SISTEMA MULTIPARAMÉTRICO PARA LA COMPARACIÓN FORENSE DE HABLANTES

A MULTIPARAMETRIC SYSTEM FOR FORENSIC SPEAKER COMPARISON

EUGENIA SAN SEGUNDO
Department of Criminal Science and Technology, Shanxi Police College (China)
eugenia@sxpc.edu.cn

PEDRO UNIVASO
BlackVOX
(Argentina)
punivaso@blackvox.com.ar

JORGE GURLEKIAN
Laboratorio de Investigaciones Sensoriales, INIGEM, UBA-CONICET (Argentina)
jgurlekian@fmed.uba.edu

Artículo recibido el día: 26/02/2019
Artículo aceptado definitivamente el día: 05/06/2019
Estudios de Fonética Experimental, ISSN 1575-5533, XXVIII, 2019, pp. 13-45

ABSTRACT

In Forensic Speaker Comparison (FSC) several different parameters are commonly analysed. In this investigation we propose a multiparametric system combining long-term features (F0, voice quality and durational aspects) with short-term features (MFCCs), used by a standard automatic system based on i-vector/PLDA approaches (baseline system). The objective was to determine if the performance of the new FSC system is better than that of the baseline system. For this, three experimental designs were carried out –allowing us to evaluate the new multiparametric system in extreme conditions, as if it was a stress test–: (1) use of forensically-realistic characteristics (e.g. background noise, reverberation, intra-speaker variability, signal compression); (2) voice comparison of 12 monozygotic twin pairs; and (3) comparison of disguised voices through nose pinching. The results show that the new system performs better than the baseline system although the mean contribution of long-term features to the new system was 6.5%, with the short-term features being responsible for the remaining 93.5%.

Keywords: *MFCCs, voice quality, stress test, twins, disguise*.

RESUMEN

En la comparación forense de hablantes se pueden examinar diversos parámetros. En el presente trabajo se utilizó un sistema multiparamétrico que combina parámetros acústicos de largo plazo (F0, cualidad de voz y aspectos duracionales) con los parámetros de corto plazo (*MFCC*) empleados por un sistema automático estándar basado en el enfoque *i-vector/PLDA* (sistema base). El objetivo era determinar si el nuevo sistema de comparación de hablantes ofrece mejor rendimiento que el sistema base. Para ello se llevaron a cabo tres experimentos con diseños diferentes –que permitieron evaluar el nuevo sistema multiparamétrico en condiciones extremas, a modo de prueba de estrés–: (1) uso de grabaciones con características forenses realistas (p. ej. ruido de fondo, reverberación, variabilidad intra-hablante, compresión de la señal); (2) comparación de las voces de 12 parejas de gemelos monocigóticos; y (3) cotejo de voces con enmascaramiento mediante pinzamiento de nariz. Los resultados obtenidos con el nuevo sistema muestran una mejora de rendimiento con respecto al sistema base, si bien el aporte medio de los parámetros de largo plazo al nuevo sistema fue de un 6.5%, siendo el restante 93.5% responsabilidad de los parámetros de corto plazo.

Palabras clave: *MFCC, cualidad de voz, prueba de estrés, gemelos, disimulo*.

1. INTRODUCCIÓN

No es raro encontrar hoy en día a expertos en ciencias del habla o fonetistas que, actuando como peritos forenses, siguen usando los términos "identificación" e "individualización", o bien adjetivos como "única" para referirse a la voz de una persona. Es especialmente preocupante cuando dichos sustantivos van acompañados, en los peritajes de voz, de otras palabras como "absoluta", "incuestionable" o expresiones como "más allá de toda duda razonable". Este problema no atañe únicamente al ámbito de la voz como prueba judicial o forense. Efectivamente, Champod *et al.* (2018) debaten sobre esta cuestión en relación con el ADN, los residuos de bala y otra serie de pruebas que se pueden encontrar en la escena de un crimen o estar relacionadas con un acto delictivo. El uso de las expresiones mencionadas anteriormente, que varios autores achacan en gran medida a la propagación de series como CSI (Schweitzer y Saks, 2006), implica una supresión deliberada de la idea de 'incertidumbre'. El experto forense no debería permitir que los tribunales crean que se puede abordar un informe pericial sin tener en cuenta la incertidumbre. Para ello, es fundamental que el experto tenga ciertos conocimientos sobre probabilidad e inferencia.

