dentro de las principales asociaciones internacionales que estudian la identificación de hablantes en el ámbito forense se encuentra la Asociación Internacional de Acústica y Fonética Forense (*IAFPA – International Association for Forensic Phonetics and Acoustics*), y la Asociación Internacional de Lingüística Forense (*IAFL - International Association of Forensic Linguistic*), las cuales congregan a sendas reuniones anuales, y están representadas en la Revista Internacional de Habla, Lenguaje y Ley (*International Journal of Speech, Language and the Law*). En las mismas se congregan principalmente fonetistas, y en menor medida ingenieros de tecnologías del habla. Otra agrupación que representa a 58 laboratorios en 33 países es el Consejo Europeo de Redes de Instituciones de la Ciencia Forense (*ENFSI - Board of the European Netwotk of Forensic Science Institutions*). En este grupo los ingenieros y computadores científicos están al menos tan representados como los fonetistas y lingüistas.

En el 13º Simposio de Ciencias Forenses, que tuvo lugar en Lyon, Francia, en octubre de 2001 [42] se presentó el resultado de una encuesta realizada a los gobiernos y países miembros de la *INTERPOL*, para conocer el estado de actividad en el campo de la lingüística forense. Se recibieron 30 cuestionarios de 21 de los 190 países miembros, de los cuales 5 laboratorios indicaron que no estaban activos en esta área. Los resultados mostraron que los laboratorios llegan a procesar desde unos pocos casos de identificación de hablantes por año hasta varios cientos (ver Tabla 2).

TABLA 2. CANTIDAD DE CASOS DE IDENTIFICACIÓN DE HABLANTES POR AÑO Y POR LABORATORIO.

Laboratorio	País	Casos por año
Instituto Lituano de Investigación Forense	Lituania	500-600
Instituto Criminalístico de Praga	República Checa	100
Policía Científica de Madrid	España	100
Ministerio de Relaciones Internas	Belarrús	100
FBI (sólo investigación)	USA	100
BKA	Alemania	90-100
NFI	Holanda	50-80
Policía Federal	Bélgica	50-80
Gendarmería Nacional (sistema automático)	Francia	10
ETH Zürich (sistema automático)	Suiza	10
Guardia Civil (sistema automático)	España	2
Servicio de Ciencias Forenses	Reino Unido	0
Policía Metropolitana Forense	Reino Unido	0

Como en el resto de las ciencias forenses en general, es cada vez más común los procedimientos de aseguramiento de la calidad de los laboratorios que realizan estos tipos de análisis. Existen varios organismos encargados de certificar laboratorios, entre los que se encuentra el *ENFSI* en Europa, *UKAS* en Gran Bretaña, el Consejo de Acreditación en Holanda, grupos en USA como el *SWGDOC* (*Scientific Working Group for Document Examination*), o el *SWGMAT* (*Scientific Working Group for Materials Analysis*), auspiciados por el FBI, y el *SAG* (*Scientific Advisory Groups*) en Australia y Nueva Zelanda. Uno de los primeros laboratorios que preparó la certificación de su metodología de identificación de hablantes durante el año 2001 fue el *IRCGN* de Francia. La *IAFP* inició, también ese año, un procedimiento de acreditación práctica de fonetistas forenses.

Las bases de datos de uso forense reportadas en el 13º Simposio de Ciencias Forenses [42] fueron las siguientes: DRUG (Alemania), KISTE y TELDAT (Suiza), ETH (Suiza), AHUMADA/GAUDI conformada por 450 hablantes (España), corpus de diferentes lenguas (Austria), llamadas telefónicas anónimas (República Checa), base de datos de 250 hombres y 150 mujeres (Francia), grabaciones de eventos (Israel), CGN (Holanda), LOCOPOL (Policía Científica de España), IDEM base de datos de formantes (Italia).

Un caso particular: la Argentina

En la Argentina los casos de identificación de hablantes por medio de la voz solicitados por la justicia son procesados por la Policía Federal Argentina, la Gendarmería Nacional Argentina y por la Dirección General de la Asesoría Pericial del Poder Judicial en la provincia de Buenos Aires. La Policía Federal Argentina, a través del Gabinete de Identificación de la Voz, perteneciente a la División Scopometría de la Superintendencia de Policía Científica, emplea una metodología semi-automática por medio de la colaboración que le brinda al experto un sistema interactivo de análisis y procesamiento de señales acústicas desarrollado por el Dr. Sergey Koval. El Dr. Koval es reconocido como uno de los expertos mejor calificados en el mundo en el análisis forense de audio, y cuenta con más de 32 años de experiencia y más de 1.000 peritajes de audio forense ejecutados por su autoría y bajo su dirección. Es uno de los creadores de las metodologías de identificación de hablantes utilizadas por los órganos de seguridad interna de la Federación Rusa. La Policía Federal Argentina que durante el año 2001 realizó un total de 50 identificaciones de voz está encarando a partir del año 2011 un proyecto para la creación de un Banco Nacional de Voces Delictivas a incorporar al servicio que actualmente brinda. En la Gendarmería Nacional Argentina, la División Sónica, perteneciente al Departamento de Estudios Especiales, posee un grupo de trabajo en el cual los expertos utilizan un método de medición visual de formantes de vocales pertenecientes a las grabaciones a comparar. En cambio la Asesoría Pericial de la provincia de Buenos Aires emplea un método de identificación perceptual llevado a cabo por un equipo de fonoaudiólogos expertos para la realización de las pericias de voz que se le encomiendan.

Tanto la Policía Federal como la Gendarmería Nacional han adquirido recientemente un sistema automático de identificación de hablantes de origen ruso que compara las grabaciones con respecto a un grupo de modelos de voces de diferentes fuentes (celulares, telefonía fija y entrevistas). Las metodologías empleadas por la policía, la gendarmería y la Asesoría Pericial emplean métodos de identificación de hablantes en los cuales comparan las voces del sospechoso y la evidencia sin contrastarlas con otras bases universales de voces. Como se vio anteriormente, esta metodología está siendo cuestionada por los científicos de diversas ramas de las ciencias forenses, no sólo de la identificación de voces, sino también de la identificación de huellas dactilares, escritura, marcas de herramientas, de proyectiles de armas de fuego, cabellos, marcas dentales, etc., debido a los errores que se han encontrado en diversos casos judiciales. La ciencia forense tradicional de individualización o identificación se basa en una premisa central, denominada unicidad discernible, que es la de suponer

que dos marcas indistinguibles deben ser producidas por el mismo objeto o persona, excluyendo cualquier otro objeto o persona en el mundo que la hubiera podido producir. Esta premisa le brinda importantes beneficios prácticos a la ciencia forense, la cual se excusa de desarrollar mediciones de atributos objetivos, de recolectar costosas bases de datos universales de frecuencia de variación de dichos atributos, analizando la independencia de los atributos o calculando la probabilidad de que diferentes objetos puedan poseer los mismos atributos observables. Esta nueva visión de la ciencia de la identificación forense es denominada por Saks *et al.* [26] como el "nuevo cambio de paradigma". En el país se realizan trabajos de investigación relacionados con el reconocimiento de hablantes para el ámbito forense en la Carrera de Fonoaudiología de la Facultad de Medicina de la Universidad de Buenos Aires, en el Laboratorio de Investigaciones Sensoriales, LIS (INIGEM, UBA-CONICET) y en el Departamento de Electrónica del Instituto Tecnológico de Buenos Aires (ITBA). El sistema de reconocimiento de hablantes empleado en la Carrera de Fonoaudiología de la UBA se basa en la metodología audio-perceptual desarrollada por el Dr. Harry Hollien, la cual tampoco considera el contraste de los resultados con respecto a una base universal de voces. Un panel de fonoaudiólogos valoriza subjetivamente una serie de rasgos distintivos predeterminados: entonación, articulación, calidad de la voz, prosodia, intensidad, dialecto y otras características como las disfluencias y los desórdenes del habla. El LIS actualmente está llevando adelante un Proyecto de Investigación y Desarrollo financiado por el Ministerio de Ciencia, Tecnología e Innovación Productiva, en el cual la institución adoptante es la Gendarmería Nacional Argentina. El proyecto prevé orientar a la institución adoptante acerca de los protocolos aceptados para el diseño de bases de datos de hablantes para la construcción de una base de referencia de Argentina en forma directa y por medio del canal telefónico y proveerla de entrenamiento, recursos humanos capacitados y desarrollos de software de aplicación forense. En el ITBA se está desarrollando una herramienta tecnológica que, empleando metodologías del estado del arte en verificación de hablantes, permita utilizar las bases de datos universales de manera eficiente, flexible y simple para la elaboración de diferentes estrategias de comparación a ser empleadas en el ámbito forense. En esta casa de estudios también se ha comenzado a brindar una diplomatura en biometría que incluye el reconocimiento de locutor.

VI. CLASIFICACIÓN DE LOS SISTEMAS DE IDENTIFICACIÓN FORENSE DE HABLANTES

Se pueden clasificar las diferentes metodologías empleadas para la identificación forense de hablantes en:

1) Espectrográfico-auditivos
2) Auditivo-perceptuales
3) Fonético-lingüístico
4) Automáticos
5) Semi-automáticos
6) Combinados

1) Análisis espectrográfico-auditivo o huella de voz

Inicialmente se comenzó empleando exclusivamente el análisis espectrográfico, pero a partir de los 1970s se adicionó el análisis auditivo, quedando conformado el actual enfoque espectrográfico-auditivo que consiste en escuchar las grabaciones correspondientes al delincuente y al sospechoso (y dependiendo del protocolo también de otros hablantes de contraste), y mirando los espectrogramas respectivos, tomar una decisión en base a este examen visual y auditivo.

Ha existido un gran debate sobre si esta metodología es apropiada, principalmente porque todo el proceso se basa en el juicio subjetivo del experto que realice la identificación. Los defensores de la misma esgrimen que la experiencia de sus expertos, validada por la gran cantidad de espectrogramas que han analizado en su vida profesional, los trabajos pioneros de Potter [43] y Kersta [21] y la investigación de laboratorio llevada adelante por Tosi *et al.* [44] demuestran que su metodología es apta para ser empleada en el ámbito forense. En base a los resultados de éstos últimos han declarado en las cortes que su sistema tiene una precisión del 99% y que la huella de voz puede equipararse a la huella dactilar como recurso para proveer evidencia confiable. Los defensores de esta metodología se reúnen en la Asociación Internacional para la Identificación (*IAI – International Association for Identification*).

Por otra parte varios investigadores han refutado el trabajo realizado por el grupo de Tosi. Stevens *et al.* [22] compararon la habilidad de los sujetos para reconocer un hablante por medio del análisis espectrográfico y por el auditivo, y encontraron que la tasa de error que ellos pudieron medir era de entre el 21% y el 47%, muy superior al declarado en los trabajos de Kersta. Young y Campbell [45] también contrastaron el experimento de Kersta y hallaron valores de tasa de error muy superiores a los reportados. En su trabajo, Hazen [46] reportó tasas de error de entre 12% y 57%, muy por encima de los valores de Kersta y Tosi. Endres *et al.* [5] encontraron que los patrones espectrográficos y los niveles de la frecuencia fundamental de un mismo hablante variaban substancialmente a lo largo del tiempo y que en los casos de enmascaramiento la estructura de los formantes también mostraba cambios muy importantes. En un trabajo posterior, Reich *et al.* [47] analizaron la influencia del enmascaramiento en la identificación de hablantes basada en espectrogramas, y reportaron tasas de error que variaban entre 50% y 78%. Es de notar que ningún investigador neutral ha logrado una tasa de error cercana al 1%, predicha por Kersta, Tosi y sus seguidores, ni siquiera en condiciones de laboratorio.

La recomendación del comité de expertos convocado por la Academia Nacional de Ciencias de los Estados Unidos de América durante el año 1976 a pedido del FBI [24] fue que no

se empleara la metodología espectrográfica-auditiva en los casos que se presentaran en la corte, pero que se la podría emplear en la etapa de investigación. A pesar de este informe, la metodología es aceptada en varias cortes. Saks [48] reporta que en Estados Unidos de Norteamérica es aceptada en 6 estados, rechazada en 8, admisible en 4 cortes federales y excluida en 1. En una resolución sobre el uso de la huella de voz, la *IAFPA* en 2007, consideró que esta metodología no tiene fundamentos científicos, y por lo tanto no debería ser empleada en casos forenses [49].

2) Enfoque auditivo-perceptual

El análisis auditivo-perceptual se refiere principalmente a la participación de un ser humano que, empleando sus facultades auditivas y perceptuales, realiza la tarea de identificación de un hablante. Por lo tanto las características del oyente son, al menos, tan importantes como las del hablante y su entorno, al momento de analizar las variables que afectan la identificación.

Desde el punto de vista del hablante los problemas con los que se deberá enfrentar cualquier sistema de identificación en el ámbito forense son la singularidad y la distorsión.

La singularidad se refiere al cambio de paradigma propuesto por Saks y Koehler [26], que se preguntaron si las voces de los billones de personas en el mundo son únicas entre ellos y diferentes entre sí. Técnicamente este paradigma puede expresarse como la imposibilidad de conocer si la variabilidad intra-hablante es siempre menor o igual a la variabilidad entre-hablantes, y si esta relación es verdadera para todas las situaciones y bajo todas las circunstancias. Para acotar este problema en el ámbito forense y no tener que comparar la voz del hablante desconocido con respecto a toda la población mundial, se hace uso de bases de datos restringidas. Si el hablante desconocido es un hombre, se restringe la base de datos universal a los de ese género, lo mismo si es un niño o un anciano, o si habla en una lengua o dialecto en particular, y así sucesivamente. Es así como se llega a trabajar con un número razonable de potenciales hablantes, entre los cuales debería estar representado el hablante desconocido. A este punto es al que se refiere Morrison [41] cuando propone el "*uso de bases de datos representativas de la población relevante que reflejen las del caso bajo investigación*".

Los principales problemas debidos a la distorsión son debidas al sistema y al hablante. La distorsión del sistema incluye diferentes tipos de degradación de la señal. Uno es la reducción de la respuesta en frecuencia, viéndose afectado el ancho de banda cuando: a) alguien habla a través de una línea telefónica, b) el sistema de grabación o almacenamiento es de baja calidad, y c) se emplea un micrófono de limitada capacidad. El segundo es el ruido que puede enmascarar la voz del hablante y oscurecer elementos necesarios para su identificación. Ejemplos de ruido que podemos encontrar en los casos forenses pueden ser los generados por el viento, motores, movimiento de automóviles, fricción del micrófono con la ropa, música, otros hablantes, etc. La señal de ruido puede ser intermitente o estable, diente de sierra o térmico, etc. El tercero es la distorsión armónica o en frecuencia que puede generarse por cortocircuitos intermitentes, respuesta en frecuencia variable, distorsión armónica propiamente dicha, etc. Una forma de reducir el efecto de la distorsión del sistema es aplicando diferentes procedimientos, entre los cuales son útiles el filtrado y otras técnicas relacionadas. Pero debe tenerse mucho cuidado al aplicarlos de no eliminar parámetros de la señal que sean de utilidad para la discriminación de hablantes.

La distorsión debida a los hablantes puede provocarse por diferentes causas. Entre las principales se encuentran las debidas a: a) emociones (e.g. stress, cólera, tristeza, depresión, excitación, euforia, y alegría), b) estados inducidos por agentes externos (e.g. intoxicación por drogas o alcohol), c) comportamientos intencionales (e.g. engaño, falsedades, enmascaramiento, insolencia, prevención), d) estados de salud (e.g. resfrío, gripe, fatiga), y e) envejecimiento del hablante. En los casos de intoxicación o stress la reducción del efecto de la distorsión del hablante es muy difícil de lograr [50]. Sin embargo, el conocimiento de las diferentes causas y sus efectos en la señal de habla a analizar permiten valorizar la confiabilidad de los resultados obtenidos.

Ahora se analizarán los problemas que se encuentran desde el punto de vista del oyente, que es quien debe determinar la identidad del hablante desconocido. El primer factor a tener en cuenta es el tipo de oyente involucrado, que puede ser: a) un miembro del público, b) una persona interesada en el proceso, pero que es relativamente iniciado en el tema (estudiante), y c) un experto entrenado (generalmente un fonetista forense). Según el trabajo comparativo realizado por Hollien y Schwartz [51], la tasa de identificación correcta de los expertos fonetistas fue un 21% superior a la de los estudiantes, al actuar ambos como oyentes y empleando la metodología auditivo-perceptual. El segundo factor a tener en cuenta es la capacidad humana de recordar voces (memoria). Para ello se debe diferenciar si la voz es conocida o desconocida. En el caso de voces desconocidas, McGeehe [52] reportó obtener una tasa de reconocimiento de voces del 83% (correcto) después de un lapso de un día y hasta una semana, entre que se oyó la voz y se la reconoció. A las 2 semanas la tasa se redujo al 68%, a los 3 meses era del 35% y a los 5 meses el oyente solo podía identificar correctamente al 13% de los hablantes. El tercer factor a tener en cuenta es la capacidad del oyente para identificar muestras de habla del mismo hablante no-contemporáneas. La no-contemporaneidad se refiere a muestras de habla de un mismo hablante extraídas en distintos instantes de tiempo, las cuales serán posteriormente usadas en un proceso de identificación. Rothman [53] llevó adelante un experimento en el que comparó voces del mismo hablante grabadas con una diferencia de una semana, empleando el método auditivo-perceptual. Sus resultados fueron totalmente inesperados al encontrar que la tasa de reconocimiento promedio era de solo el 42%. En 1995 Schwartz [54] realizó un experimento para corroborar los datos obtenidos por Rothman y obtuvo resultados similares, donde la tasa de identificación correcta para datos contemporáneos alcanza valores del 95%, reduciéndose a las cuatro semanas en casi un

25%. Es de notar que entre las cuatro semanas y los cinco años la tasa se mantiene en un entorno acotado, produciéndose una disminución abrupta de un 50% adicional a los veinte años, la cual podría estar debida al fenómeno de envejecimiento. El envejecimiento produce cambios fisiológicos del hablante que se manifiestan en la voz del adulto de varias formas: disminución de la frecuencia fundamental, cambios en el timbre o calidad tonal, junto a una creciente inestabilidad del tono y la intensidad del habla, incremento de la respiración, y disminución de la precisión articulatoria [55]. Se definen, en base a estos datos, y según la tasa de identificación y el período de tiempo entre muestras, tres estadios denominados: a) contemporáneas (entre 0 y 2 semanas), b) no-contemporáneas (entre 2 semanas y 10 años), y c) envejecimiento (mayor a 10 años).

En el campo de la identificación forense de hablantes [31], normalmente transcurre un largo tiempo entre la evidencia y la grabación de la voz del sospechoso (e.g. en el caso del destripador bromista [56], la brecha fue de 27 años entre la "escena del crimen" y la grabación de custodia). En tales casos, indudablemente se debe tener en cuenta el envejecimiento del hablante.