La Red Europea de Institutos de Ciencias Forenses (*European Network of Forensic Science Institutes, ENFSI*) publicó en 2015 unas directrices para estandarizar y mejorar los informes periciales (de tipo evaluativo) en el conjunto de las disciplinas forenses (ENFSI, 2015). El reconocimiento de hablantes figura como una de las áreas en la que son aplicables dichas directrices. Por otro lado, entre los 69 miembros repartidos en 37 países que forman parte del ENFSI, encontramos tanto la Comisaría General de Policía Científica de la Policía Nacional de España como el Servicio de Criminalística de la Guardia Civil. Este último figura, además, como uno de los principales artífices en la elaboración de las directrices mencionadas anteriormente.

El punto 2.4 de las directrices del ENFSI establece que la evaluación forense, independientemente de la disciplina concreta de la que se trate, se basará en la asignación de una ratio o relación de verosimilitud (inglés: *likelihood ratio; LR* de ahora en adelante). Esta relación mide la fuerza de apoyo que los resultados proporcionan para discriminar entre las proposiciones de interés. Generalmente estas proposiciones son la hipótesis de que las muestras de habla proceden del mismo hablante (H0 o hipótesis del fiscal) y la hipótesis de que proceden de distinto hablante (H1 o hipótesis de la defensa). Por otro lado, las *LR* "están científicamente aceptadas y proporcionan una forma lógica de lidiar con el razonamiento inferencial" (ENFSI 2015:6; nuestra traducción).

Actualmente se utilizan los términos "comparación forense del habla" o "comparación forense del hablante" para referirse al área forense que se centra en la voz como prueba pericial. Al menos así ocurre en la bibliografía especializada escrita en inglés. Desde hace algunos años la denominación previa "identificación de hablantes" viene siendo criticada por sus implicaciones semánticas. Desde el punto de vista defendido por el ENFSI y por otros autores anteriormente (Rose, 2002; Meuwly, 2006; Morrison, 2009a), el uso de la palabra "identificación" implica que el científico forense puede dar un veredicto con respecto a la identificación de un sospechoso, cuando en realidad ese es el papel del juez. Al usar el término neutro "comparación" no se incurre en el error de determinar probabilidades *a posteriori* (véase la falacia del fiscal en Thomson y Schuman, 1987 o Evett, 1995). De este modo, se respeta uno de los requisitos fundamentales que deben cumplir los informes forenses; esto es, el de la "lógica": "Los informes evaluativos deben abordar la probabilidad de los resultados dadas las proposiciones y el contexto relevante; y no la probabilidad de las proposiciones dados los resultados y el contexto relevante" (ENFSI, 2015:10; nuestra traducción).

En el contexto de la voz, la función del científico es la de responder a la siguiente pregunta: ¿cuánto más probable es que las diferencias observadas entre las muestras indubitada (muestra de origen conocido) y dubitada (muestra de origen desconocido) ocurran bajo la hipótesis de que ambas muestras tienen el mismo origen que bajo la hipótesis de que estas tienen un origen distinto? La manera de responder cuantitativamente a esta pregunta viene dada por la expresión de conclusiones en forma de una relación de verosimilitud. Esta perspectiva metodológica presenta la ventaja de: (1) tener en cuenta la variabilidad inter e intralocutor y (2) evaluar no solo la similitud entre las muestras de habla sino también su tipicidad con respecto a una población de referencia apropiada.