En otro experimento Schwartz comparó la capacidad de identificación de hablantes cuando se introducen distractores de voces muy similares (padres, hermanos, hijos, etc.) entre dos grupos de oyentes. Los resultados mostraron que la tasa de identificación se mantiene aún con estos distractores, siempre y cuando las muestras sean contemporáneas. En cambio, si las muestras son no-contemporáneas la tasa de identificación cae a valores similares a los obtenidos por Rothman.

Podemos concluir que si las muestras son contemporáneas (menores a 2 semanas) puede obviarse el fenómeno de no-contemporaneidad, en cambio si las mismas superan ese valor deberá tenerse en cuenta que los resultados pueden incluir un error del orden del 15-20% debido a este fenómeno, el cual puede incrementarse un 50% adicional a partir de los 10 años debido al envejecimiento del hablante. En los casos forenses es poco usual encontrar períodos de tiempo entre muestras de habla del orden de las décadas. En cambio no es inusual trabajar con muestras con diferencias de aproximadamente un año.

En un experimento actual Künzel [57] analizó muestras de hablantes grabadas con intervalos de 11 años y encontró que para la mayoría de los hablantes la no-contemporaneidad de las muestras no era un problema para la identificación correcta. Los experimentos que fueron realizados con el método auditivo-perceptual y con un sistema automático de reconocimiento de hablantes, mostraron resultados similares.

El cuarto factor que se estudió es la influencia del género del oyente en los resultados de identificación correcta. En el trabajo realizado por Hollien *et al.* [51] no encontraron diferencias entre los resultados obtenidos por hombres y mujeres actuando como oyentes y empleando el método auditivo-perceptual.

El quinto factor a tener en cuenta es la familiaridad del oyente con el hablante a identificar. Se considera que una voz es familiar cuando se la ha escuchado frecuentemente por un período de tiempo de al menos dos años. Hollien *et al.* [58] realizaron un experimento auditivo-perceptual para mostrar el comportamiento de oyentes familiarizados y no, con la voz a reconocer, y en diversas condiciones del hablante (normal, con stress y enmascarando su voz). En la Fig. 1 pueden verse los resultados del mismo, notándose que el oyente familiarizado con el hablante obtiene en todos las condiciones tasas de identificación muy superiores a las de los oyentes que desconocen al hablante. Esta característica particular de los oyentes debe tenerse en cuenta especialmente en los casos forenses de ruedas de voces.

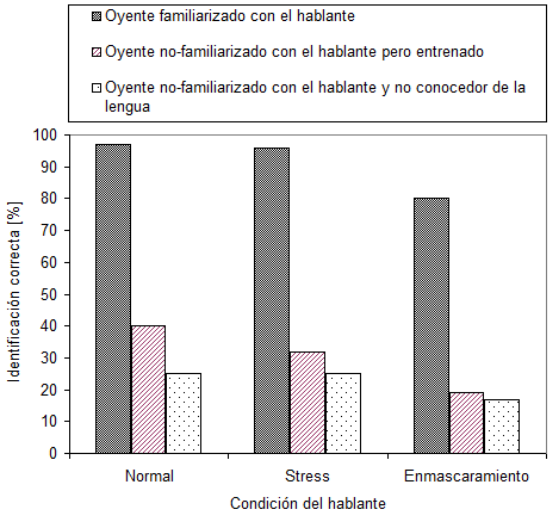
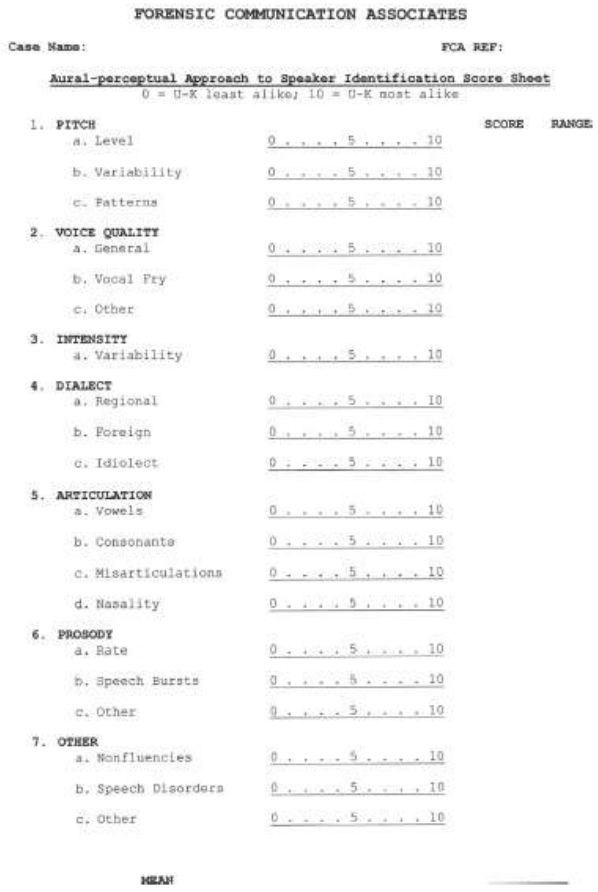


Figura 1. Tasa de reconocimiento correcta de hablantes bajo las condiciones de habla: a) normal, b) bajo stress, y c) de enmascaramiento, realizada por tres grupos de oyentes. Adaptado de Hollien [43]

El enfoque estructural propuesto Hollien [43] para implementar la metodología auditivo-perceptual incluye parámetros auditivos, fundamentalmente suprasegmentales. El método propuesto comienza con la toma de muestras, el tipo de material de habla requerido, la preparación de los pares de voces a identificar, para finalmente finalizar con un procedimiento estructurado de comparación de las pares de voces a identificar, empleando el test ABX [59] de discriminación de pares. En la Fig. 2 puede verse el formulario que debe completar cada oyente con cada par de voces a identificar, en el cual debe ingresar un valor del 0 al 10 (0 cuando las voces son percibidas como totalmente diferentes y 10 cuando se perciben iguales). Los parámetros empleados son: tono, calidad de voz, intensidad, dialecto, articulación, prosodia, y otros (disfluencias y desórdenes particulares en el habla). La comparación se realiza de a un parámetro a la vez y se extrae un valor medio y una variabilidad final. Un valor medio alto (en el rango de 7-10), con una baja variabilidad (± 2), permite asegurar que ambas voces pertenecen al mismo hablante, mientras valores medio bajos (en el rango de 0-3), y con una baja variabilidad (± 2), nos estarían confirmando que ambas voces pertenecen a diferentes personas. Valores medio intermedios o variabilidades muy altas no permitirían afirmar ninguna de ambas aseveraciones.



FORENSIC COMMUNICATION ASSOCIATES

Case Name: FCA REF:

Aural-perceptual Approach to Speaker Identification Score Sheet

0 = U-K least alike; 10 = U-K most alike

		SCORE	RANGE
1. PITCH			
	a. Level	0 . . . . 5 10	
	b. Variability	0 . . . . 5 10	
	c. Patterns	0 . . . . 5 10	
2. VOICE QUALITY			
	a. General	0 . . . . 5 10	
	b. Vocal Fry	0 . . . . 5 10	
	c. Other	0 . . . . 5 10	
3. INTENSITY			
	a. Variability	0 . . . . 5 10	
4. DIALECT			
	a. Regional	0 . . . . 5 10	
	b. Foreign	0 . . . . 5 10	
	c. Idiolect	0 . . . . 5 10	
5. ARTICULATION			
	a. Vowels	0 . . . . 5 10	
	b. Consonants	0 . . . . 5 10	
	c. Misarticulations	0 . . . . 5 10	
	d. Nasality	0 . . . . 5 10	
6. PROSODY			
	a. Rate	0 . . . . 5 10	
	b. Speech Bursts	0 . . . . 5 10	
	c. Other	0 . . . . 5 10	
7. OTHER			
	a. Nonfluencies	0 . . . . 5 10	
	b. Speech Disorders	0 . . . . 5 10	
	c. Other	0 . . . . 5 10	

MEAN

Figura 2. Formulario empleado con la metodología auditivo-perceptual, desarrollado por Hollien [43].

La validez y confiabilidad de esta metodología se estableció por medio de una serie de evaluaciones de laboratorio y de casos forenses reales, con datos recolectados durante 23 años. El nivel de confianza para los casos de laboratorio resultó ser del 81-88%, mientras que para los casos de campo el porcentaje se redujo al 77-79%. En la mitad de los casos reales se suplementó el análisis auditivo-perceptual con el análisis acústico, el cual extrae de la señal de habla diferentes parámetros acústicos útiles para la identificación de hablantes por medio de un experto.

3) Enfoque fonético-lingüístico

Una primera distinción que propuso Rose [60] cuando se habla de parámetros fonético-forenses es si son acústicos o auditivos, y luego si son lingüísticos o no-lingüísticos (estos conceptos se definirán más adelante). Con lo cual se puede clasificar un parámetro empleado para realizar una comparación de voces en el ámbito forense como: acústico-lingüístico, acústico-no-lingüístico, auditivo-lingüístico y auditivo-no-lingüístico. En el caso de los parámetros acústicos otra clasificación que suele utilizarse es si son de corto o largo plazo.

A pesar que las grabaciones que se emplean en el ámbito forense pueden contener risas, llantos, toses, o gritos, la mayoría pertenecen a la categoría de habla. El habla es el medio más común en el cual el lenguaje se realiza (la escritura es el otro). El lenguaje es un complejo código que enlaza el significado que el hablante quiere transmitir con los sonidos que produce. Los parámetros lingüísticos se agrupan en: fonológicos, morfológicos y sintácticos. En cambio los parámetros no-lingüísticos se caracterizan por no influir en la organización o producción de los sonidos de habla específicos. Como ejemplos de parámetros no-lingüísticos, al menos para el Español, se pueden citar: la calidad de la voz (nasalizada, ronca, áspera, susurrada, etc.), la velocidad de habla, el tono y la intensidad.