Morrison (2009b) defiende un cambio de paradigma en la comparación forense de hablantes que emule los cambios adoptados en el ámbito del ADN, si bien ya una década antes Champod y Meuwly (2000) afirmaban que el marco de interpretación bayesiano basado en *LR* era completamente necesario para evaluar la evidencia de voz. Un amplio número de investigaciones han demostrado que es posible abordar la disciplina de la comparación forense de hablantes desde esta perspectiva, no solo en estudios enmarcados en el reconocimiento automático de hablantes (Brümmer y du Preez, 2006; van Leeuwen y Brümmer, 2007) sino también usando parámetros fonético-acústicos extraídos mediante métodos propios de la fonética experimental (González-Rodríguez *et al.*, 2007; Hughes *et al.*, 2017). Para una

descripción detallada del marco bayesiano para la evaluación de evidencias forenses, véanse Berger *et al.* (2010) o Ramos-Castro (2007).

2. ESTADO DE LA CUESTIÓN Y OBJETIVO DEL ESTUDIO

En la comparación forense de hablantes se pueden examinar diversos parámetros, dada la amplia variedad de características fonéticas susceptibles de análisis. Generalmente se aborda el análisis de aspectos acústicos, tanto segmentales como suprasegmentales, que van desde la frecuencia fundamental y las frecuencias formánticas hasta la velocidad de articulación (Gold, 2018) y la cualidad de voz (San Segundo *et al*., 2018) –por citar estudios recientes– si bien el análisis de algunos de estos parámetros, particularmente los de tipo suprasegmental o prosódico, en ocasiones tiene un marcado carácter híbrido acústico-auditivo, o incluso una naturaleza totalmente perceptiva. Al no existir un procedimiento estándar consensuado entre los expertos, cada laboratorio tiende a emplear su propio enfoque particular, dependiendo de la formación e intereses de los expertos. Estos enfoques oscilan entre los puramente auditivo-perceptuales (Hollien 2002, San Segundo y Mompeán, 2017) y aquellos con un alto componente de automatización (véase Hansen y Hasan, 2015). Los llamados métodos semiautomáticos se basan principalmente en el análisis de formantes vocálicos. El empleo conjunto de diferentes métodos complementarios permite considerar diferentes fenómenos de la cadena hablada, consiguiendo así resultados más precisos, de manera similar a los métodos de ensamble empleados en la minería de datos (Zhou, 2012).

En el presente trabajo se utiliza un sistema de comparación de hablantes que adiciona parámetros acústicos de largo plazo (frecuencia fundamental, cualidad de voz y aspectos duracionales) a los parámetros de corto plazo (*Mel Frequency Cepstral Coefficients, MFCC*) empleados por un sistema automático estándar basado en el enfoque *i-vector/PLDA*, considerado el estado del arte de acuerdo con la última evaluación del *NIST*[1] del año 2016, y que denominaremos sistema base. Uno de los sistemas que obtuvo mejor rendimiento en dicha competición fue el *I4U* (Lee *et al.*, 2017), desarrollado colaborativamente entre 16 instituciones y universidades de los 4 continentes, el cual empleó 15 sistemas basados en la metodología *i-vector/PLDA* de los 17 sistemas presentados (véase § 3.4.1).

[1] Instituto Nacional de Patrones y Tecnología (NIST, por sus siglas en inglés: *National Institute of Standards and Technology*).

El objetivo del presente estudio es comprobar si el nuevo sistema forense propuesto, que incluye características del hablante de corto y largo plazo, ofrece mejor rendimiento que el sistema base. Para poner a prueba la robustez del sistema de comparación multiparamétrico propuesto, se han llevado a cabo tres experimentos con diseños distintos. En primer lugar, para el entrenamiento del sistema se ha utilizado el corpus de grabaciones en condiciones extremas *SITW (Speakers In The Wild*, McLaren *et al.*, 2016*).* Estas grabaciones presentan características realistas desde el punto de vista forense (p. ej. ruido de fondo, reverberación, variabilidad intra-hablante, distintos tipos de compresión de la señal). En segundo lugar, se han comparado las voces de 12 parejas de gemelos monocigóticos pertenecientes al corpus de gemelos descrito en San Segundo (2013, 2014). En tercer lugar, se ha utilizado una tarea de lectura que presenta enmascaramiento de la voz mediante la técnica del pinzamiento de nariz. Estas tres condiciones experimentales permiten evaluar el rendimiento del sistema propuesto en condiciones extremas. De esta manera, se pretende que los resultados sirvan para que el analista forense evalúe la utilidad que el sistema paramétrico propuesto pueda tener aun en los casos más complejos que puedan darse en el ámbito forense.