La metodología fonético-lingüística empleada en el ámbito forense propone emplear los diferentes tipos de parámetros descritos anteriormente. Es decir, emplea tanto parámetros acústicos como auditivos, y tanto parámetros lingüísticos como no-lingüísticos. Para determinar esos parámetros se emplean diferentes tipos de análisis.

El análisis auditivo emplea las características discriminatorias de un oyente para: realizar la transcripción fonética y fonémica (generalmente recibe de la corte las transcripciones ortográficas junto a las grabaciones de las voces), separar los segmentos de habla de los extra-lingüísticos, determinar el tono y la calidad de la voz, el sociolecto, idiolecto y etnolecto, y el enmascaramiento por parte del hablante, entre otros.

El análisis acústico extrae de la señal de habla diferentes parámetros acústicos útiles para la identificación de hablantes por medio de un experto. Esta tarea se realiza generalmente por medio de un computador. Los parámetros generalmente empleados son las frecuencias y el ancho de banda de los formantes de las regiones sonoras del habla (mayormente en las vocales), la frecuencia fundamental y su distribución de largo plazo (media, mediana, desviación estándar, etc.), el espectro promedio de largo plazo (*LTAS –Long-Term Average Spectrum*), y el espectrograma, entre otros. En las Fig. 3 a 6 pueden verse gráficas de los parámetros expuestos.

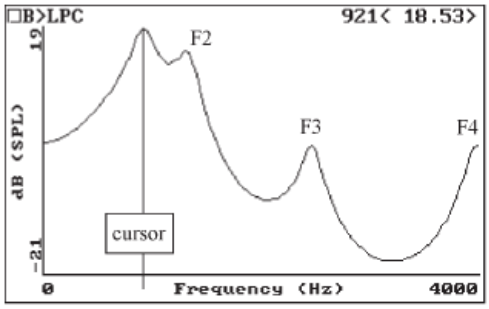


Figura 3. Análisis de formantes empleando el espectro de predicción lineal (*LPC*) para una vocal emitida por un hablante masculino [60].

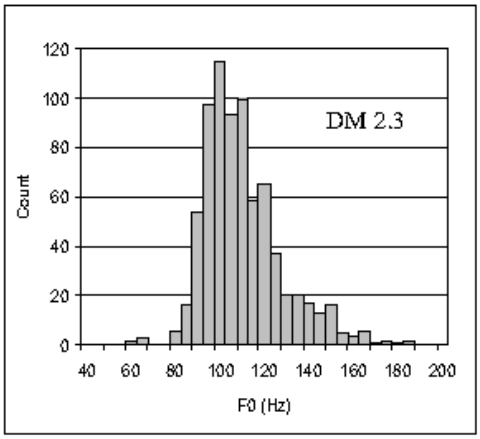


Figura 4. Histograma de la frecuencia fundamental de largo plazo de un hombre australiano adulto durante la lectura de un pasaje leído de 30 segundos de duración [60].

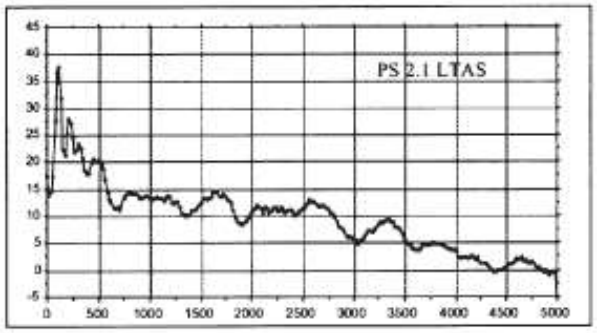


Figura 5. Espectro de largo plazo (*LTAS*) de un hombre australiano adulto durante la lectura de un pasaje leído de 30 segundos de duración [60].

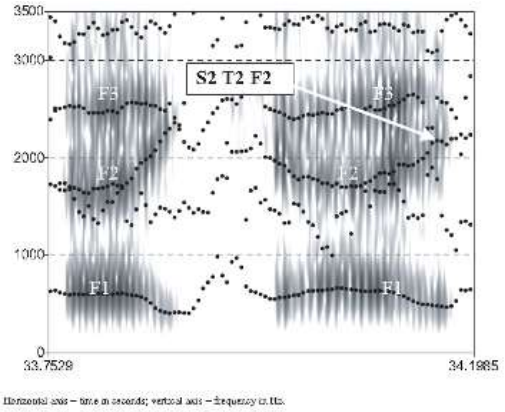


Figura 6. Espectrograma de la palabra <okay> donde se muestra el parámetro S2 T2 F2, donde S2: segunda sílaba, T2: objetivo del segundo diptongo, y F2: segundo formante [60].

Otra postura alternativa es la de French y Harrison que propusieron separar las muestras de habla recolectadas en sus constitutivos fonéticos y acústicos (e.g. calidad de voz, entonación, ritmo, tempo, velocidad de habla, realizaciones de vocales y consonantes) y analizarlas cada una en forma separada [28]. Dentro de los parámetros considerados para la comparación de voces se proponen los siguientes:

a. Calidad de voz: análisis de la zona estable de las vocales considerando hasta 38 elementos individuales.

b. Entonación: análisis del tono en el núcleo, la cabeza y la cola de las unidades entonativas.

c. Tono: medición del valor medio y la variación de la frecuencia fundamental.

d. Velocidad de habla.

e. Parámetros de ritmo.

f. Procesos de habla concatenada: patrones de asimilación y elisión.

g. Parámetros consonánticos: energía de las fricativas y ruidos de las oclusivas (*loci*), duración de las nasales, líquidas y fricativas en entornos fonológicos específicos, tiempo de sonorización de las oclusivas (*VOT – Voice Onset Time*), presencia/ausencia de pre-sonorización de las oclusivas, y variables sociolingüísticas.

h. Parámetros vocálicos: una gran cantidad de parámetros, entre los cuales se incluyen la configuración de los formantes, frecuencias centrales, densidades, anchos de banda, y variables sociolingüísticas.

i. Información lingüística de alto nivel: marcadores de discurso, elecciones lexicales, variantes morfológicas y sintácticas, comportamientos pragmáticos (e.g. sesión de turno, hábitos al contestar un llamado telefónico, aspectos de comportamiento multilingüal como alternancias)

j. Evidencias de impedimentos en el habla, la voz o patologías del lenguaje.

k. Parámetros no-lingüísticos: ronquera audible, carraspeo, ruidos de lengua, y fenómenos de vacilación como silencios prolongados y disfluencias.

4) Métodos automáticos

A partir de los desarrollos en el reconocimiento automático de hablantes en la última década [34], mayormente orientados al área comercial, hay un necesidad creciente de distinguir si es apropiado o no su uso en el área forense [61]. Los resultados impresionantes que han logrado este tipo de sistemas se ha visto reflejado en las evaluaciones de sistemas de reconocimiento de hablantes (*SRE*) que desarrolla el Instituto Nacional de Estándares y Tecnología (*NIST*) de los Estados Unidos de Norteamérica [62]. Dentro de los principales avances se puede destacar la aparición de una nueva técnica para modelar las variaciones de las muestras de habla entre sesiones, como el análisis factorial (*FA – Factor Analysis*) [63, 64], el análisis factorial conjunto (*JFA*) [16], la proyección de atributos perjudiciales (*NAP – Nuisance Attribute Projection*) [66], las aproximaciones por vectores de factor total (*i-vectors*) [17], y el resurgimiento de las redes neuronales profundas (*DNN – Deep Neural Networks*) en reemplazo de los *GMM*. El uso de las redes neuronales profundas, que son perceptrones multicapa de gran cantidad de capas ocultas, están siendo analizadas con gran interés por la comunidad científica del procesamiento del lenguaje natural, y el reconocimiento de habla e imágenes, debido a la factibilidad de implementación de las mismas gracias al descubrimiento de nuevas técnicas de inicialización, al uso del

procesamiento paralelo y a la capacidad de procesamiento de la tecnología actual [67, 68].

Los niveles de resultados obtenidos permiten suponer factible la aplicación de los sistemas automáticos de reconocimiento de hablantes en el ámbito forense.

Las evaluaciones *SRE* de *NIST* comenzaron en 1996 y se vienen desarrollando anualmente, con pocas excepciones. El principal objetivo de las mismas es proveer un marco integrado para evaluar científicamente los enfoques y sistemas en el campo del reconocimiento de hablantes: los participantes trabajan sobre un mismo corpus y protocolo, el mismo criterio de evaluación, y con un límite temporal acotado y sincronizado. Se han llegado a presentar hasta 40 laboratorios de diferentes países en la evaluación del año 2008. Una de las tareas, de mayor interés para el ámbito forense, es la que involucra el reconocimiento de hablantes independiente del texto y basada principalmente en segmentos de habla en conversaciones telefónicas de aproximadamente 2 minutos de duración promedio.

La metodología *GMM-UBM* fue la dominante en las evaluaciones independientes del texto [69] durante los primeros años del segundo milenio.

Una evolución de la metodología *GMM-UBM* es la incorporación de clasificadores basados en máquinas de soporte vectorial (*SVM – Support Vector Machine*), que generan supervectores *GMM-SVM* con funciones núcleo lineales (*GSL*) [70]. La metodología *GMM-UBM* se emplea para modelar los datos de entrenamiento y prueba. Cada grabación queda resumida en un supervector extractado del *GMM* correspondiente, compuesto por la concatenación de los coeficientes medios de todos los componentes del *GMM*. Los supervectores son luego utilizados como atributos de un clasificador *SVM* con una función núcleo (*kernel*) lineal basada en la distancia de Kullback-Leibler (*KL*) entre dos *GMMs*.

Durante la última década, una gran parte del esfuerzo en el campo del reconocimiento de hablantes se ha dedicado a las variaciones de las muestras entre las diferentes sesiones. Estas variaciones son debidas a diversos factores, exceptuando las variaciones intra-hablante: ambiente, micrófono, canal de transmisión; estado psicológico y patológico del hablante, contenido lingüístico, envejecimiento del hablante, etc. Por ejemplo, en las Figs. 7 y 8 se muestran como varía la tasa de error cuando existen diferencias de canal y de distancias al micrófono entre sesiones.

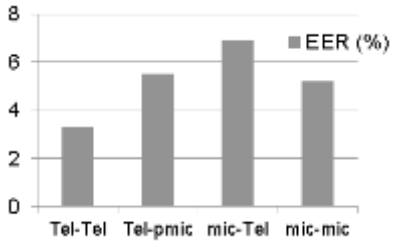


Figura 7. Tasa de igual error para diferentes combinaciones de canal entre sesiones: a) Tel-Tel: teléfono-teléfono, b) Tel-pmic: teléfono-micrófono profesional, c) mic-Tel: micrófono-teléfono, y d) mic-mic: micrófono-micrófono [71].

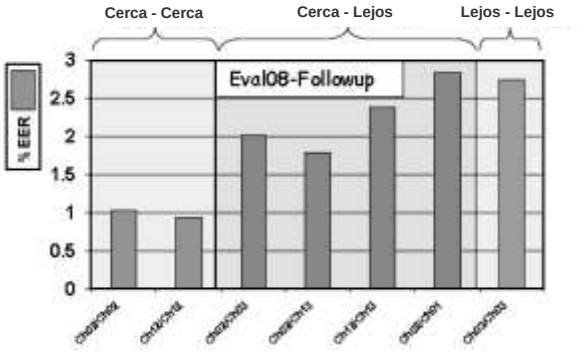


Figura 8. Tasa de igual error para diferentes combinaciones de la distancia del hablante con el micrófono entre sesiones: a) Cerca-Cerca, b) Cerca-Lejos, y c) Lejos-Lejos [71].

Los diferentes trabajos proponen soluciones parciales referidas a los primeros factores detallados anteriormente. Algunos se enfocan en el nivel acústico [72] y otros en el nivel de los resultados [73]. Más recientemente se han propuesto dos nuevos enfoques, uno basado en el análisis factorial (*FA*), dentro del paradigma *GMM* [63, 64], y otro en el marco de los clasificadores *SVM* empleando la proyección de atributos perjudiciales (*NAP*) [66]. La idea común entre ambos es la de modelar directamente las variaciones entre sesiones, antes que compensar sus efectos. Este modelado requiere el uso de grandes bases de datos con, por ejemplo, varias grabaciones de un mismo hablante utilizando diferentes micrófonos. Así como este ejemplo se focaliza en las variaciones de las sesiones debidas a diferentes micrófonos, lo mismo se debería hacer con el resto de los factores que las afecten. Se ha de notar que, para ambos casos *FA* y *NAP*, el problema subyacente se encuentra dentro del espacio del supervector, con una dimensión muy grande (de más de 25.000 en algunos casos).

En el año 2009 Dehak *et al.* [74] presentaron un nuevo método que combinaba el análisis factorial conjunto (*JFA*) con las máquinas de soporte vectorial (*SVM*). El cual consistía en usar directamente los factores del hablante estimados por el *JFA* como entradas al *SVM*. Posteriormente propusieron un nuevo sistema de reconocimiento de hablantes basado en el análisis factorial como extractor de parámetros [17]. El análisis factorial es empleado para definir un nuevo espacio de baja dimensionalidad denominado espacio de variabilidad total (del orden de 500). En este nuevo espacio, una emisión de habla es representada por un nuevo vector denominado de factor total o *i-vector*. La compensación del canal se realiza en este nuevo espacio de dimensiones reducidas, en contraposición con el espacio de alta dimensionalidad de los supervectores *GMM* que utilizan el clásico análisis *JFA*. El resultado de la identificación del hablante se obtiene de la comparación de los dos *i-vectors* que representan al hablante desconocido y al sospechoso. Uno de los métodos que ha reportado muy buenos rendimientos es el que utiliza la distancia coseno entre los *i-vectors* normalizados y las máquinas de soporte vectorial como clasificadores. Para compensar la variabilidad entre sesiones se han propuesto diferentes técnicas de normalización de los *i-vectors*. Entre las

principales metodologías empleadas se encuentra la normalización *EFR* (*Eigen Factor Radial*) [75], la normalización esférica de atributos perjudiciales (*sphNorm – Spherical Nuisance Normalization*) [76], la normalización de longitud (*lenght norm*), que es un caso particular de *EFR*, la normalización de la covarianza entre-clases (*WCCN – Within-Class Covariance Normalization*) [77], el análisis discriminativo lineal (*LDA – Linear Discriminative Analysis*) [78, 75] y el análisis discriminativo lineal probabilístico (*PLDA*) [18]. Los sistemas basados en *i-vectors/PLDA* son considerados en la actualidad el estado del arte y sus diferentes variantes son las que se disputan la supremacía en las últimas evaluaciones *NIST-SRE*.

A pesar de los buenos resultados obtenidos, son varios los factores que afectan la variabilidad del rendimiento de los sistemas automáticos de reconocimiento de hablantes. Algunos dependen de los hablantes y otros no, mientras que otros factores son difíciles de aislar. Uno de esos factores, que también afecta a las metodologías no-automáticas, es la no-contemporaneidad de las muestras, tema de investigación en curso. Otro factor es la duración y la cantidad de las muestras de entrenamiento que al aumentar producen una reducción en la tasa de error de reconocimiento. Un factor al que debe prestarse atención cuando se intenta utilizar un sistema automático en el ámbito forense, es la selección de las muestras de entrenamiento. Se ha visto que, para el caso de sistemas basados en *GMM*, la tasa de error varía cuando se intercambian los datos de entrenamiento por los de prueba para un mismo hablante [79]. Los factores de variación también tienen un impacto muy importante en el rendimiento de los sistemas, para el caso de sistemas del tipo *GSL-FA*, llegando a encontrarse incrementos de la *EER* de un factor de hasta 9 veces entre dos condiciones diferentes [61]. En cambio las tasas de error de los sistemas actuales basados en *i-vectors*/*PLDA* resultan más estables frente a las variaciones en las muestras de entrenamiento [62, 65].

Un enfoque contrapuesto a la problemática de las variabilidades entre sesiones es el empleado por Morrison [80] en su laboratorio de identificación forense de hablantes. En lugar de intentar modelar las variaciones entre sesiones (como en las metodologías de supervectores e *i-vectors*) propone compensar sus efectos, para lo cual trabaja con bases de datos que contengan múltiples grabaciones no-contemporáneas de un gran número de hablantes, con cada hablante grabado con múltiples estilos de habla y en diferentes ocasiones. Posteriormente selecciona las grabaciones que mejor se asemejan a los estilos del hablante desconocido y al sospechoso (e.g. conversación telefónica y entrevista) del caso en particular en el que está trabajando. El caso es rechazado si en su laboratorio no posee ese tipo de estilo de habla (e.g. no tiene base de datos de habla susurrada). La selección de las grabaciones es realizada por medio de un procedimiento que hace uso de un panel de oyentes. La base de datos está compuesta por grabaciones en diferentes lenguas y dialectos, las cuales han sido realizadas con equipamiento de alta calidad. Posteriormente son procesadas para reflejar las condiciones de grabación del hablante desconocido y del sospechoso, por ejemplo adicionándole reverberación y ruido, o haciendo pasar la señal por diferentes canales (telefonía de línea fija, telefonía móvil, etc.) u otro tipo de simulación. Las bases así seleccionadas y procesadas pueden ser empleadas por cualquiera de los métodos de identificación forense de hablantes habituales.

Otro problema al que debe prestarse atención al emplear un sistema automático de reconocimiento de hablantes es el error de calibración. A pesar que un sistema presente una tasa de error muy baja, el resultado realmente producido está sujeto a diferentes niveles de variación. Estas variaciones pueden producirse por la selección del punto de trabajo utilizado, el cual generalmente se determina por medio de un procedimiento de calibración con datos de hablantes conocidos. Por lo tanto es necesario contar con suficientes datos, similares a los empleados durante el entrenamiento del sistema. En algunos usos de los sistemas automáticos de reconocimiento de hablantes este procedimiento puede ser adecuado, como en el ámbito comercial, pero puede ser muy difícil obtener una buena calibración en muchos casos forenses donde no coincide el proceso de recolección de datos de entrenamiento e identificación. Se han propuesto diferentes técnicas para mitigar este problema, las cuales se basan en obtener distribuciones de resultados más predecibles que permitan calibrar los sistemas de tal manera que puedan generalizar a condiciones no vistas durante el entrenamiento de calibración.