Son numerosos los estudios que contemplan las condiciones experimentales que acabamos de mencionar, aunque casi siempre de manera independiente. Por ejemplo, el uso de características realistas desde el punto de vista forense se recomienda ya en Morrison *et al*. (2012) para la recogida de corpus de voces con fines judiciales. Para el español, en concreto, varios estudios se han publicado usando el corpus CIVIL (San Segundo *et al*., 2013), que sigue las directrices establecidas por Morrison *et al*. (2012). En cuanto a la utilidad de probar los sistemas de comparación de hablantes con gemelos, San Segundo (2014) recoge varias referencias bibliográficas al respecto, aunque en los últimos cinco años no han dejado de aparecer nuevas publicaciones. Por citar algunas de las más recientes, da Costa Fernandes (2018) o Sabatier *et al.* (2019) ponen a prueba sus respectivos sistemas automáticos comparando parejas de gemelos, pues consideran que estos pueden servir como la prueba de estrés definitiva de los sistemas automáticos actuales. Es decir, se trataría de probar los límites funcionales de un sistema sometiéndole a condiciones extremas, en este caso, de similitud entre hablantes. Finalmente, los estudios recientes que se han ocupado de explorar el efecto del disimulo de la voz en los sistemas automáticos no son muy numerosos, a pesar de que es ampliamente conocido (cf. Masthoff, 1996; Künzel, 2000, entre otros) que, en el contexto forense, los hablantes no siempre son cooperativos y desean ser reconocidos. Hautamäki *et al.* (2017) estudia un tipo concreto de disimulo: aquel que se da cuando un hablante enmascara su identidad impostando

una voz más joven o más vieja que la suya. Este tipo de estudios se suelen enmarcar en un contexto acústico-perceptivo. Es decir, el objetivo es conocer cómo influye el disimulo en una serie de parámetros acústicos (F0, F1, F2, F3 y F4 en el caso de Hautamäki *et al.* (2017)), y también cómo afecta dicho disimulo a la capacidad de reconocimiento de voces por parte de una serie de oyentes o jueces perceptivos. Es en esta línea perceptiva en la que se inscribe el único estudio que conocemos para el español que estudia el pinzamiento de nariz como estrategia para el enmascaramiento de la voz (Gil y San Segundo, 2013).

3. METODOLOGÍA

3.1. Corpus

Para las comparaciones de hablantes se utilizó el corpus de gemelos masculinos *Twin Corpus* (San Segundo, 2013, 2014). Dicho corpus está conformado por grabaciones de 24 gemelos idénticos (es decir, monocigóticos), con edades comprendidas entre los 18 y los 52 años (media de 29 años). Todos los hablantes de dicho corpus son hablantes nativos de español peninsular (región norte-central), sin patologías en el habla o dificultades auditivas. Se obtuvieron grabaciones de dos textos leídos en dos ocasiones diferentes, separadas por 2-3 semanas, de manera que se tuvo en cuenta la variación intra-hablante, obteniéndose así muestras de habla no contemporáneas. Los gemelos concurrieron juntos a las sesiones de grabación que tuvieron lugar en el Laboratorio de Fonética del Consejo Superior de Investigaciones Científicas de España. Las grabaciones se realizaron con un micrófono de condensador omnidireccional de respuesta plana en frecuencia. Las características técnicas de la grabación fueron las siguientes: frecuencia de muestreo de 44,1 KHz, 16 bits de resolución y canal monofónico. Para el presente estudio se utilizó una frecuencia de re-muestreo de 16,0 KHz, con el fin de simular el ancho de banda empleado habitualmente en el ámbito forense. Cada miembro de una pareja de gemelos fue grabado en una sala diferente, aisladas acústicamente, aunque conectadas telefónicamente para la ejecución de tareas colaborativas. A pesar de que las grabaciones fueron de alta calidad, este montaje replica aproximadamente las condiciones reales de las grabaciones del ámbito forense. Adicionalmente se pidió a cada hablante que repitiera los dos textos ya leídos, pero con pinzamiento de la nariz realizada con su propia mano. De esta manera se obtuvieron grabaciones de habla leída con uno de los tipos de enmascaramiento o disimulo de naturaleza más paradójica en los casos forenses (Gil y San Segundo, 2013:332).

A partir de las grabaciones obtenidas se generaron dos tareas de alta exigencia para la comparación de hablantes en el ámbito forense. La primera tarea, denominada *GEMELOS*, busca analizar el impacto del uso de gemelos en el sistema automático y la segunda, denominada *DISIMULO*, pretende averiguar el efecto que produce comparar hablantes no gemelares pero que intentan enmascarar sus voces a través del pinzamiento de nariz.

Para el entrenamiento del nuevo sistema se utilizó la base de datos *SITW* (McLaren *et al.*, 2016), desarrollada especialmente para evaluar sistemas de reconocimiento de habla independiente del texto en grabaciones mono- y multi-hablante recogidas en condiciones "salvajes" (*wild*). La base de datos consiste en grabaciones de 299 hablantes con un promedio de 8 sesiones por hablante tomadas en ambientes reales de alto ruido, reverberación, variabilidad intralocutor, y en diferentes canales y tipos de compresión de señal. Estos factores hacen de *SITW* un gran desafío para el reconocimiento de hablantes. En el año 2016 el laboratorio *SRI International* convocó a una evaluación internacional de sistemas de reconocimiento de hablantes, denominada *The 2016 Evaluation Leaderboard*, a la que se presentó el sistema base utilizado en esta investigación. Se recibieron 45 presentaciones correspondientes a 11 equipos de diferentes países. El sistema presentado ocupó el séptimo puesto, siendo el resultado obtenido similar al promedio de los participantes.

En concreto, para el entrenamiento del sistema que emplea parámetros de largo plazo se utilizó un subgrupo de datos del *SITW* correspondiente a grabaciones de micrófono de solapa (*lapel*) que poseen características similares a los de la base de datos de prueba (*Twin Corpus*) empleada en el presente trabajo, como se explicó más arriba.

3.2. Parámetros

Los sistemas automáticos de comparación forense de hablantes generalmente emplean una serie de métodos que incluyen el preprocesamiento de la señal, la detección de actividad vocal (*VAD – Voice Activity Detection*), la extracción de parámetros –que hemos llamado de corto plazo–, el modelado y la medida de similitud, para finalmente obtener el cómputo de *LR* (véase § 1). La particularidad de este trabajo consiste en la incorporación a los parámetros de corto plazo, expresados en *LR*, de una serie de parámetros acústicos de largo plazo. Empezaremos describiendo el primer tipo de parámetros (coeficientes cepstrales en escala Mel) y a continuación los tres tipos de parámetros de largo plazo considerados para esta investigación (véanse § 3.2.2 a 3.2.4). Como

puntualización, hay que destacar que el criterio para la selección de parámetros de largo plazo fue que se pudieran capturar sus aportes en términos de *LR* y que pudieran ser extraídos de forma automática (sin transcripción previa de las emisiones).