Entre los mismos se encuentra la compensación de parámetros de clasificación, como los diferentes enfoques basados en la metodología de sustracción de la media cepstral (*Cepstral Mean Subtraction*) [81, 82], y el filtrado espectral relativo (*RASTA*) [83], que buscan remover los efectos residuales de las diferencias introducidas en la señal del habla por el canal de audio; el modelado de parámetros (*FA* y *NAP*), y las diferentes normalizaciones de resultados (*Z-norm*, *T-norm*, etc.) [12, 73]. Se han obtenido mejoras sustanciales de la tasa de error en las evaluaciones *NIST*, pero la calibración de esos sistemas varía dramáticamente cuando se analizan diferentes tipos de micrófonos. Por ejemplo en la evaluación *NIST-SRE-2008* presentada por el Instituto Tecnológico de Massachusetts (*MIT*), en un caso de diferentes micrófonos, se obtuvo una mejora porcentual del 5% entre el costo de detección C_{Det} mínimo y el real; sin embargo en otro caso se observó una diferencia del 160%. Para ambos casos la tasa de igual error era menor al 2%. Estos resultados muestran que la calibración en diferentes condiciones es aún un tema de investigación y afecta de manera muy importante el uso de los sistemas automáticos de reconocimiento de hablantes, especialmente en el ámbito forense.

Una de las problemáticas en el desarrollo de los sistemas automáticos es el uso de la tasa de igual de error (*EER*) como único criterio de evaluación. Los profesionales de la ingeniería y la computación se enfocan en la reducción de este parámetro sin tener en cuenta otros criterios que podrían ayudar a determinar la validez y confiabilidad de los sistemas que deben emplearse en casos reales y particularmente en el rubro forense. Bonastre *et al.* [84] presentaron resultados alarmantes

de la degradación del rendimiento de un sistema de reconocimiento automático cuando el habla de varios impostores era manipulada por medio de técnicas de síntesis de habla para simular engaños, incrementándose la *EER* de 8,5% a 35,4%. Dado que se intentaba engañar a un sistema computacional, y no a un experto humano, las voces artificiales solo requerían ser percibidas como voces naturales.

De todo lo expuesto se puede concluir que los sistemas automáticos de identificación forense de hablantes deben aplicarse con mucho cuidado, considerando cada caso en particular, e incorporando otros ámbitos teóricos y analíticos a la mera área de la ingeniería: como por ejemplo la de los expertos fonéticos.

5) Métodos semi-automáticos

Son aquellos en los que es necesaria una fuerte y continuada interacción entre el analista y la aplicación de cálculo o análisis que utiliza. En estos sistemas, el operador desempeña un papel determinante tanto en la selección de elementos para la comparación, como en la interpretación de los resultados finales del proceso. Esta metodología, que es la más empleada en el ámbito forense que utiliza sistemas automáticos de identificación de hablantes, busca aprovechar las ventajas de los mismos con la incorporación del conocimiento de los expertos fonetistas.

El sistema propuesto por Hollien [43], denominado *SAUSI* (*Semi-automatic speaker identification*), utiliza cuatro vectores que han demostrado poseer características de discriminación entre hablantes, para un grupo numeroso de sujetos y en situaciones donde la distorsión está presente. Los vectores que emplea el sistema *SAUSI* son: a) el espectro de largo plazo (*LTS – Long-Term Spectra*), b) la frecuencia fundamental (*SFF – Speaking Fundamental Frecuency*), c) la distribución temporal de la energía (*TED – Time-Energy Distribution*), y d) la distribución de formantes en las vocales (*VFT – Vowel Formant Tracking*).

Otro sistema desarrollado en la Universidad del Estado de Carolina del Norte [85] se basaba en la comparación entre el hablante conocido y el desconocido empleando las mismas secuencias fonémicas (denominadas secuencias isofonémicas). Para ello es necesario que un experto determine las secuencias isofonémicas a comparar. Por ejemplo si se quisieran contrastar dos hablantes, uno que emitió la frase: "*In the heat of the night*", y el otro: "*By the seat of your pants*", es posible seleccionar las vocales de las palabras "*heat*" y "*seat*" como las secuencias isofónemicas a comparar.

En Polonia, Majewski conformó un grupo de trabajo [86] que desarrolló una metodología basada en métodos perceptuales y en el uso de sistemas computacionales que procesan una gran cantidad de parámetros acústicos: espectro, cepstrum, coeficientes *LPC*, tasa de cruce por cero (*ZCR - Zero Crossing Rate*), y parámetros temporales, entre otros. Luego aplican diferentes algoritmos no-paramétricos, entre los cuales se destacan los vecinos más cercanos (*NN – Nearest Neighbour*) con alineamiento temporal dinámico (*DTW – Dynamic Time Warping*) para la clasificación de las muestras procesadas. Para poder trabajar con sistemas de identificación abiertos, en los cuales debe existir la posibilidad de rechazar al hablante de prueba, desarrollaron un algoritmo denominado *OSA* (*Open Set Algorithm*). El cual, primero realiza una identificación cerrada y determina una identificación posible, luego verifica la misma con respecto al modelo competitivo elegido y acepta o rechaza la identificación original de acuerdo al resultado de dicha comparación.

Univaso *et al.* [87] desarrollaron un sistema semi-automático incorporando diferentes fuentes de información suprasegmental o prosódica a la información segmental (*GMM* y *i-vector*) por medio de herramientas de minería de datos.

Otros sistemas semi-automáticos presentes en la literatura son los siguientes: *SIVE* [88, 89, 90], *IDEM* [91], *DIALECT* [92], *VOCALISE* [93], y el método descrito y validado por el grupo de Amino *et al.* [94].

Actualmente los principales laboratorios de acústica forense emplean diferentes tácticas para implementar la metodología semi-automática, aunque debido al secreto con que normalmente trabajan no se posee información fehaciente de las mismas.

6) Métodos combinados

Conocidas las principales metodologías que sustentan la identificación forense del habla (espectrográfica, fonético-lingüística, semi-automática y automática) se presentará la que es considerada por la comunidad forense de vanguardia, la alternativa metodológica de mayor fiabilidad: los denominados métodos combinados. Dicha denominación, se deriva de la que es su característica más representativa, puesto que en todos los casos, y sea cual fuere la versión de los mismos, los métodos combinados vendrán siempre configurados por la conjugación de los cuatro sistemas básicos anteriormente descritos. Pero la utilización exclusiva de alguna de las cuatro aproximaciones generales citadas carece, en cualquier caso, del máximo rigor y eficacia que hoy en día debe exigirse a la técnica forense de identificación de hablantes. Cuando cualquier procedimiento científico aborda un objeto de estudio con carácter variable suele utilizar el mayor número de perspectivas de estudio posible para que las conclusiones alcanzadas posean un alto grado de objetividad. Por esta razón, se propone como el mejor acercamiento al problema la utilización de un método combinado, que permita utilizar un grupo de procedimientos complementarios pero de distinta o similar naturaleza.

En 1999, fue celebrada en Wiesbaden una reunión de expertos en acústica forense europeos, como consecuencia del proyecto de estandarización desarrollado para tal materia por el Grupo de Trabajo de Cooperación Policial (*PCWG - Police Cooperation Working Group*) de la Unión Europea. Meses después, tuvo lugar en Madrid el segundo encuentro del grupo de trabajo para habla y audio forense de la red *ENFSI*. En ambas reuniones, los métodos combinados de identificación forense de hablantes fueron señalados como los más funcionales, eficaces y fiables. Igualmente, fue consensuada la definición del ámbito y sistemas de análisis que quedan integrados dentro de los mismos: "*se consideran métodos*

combinados de identificación forense de locutores, aquellos que entre sus aproximaciones de estudio incluyen, al menos, el enfoque perceptivo-auditivo, el análisis acústico (sonográfico, oscilográfico, espectrográfico) y el análisis fonético-lingüístico". La aceptación institucional de los métodos combinados como los más idóneos resulta de extraordinaria importancia, ya que son el inicio del final formal del tradicional cisma metodológico entre fonetistas e ingenieros. Por otra parte, esta nueva concepción no excluye en ningún caso el complemento de análisis representado por los enfoques semi-automáticos y automáticos. Muy al contrario, el espíritu de los asistentes a las citadas reuniones se manifestaba totalmente abierto a este tipo de opciones, si bien, en el momento actual se considera que solamente pueden ser apreciadas como un elemento complementario a los sistemas referidos. Sin embargo, el hecho de estimar como mejor solución la utilización de una metodología combinada, no tiene otro significado que el de señalar cuales son los ingredientes básicos que no deben faltar en el desarrollo de los distintos modelos forenses de identificación de hablantes. A partir de aquí cada laboratorio o experto, en función de sus posibilidades, necesidades, entorno legal, etc. diseñan el modelo que estiman más adecuado.

Dentro de los pocos trabajos que comparan los diferentes métodos se encuentra el de Alexander *et al.* [95], quienes estudiaron un método automático basado en *GMM-UBM* y el enfoque auditivo-perceptivo llevado a cabo por oyentes no entrenados en fonética o identificación de hablantes. Las conclusiones a las arribaron fueron que cuando se tratan de muestras de sesiones coincidentes los sistemas automáticos mostraban rendimientos mucho mejores que los basados en reconocimiento auditivo-perceptual. Sin embargo, en condiciones de sesiones no-coincidentes, encontraron que el método automático presentaba resultados levemente inferiores a los basados en humanos. Otro resultado llamativo entre ambos sistemas es que los mejores rendimientos se encuentran con muestras de sesiones de canal telefónico de línea fija para el caso auditivo-perceptual, mientras que el sistema automático se comporta mejor con sesiones de la red celular. En el mismo trabajo se preguntó a los oyentes cuáles eran las características de las voces que tuvieron en cuenta al realizar la identificación. Las principales características nombradas, para todas las condiciones de canal, fueron parámetros no-lingüísticos: acento, timbre, entonación, velocidad de habla, y anomalías en el habla. Todas características robustas que pueden ayudar a la identificación de hablantes, especialmente en los casos forenses donde las condiciones están severamente degradadas.