3.2.1. Parámetros de corto plazo: coeficientes cepstrales en escala Mel

Los parámetros extraídos de la envolvente espectral de corto plazo se centran en la forma espectral de la señal de voz derivada de una porción corta (trama) de la señal. Su objetivo es examinar la influencia del tracto vocal (trama por trama) ignorando la influencia de la fuente de voz, en particular la frecuencia fundamental. Los parámetros mayormente empleados son los coeficientes cepstrales en escala Mel (*MFCC*) (Davis y Mermelstein, 1980), aunque también suelen utilizarse los coeficientes de codificación predictiva lineal (*LPC*) (Makhoul, 1975) y los coeficientes de predicción lineal perceptual (*PLP*) (Hermansky, 1990). En el presente trabajo se emplearon los coeficientes *MFCC*.

Primeramente, la señal fue preprocesada por medio del paquete *Qualcomm-ICSI-OGI*, que realiza un filtrado *Wiener* y filtros *RASTA-LDA* (Adami *et al.*, 2002). La extracción de parámetros se realizó con el paquete desarrollado por la Universidad de Cambridge: *HTK Toolkit ver. 3.4* (Young *et al.*, 2006). Los 12 parámetros *MFCC* y el logaritmo de la energía, junto a la primera y segunda derivadas fueron extraídos cada 10 ms, generándose un vector de 39 parámetros por trama. La selección de segmentos de habla que fueron procesados se realizó por medio del detector de actividad vocal (*VAD*) basado en energía incluido en el paquete *Alize* (Bonastre *et al.*, 2005), al que se añadió un algoritmo heurístico de restricción de duraciones. Posteriormente se aplicó una normalización cepstral media (*CMN*) a cada segmento.

3.2.2. Parámetros de largo plazo (i): frecuencia fundamental

Dentro de los diferentes parámetros relacionados con la fuente glótica que son susceptibles de análisis con fines forenses, hemos considerado los relacionados directamente con la frecuencia fundamental, como son las medidas de tendencia central del F0 (es decir, valor medio, mediana y moda), los límites del F0 (es decir, mínimo y máximo) y la variabilidad del F0 (es decir, desviación estándar y coeficiente de variación; esto es, desviación estándar dividida por el valor medio). Estos parámetros se obtuvieron empleando el software de análisis y síntesis de señales de habla *Praat* (Boersma y Weenink, 2005) con todos los coeficientes estándar.

3.2.3. Parámetros de largo plazo (ii): cualidad de voz

La evaluación de diferentes cualidades de voz es requerida en el diagnóstico médico de las disfunciones vocales. Dicho análisis se realiza por medio de juicios perceptuales y medidas objetivas, tales como características acústicas y aerodinámicas, para lo cual existe una gran variedad de técnicas. Por ejemplo, el índice de perturbación, que mide el riesgo vocal (Gurlekian y Molina, 2012), emplea los parámetros correspondientes a la perturbación de la amplitud (*Shimmer*), la perturbación de la frecuencia fundamental (*Jitter*), la armonicidad o relación armónicos-ruido o segmentos sonoros-sordos (*HNR – Harmonic-to-Noise-Ratio*), y la amplitud del Cepstrum. Una de las técnicas perceptuales más empleadas es la propuesta por la Sociedad Japonesa de Logopedia y Foniatría (Hirano, 1981), conocida como escala *GRBAS* (G de *Grade*, R de *Roughness*, B de *Breathiness*, A de *Astheny*, y S de *Strain*). *Grade* se corresponde con el nivel general de desvío de la voz con respecto a una regular o modal, *Roughness* con la fluctuación irregular de la frecuencia fundamental, *Breathiness* con el ruido turbulento producido por el pasaje de aire, *Astheny* se reserva para aquellas voces con poca energía en la emisión, y *Strain* para las voces forzadas o que poseen tensión muscular.