Recientemente Hansen y Hasan [96] compararon el desempeño de los humanos con respecto a las máquinas, con la salvedad de que es muy difícil realizar tal comparación de una manera estadísticamente significativa, debido al reto que es para los humanos la evaluación confiable de grandes cantidades de emisiones. La mejora de los algoritmos actuales ha logrado que los sistemas automáticos posean un rendimiento superior al de los humanos. A pesar de ello, consideran que este resultado no puede ser considerado como una prueba definitiva. En muchas circunstancias donde la información paralingüística es importante o cuando se deben reconocer voces conocidas, los humanos pueden tener una performance mejor que la de las máquinas.

VII. CONCLUSIÓN

Como hemos mostrado se ha producido un gran avance en el reconocimiento de hablantes durante estos últimos años. Pero todavía no ha podido afianzarse como una tecnología de amplia utilización, aunque algunas aplicaciones comerciales la emplean; pero es especialmente en el ámbito forense donde mayor resistencia encuentra en comparación con otras técnicas biométricas de mayor confiabilidad como las de ADN y las huellas dactilares.

El uso de los sistemas automáticos de reconocimiento de hablantes en aplicaciones forenses debe realizarse con gran cautela y son mayormente empleados como medio de investigación o en sistemas semi-automáticos como apoyo del experto.

A pesar de los importantes avances tecnológicos en la materia y en contra de la visión simplista y superadora que nos transmiten algunas películas cinematográficas y series televisivas policíacas, podemos concluir que aunque se ha producido un gran avance en la identificación de hablantes durante estos últimos años, especialmente en algunas aplicaciones comerciales, es en el ámbito forense donde mayor resistencia encuentra en comparación con otras técnicas biométricas de mayor confiabilidad como las de ADN y huellas dactilares.

REFERENCIAS

[1] I. Pollack, J. M. Pickett and W. H. Sumby, "On the Identification of Speakers by Voice", *Journal of the Acoustical Society of America*, vol. 26, pp. 403-406, 1954.

[2] J. N. Shearme and J. N. Holmes, "An Experiment Concerning the Recognition of Voices", *Language and Speech*, 2, pp. 123-131, July/September 1959.

[3] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition", *Journal of the Acoustical Society of America*, vol. 35, pp. 354-358, 1963.

[4] G. R. Doddington, "A method of speaker verification", *Journal of the Acoustical Society of America*, vol. 49, 139 (A), 1971.

[5] W. Endres, W. Bambach and G. Flösser, "Voice spectrograms as a function of age, voice disguise, and voice imitation", *Journal of the Acoustical Society of America*, vol. 49(6B), pp. 1842-1848, 1971.

[6] S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition", *Electronics and Communications in Japan*, 57-A, pp. 34-41, 1974.

[7] P. Univaso, E. Rosso and H. Franco, "Automatic Recognition of Isolates Spanish CV Syllables". In *111 th Meeting of the Acoustical Society of America*, Cleveland, Ohio, USA, 1986.

[8] J. M. Naik, L. P. Netsch and G. R. Doddington, "Speaker verification over long distance telephone lines", in *Proc. ICASSP Acoustics, Speech, and Signal Processing*, pp. 524-527, 1989.

[9] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation", in *Proc. ICASSP Acoustics, Speech, and Signal Processing*, pp. 293-296, 1990.

[10] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", in *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 27-30, 1994.

[11] J. Oglesby and J. S. Mason, "Optimization of neural models for speaker identification", in *Proc. ICASSP Acoustics, Speech, and Signal Processing*, vol. 1, pp. 261-264, 1990.

[12] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring", in *Proc. ICASSP, Acoustics, Speech, and Signal Processing*, vol. 1, pp. 595-598, 1998.

[13] G. R. Doddington, "Speaker recognition based on idiolectal differences between speakers", in *Proc. Eurospeech*, pp. 2521-2524, 2001.

[14] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification", in *Neural Networks Signal Process, Proc. 2000 IEEE Signal Processing Workshop*, Vol. 2, pp. 775-784, 2000.

[15] A. Stolcke, E. Shriberg, L. Ferrer, S. Kajarekar, K. Sonmez and G. Tur, "Speech Recognition as Feature Extraction for Speaker Recognition", in *Proc. of the IEEE Workshop on Signal Processing Applications for Public Security and Forensics*, pp. 39-43, 2007.

[16] P. J. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition", in *Audio, Speech, and Language Processing, IEEE Transactions on*, 15, 4, pp. 1435-1447, 2007.

[17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification", in *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788-798, 2011.

[18] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity", in *Computer Vision, ICCV 2007. IEEE 11th International Conference on*, pp. 1-8, 2007.

[19] S. Furui, "Acoustic and Speech Engineering", *Kindai Kagaku-sha Publishing Company*, Tokyo, 1992.

[20] National Research Council, "On the theory and practice of voice identification", National Academy of Science, Washington, pp. 3-13, 1979.

[21] L. G. Kersta, "Voiceprint identification", *Nature*, vol. 196, no. 4861, pp. 1253-1257, 1962.

[22] K. N. Stevens, C. E. Williams, J. R. Carbonell and B. Woods, "Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material", *Journal of the Acoustical Society of America, vol.* 44, pp. 1596-1607, 1968.

[23] D. L. Faigman, D. H. Kaye, M. J. Saks and J. Sanders, "Modern scientific evidence: Vol. 2", In *The law and science of expert testimony*, pp. 430-446, 2002

[24] B. Koenig, Federal Bureau of Investigation and US Dept of Justice, "Speaker Identification-Part 2-Results of the National Academy of Sciences' study", FBI Law Enforcement Bulletin, vol 49, no. 2, pp. 20-22, 1980.

[25] J. F. Bonastre, F. Bimbot, L. J. Boë, J. P. Campbell and D. A. Reynolds, I. Magrin-Chagnolleau, "Person Authentication by Voice: A Need for Caution", in *Proc. Eurospeech*, pp. 1-4, 2003.

[26] M. J. Saks and J. J. Koehler, "The coming paradigm shift in forensic identification science", in *Science*, 309, 5736, pp. 892-895, 2005.

[27] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection", *Computer Speech Language*, 20(2-3), pp. 230-275, 2006.

[28] P. French and P. Harrison, "Position Statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases, with a foreword by Peter French & Philip Harrison", *International Journal of Speech Language and the Law*, vol. 14, no. 1, pp. 137-144, 2007.

[29] P. Rose and G. Morrison, "A response to the UK position statement on forensic speaker comparison", *International Journal of Speech Language and the Law*, vol. 16, no. 1, pp. 139, 2009.

[30] P. French, F. Nolan, P. Foulkes, P. Harrison and K. McDougall, "The UK position statement on forensic speaker comparison; a rejoinder to Rose and Morrison", *International Journal of Speech Language and the Law*, vol. 17, no. 1, pp. 143-152, 2010.

[31] P. Rose, "Forensic Speaker Identification", *London: Taylor & Francis*, 2002.

[32] J. Gibbons and M. T. Turell (Eds.), "Dimensions of forensic linguistics", *Amsterdam/Philadelphia: John Benjamins Publishing*, 2008.

[33] C. Greenberg, A. Martin, L. Brandschain, J. P. Campbell, C. Cieri, G. R. Doddington and J. Godfrey, "Human assisted speaker recognition in NIST SRE10", submitted to special session on Human Assisted Speaker Recognition, *Proceedings of IEEE ICASSP*, Prague, 2011.

[34] D. A. Reynolds, "An overview of automatic speaker recognition", in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Washington, DC: IEEE Computer Society, pp. 4072-4075, 2002.

[35] National Institute of Standards and Technology, "2012 NIST Speaker Recognition Evaluation Results", May 15, 2013. Disponible: http://www.nist.gov/itl/iad/mig/sre12results.cfm. Consultado el 3 de octubre de 2015.

[36] R. Saeidi, K. A. Lee, T. Kinnunen, T. Hasan, B. Fauve, P. M. Bousquet, ... and E. Ambikairajah, "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification", 2013.

[37] National Institute of Standards and Technology, "2012 NIST Speaker Recognition Evaluation", May 30, 2012. Disponible: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf. Consultado el 1 de junio de 2014.

[38] M. Jessen, "The forensic phonetician", in *The Routledge Handbook of Forensic Linguistics*, pp. 378-394, 2010.

[39] R. Greisbach, "Estimation of speaker height from formant frequencies", in *Forensic Linguistics: The International Journal of Speech Language and the Law*, 6, 2, pp. 265-277, 1999.

[40] H. F. Hollien, "Forensic voice identification", in *Academic Press*, 2002.

[41] G. S. Morrison, "Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison", in *Science & Justice*, 54, 3, pp. 245-256, 2014.

[42] A. P. A. Broeders, "Forensic speech and audio analysis forensic linguistics", in *13th INTERPOL Forensic Science Symposium*, Lyon, France, 26, 2001.

[43] R. K. Potter, "Visible Patterns of Sound", in *Science*, 1945.

[44] O. Tosi, H. Oyer, W. Lashbrook, C. Pedrey, J. Nicoland E. Nash, "Experiment on voice identification", in *Journal of the Acoustical Society of America*, 51, 6B, pp. 2030-2043, 1972.

[45] M. A. Young and R. A. Campbell, "Effects of context on talker identification", in *Journal of the Acoustical Society of America*, 42, 6, pp. 1250-1254, 1967.

[46] B. Hazen, "Effects of differing phonetic contexts on spectrographic speaker identification", in *Journal of the Acoustical Society of America*, 54, 3, pp. 650-660, 1973.

[47] A. R. Reich, K. L. Moll and J. F. Curtis, "Effects of selected vocal disguises upon spectrographic speaker identification", in *Journal of the Acoustical Society of America*, 60, 4, pp. 919-925, 1976.

[48] M. J. Saks, "Merlin and Solomon: Lessons from the Laws Formative Encounters with Forensic Identification Science", in *Hastings Law Journal*, 49, 4, pp. 1069-1141, 1998.

[49] International Association for Forensic Phonetics and Acoustics: Resolution on voiceprints (2007). Disponible: http://www.iafpa.net/voiceprintsres.htm. Consultado el 24 de septiembre 2014.

[50] H. F. Hollien, R. Huntley Bahr and J. D. Harnsberger, "Issues in Forensic Voice", in *Journal of Voice*, 28, 2, pp. 170-184, 2014.

[51] H. F. Hollien and R. Schwartz, "Aural-perceptual speaker identification: problems with noncontemporary samples", in *Forensic Linguistics*, 7, pp. 199-211, 2000.

[52] F. McGehee, "The reliability of the identification of the human voice", in *The Journal of General Psychology*, 17(2), 249-271, 1937.

[53] H. B. Rothman, "A perceptual (aural) and spectrographic identification of talkers with similar sounding voices", in *International Conference on Crime Countermeasures-Science and Engineering*, pp. 37-42, 1977.

[54] R. Schwartz, "Effect of non-contemporary speech on aural–perceptual speaker identification", in *IAFP-95, Congress of the International Association of Forensic Phonetics*, Orlando, Florida, USA, 1995.

[55] F. Kelly, A. Drygajlo and N. Harte, "Speaker verification in score-ageing-quality classification space", in *Computer Speech & Language*, 27, 5, pp. 1068-1084, 2013.

[56] P. French, P. Harrison and J. Windsor-Lewis, "R v John Samuel Humble: The Yorkshire Ripper Hoaxer trial", in *The International Journal of Speech, Language and the Law*, 13, 2, pp. 256-273, 2006.

[57] H. J. Künzel, "Non-contemporary speech samples: auditory detectability of an 11 year delay and its effect on automatic speaker identification", in *International Journal of Speech Language and the Law*, 14, 1, pp. 109-136, 2007.

[58] H. Hollien, W. Majewski and E. T. Doherty, "Perceptual identification of voices under normal, stress and disguise speaking conditions", in *Journal of Phonetics*, 10, pp. 139-148, 1982.

[59] G. White and G. J. Louie, "The Audio Dictionary: Revised and Expanded", in *University of Washington Press* (2005)

[60] P. Rose, “The Technical Comparison of Forensic Voice Samples”. En Selby H. y Freckelton I. (eds.): *Expert Evidence 99*, in Thompson Lawbook Co., Sydney, 2003.

[61] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. F. Bonastre and D. Matrouf, “Forensic speaker recognition: A need for caution”, in *IEEE Signal Processing Magazine*, 26, 2, pp. 95-103, 2009.

[62] M. Przybocki and A. F. Martin, “NIST speaker recognition evaluation chronicles”, in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[63] P. J. Kenny and P. Dumouchel, “Disentangling speaker and channel effects in speaker verification”, in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, pp. 37-40, 2004.

[64] P. J. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, “Factor Analysis Simplified”, in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 1, pp. 637-640, 2005.

[65] P. J. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam and P. Dumouchel, “PLDA for speaker verification with utterances of arbitrary duration”, in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7649-7653, 2013.

[66] A. Solomonoff, W. M. Campbell and I. Boardman, “Advances In Channel Compensation For SVM Speaker Recognition”, in *ICASSP*, 1, pp. 629-632, 2005.

[67] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, ... and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups”, in *Signal Processing Magazine*, IEEE, 29, 6, pp. 82-97, 2012.

[68] A. R. Mohamed, G. E. Dahl and G. Hinton, “Acoustic modeling using deep belief networks”. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20, 1, pp. 14-22, 2012.

[69] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, ... and D. A. Reynolds, “A tutorial on text-independent speaker verification”, in *EURASIP journal on applied signal processing 2004*, pp. 430-451, 2004.

[70] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff, “SVM-based speaker verification using a GMM supervector kernel and NAP variability compensation”, in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, I, pp. 97–100, 2006.

[71] J. P. Campbell, “Speaker Recognition for Forensic Applications”, in *Odyssey Keynote: FSR-2*, 2014.

[72] D. A. Reynolds, “Channel robust speaker verification via feature mapping”, in *Acoustics, Speech, and Signal Processing*, 2003. *Proceedings.(ICASSP'03). 2003 IEEE International Conference on* , 2, pp. II-53, 2003.

[73] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems”, in *Digital Signal Processing*, 10, 1, pp. 42-54, 2000.

[74] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, ... and F. Castaldo, “Support vector machines and joint factor analysis for speaker verification”, in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4237-4240, 2009.

[75] P. M. Bousquet, D. Matrouf and J. F. Bonastre, “Intersession compensation and scoring methods in the i-vectors space for speaker recognition”, in *Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 485-488, 2011.

[76] P. M. Bousquet, A. Larcher, D. Matrouf, J. F. Bonastre and O. Plchot, “Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis”, in *Odyssey Speaker and Language Recognition Workshop*, 2012.

[77] A. O. Hatch, S. S. Kajarekar and A. Stolcke, “Within-class covariance normalization for SVM-based speaker recognition”, in *Interspeech*, 2006.

[78] K. Fukunaga, “Introduction to Statistical Pattern Recognition 2nd ed.”, in *NewYork: Academic Press*, ch.10, 1990.

[79] S. E. Mezaache, J. F. Bonastre and D. Matrouf, “Analysis of impostor tests with high scores in NIST-SRE context”, in *Interspeech 2008*, pp. 367-370, 2008.

[80] G. S. Morrison, P. Rose and C. Zhang, “Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice”, in *Australian Journal of Forensic Sciences*, 44, 2, pp. 155-167, 2012.

[81] D. Naik, “Pole-filtered cepstral mean subtraction”, in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95, 1995 International Conference on*, 1, pp. 157-160, 1995.

[82] A. A. García and R. J. Mammone, “Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping”, in *Acoustics, Speech, and Signal Processing, 1999. Proceedings 1999 IEEE International Conference on*, 1, pp. 325-328, 1999.

[83] H. Hermansky, N. Morgan, A. Bayya and P. Kohn, “RASTA-PLP speech analysis technique”, in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1, pp. 121-124, 1992.

[84] J. F. Bonastre, D. Matrouf and C. Fredouille, ”Artificial impostor voice transformation effects on false acceptance rates”, in *Interspeech*, pp. 2053-2056, 2007.

[85] R. Rodman, “Semi-automatic Speaker Recgnition of Disguised Voices”, invited paper presented at the *American Academy of Forensic Sciences*, Orlando, Florida, 1998.

[86] C. S. Basztura, W. Majewski and J. Jurkiewicz, “Automatic Voice Recognition in Open Sets”, in *Arch. Acoust.*, 13, pp. 205-218, 1978.

[87] P. Univaso, J. M. Ale and J. A. Gurlekian, ”Data Mining applied to Forensic Speaker Identification”, in *Latin America Transactions, IEEE (Revista IEEE America Latina)*, 13(4), 1098-1111, 2015.

[88] A. Lipeika and J. Lipeikiene, “The use of pseudostationary segments for speaker Identification”, in *Proceedings of the 3rd European Conference on Speech Communication and Technology*. Berlin, pp. 2303-2306, 1993.

[89] A. Lipeika and J. Lipeikiene, “Speaker identification using vector quantization”, in *Informatica 1995*, 6, 2, pp. 167-180, 1995.

[90] A. Lipeika, J. Lipeikiene and B. Salna, “On usefulness of the LPC residue in speaker Identification”, in *Proceedings of the International Conference Biomedical Engineering*, Kaunas 1997, pp. 61-64, 1997.

[91] M. Falcone and N. De Sario, “ A PC based speaker identification system for forensic use: IDEM”, in *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, pp. 169-172, 1994.

[92] H. F. Popov, A. N. Linkov and M. V. Baicharov, “Personal identification by Russian speech phonograms on the automatized system DIALECT: Manual for experts”, in *Fesenko A.F. (Ed) Military unit ni 34435*, Moscú, 1996.

[93] A. Alexander, O. Forth, M. Jessen and M. Jessen, “Speaker recognition with Phonetic and Automatic Features using VOCALISE software”, in 22th Annual Conference of the International Association for Forensic Phonetics and Acoustics, Tampa, 2013.

[94] K. Amino, T. Osanai, T., Kamada, H. Makinae and T. Arai, “Historical and procedural overview of forensic speaker recognition as a science”, in *Forensic speaker recognition, Springer New York*, pp. 3-20, 2012.

[95] A. Alexander, F. Bottib, D. Dessimozb and A. Drygajloa, “The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications”, in *Forensic Science International*, 146S, pp. S95-S99, 2004.

[96] J. H. Hansen and T. Hasan, “Speaker Recognition by Machines and Humans: A tutorial review”, in *Signal Processing Magazine, IEEE*, 32, 6, pp. 74-99, 2015.

Pedro Univaso nació en Buenos Aires, Argentina, el 4 de marzo de 1959. Se graduó en la Facultad de Ingeniería de la Universidad de Buenos Aires (UBA) como Ingeniero Electromecánico orientación Electrónica y es candidato al doctorado por la misma universidad. Es titular de BlackVOX, empresa de base tecnológica incubada en el Laboratorio de Investigaciones Sensoriales (LIS), perteneciente al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y a la UBA, de la cual es Investigador Invitado. Sus temas de investigación son el reconocimiento automático de habla y hablantes, la identificación de hablantes en el ámbito forense y la minería de datos.