La correlación entre las medidas objetivas y las perceptuales de la escala *GRBAS* ha sido profundamente estudiada, pero aún no se ha llegado a un acuerdo general entre los investigadores. Por ejemplo, Dejonckere *et al.* (1996) determinaron que los parámetros con mayor correlación son: G con el cociente entre *Shimmer* y *HNR*; R con *Jitter*; y B con *Shimmer*. El trabajo de Martin *et al.* (1995) relacionó R con *HNR* y *Shimmer*, y B con una combinación de *Jitter*, *Shimmer* y *HNR*. Por otra parte, Bhuta *et al.* (2004) encontraron que R estaba solo correlacionado con *HNR*. En el presente trabajo hemos utilizado los siguientes parámetros de cualidad de voz: *Jitter*, *Shimmer* y *HNR*. Estos parámetros se extrajeron de forma automática usando el comando "Voice Report" de *Praat*. El tipo de medición del jitter empleado fue *Jitter (local)*, que es el promedio de las diferencias absolutas entre períodos consecutivos, dividido por el período promedio. Para el shimmer se empleó *Shimmer (local)*, que es el promedio de las diferencias absolutas de las amplitudes entre períodos consecutivos, dividido por la amplitud promedio. Para ambas mediciones Praat considera únicamente los segmentos sonoros. Por un lado, se consideraron los parámetros de forma aislada; por otro lado, se calculó el cociente entre *Shimmer* y *HNR* para obtener un valor objetivo que correspondiese lo más aproximadamente posible a la valoración perceptiva *Grade*, de acuerdo con los resultados de Dejonckere *et al*. (1996), como se ha especificado anteriormente.

3.2.4. Parámetros de largo plazo (iii): velocidad de articulación y ritmo

Un parámetro basado en la duración de la emisión es la velocidad de articulación, que se define como la cantidad promedio de sílabas por segundo, excluyendo los silencios. El porcentaje promedio de segmentos sonoros y sordos de una sílaba es otro parámetro de duración, relacionado con el ritmo. Para el cálculo de ambas medidas se utilizó el script para *Praat*, basado en el núcleo silábico, desarrollado por De Jong y Wempe (2008).

La tabla 1 recoge todos los parámetros acústicos de largo plazo extraídos en este estudio para su incorporación al sistema base de comparación forense de hablantes.

Nro.	Parámetro	Descripción	Tipo
1	$F0_{mean}$	Valor medio del F0	Tono
2	$F0_{min}$	Valor mínimo del F0	
3	$F0_{max}$	Valor máximo del F0	
4	$F0_{sd}$	Desviación estándar del F0	
5	$F0_{sd/mean}$	Coeficiente de variación del F0 ($F0_{sd}$ / $F0_{mean}$)	
6	*Jitter*	Perturbación del F0	Cualidad de voz
7	*Shimmer*	Perturbación de la amplitud	
8	*HNR*	Armonicidad o relación armónicos-ruido	
9	*Grado*	*Shimmer* / *HNR*	
10	*Ritmo*	Porcentaje promedio de segmentos sordos	Duración
11	*VA*	Velocidad de articulación	

Tabla 1. *Parámetros de largo plazo utilizados en el sistema de reconocimiento de hablantes propuesto, categorizados según la clasificación de Lehiste (1970).*

3.3. Modelado del hablante

Esta etapa tiene como finalidad generar un modelo probabilístico de la voz del hablante, a partir de una o varias emisiones de habla de dicho hablante. El modelo empleado con más frecuencia en los sistemas de reconocimiento automático de

hablantes es el modelo de mezclas de gaussianas (*GMM*), del cual surgen los enfoques *GMM-UBM*, supervectores y las aproximaciones por vectores de factor total (*i-vectors*) (Dehak *et al.*, 2011), así como las redes neuronales profundas (*DNN – Deep Neural Networks*) (Hinton *et al.*, 2012) que han resurgido en la actualidad debido a la implementación de nuevas técnicas de inicialización, al uso del procesamiento paralelo y a la capacidad de procesamiento de la tecnología. En el enfoque de supervectores los parámetros del modelo *GMM* son proyectados en un espacio de alta dimensionalidad, mientras que la aproximación *i-vector*, que es la que utilizamos en este estudio, reduce dicha dimensionalidad tratando de mantener la información esencial del hablante. La compensación del canal se realiza en este nuevo espacio de dimensiones reducidas empleando el análisis discriminativo lineal probabilístico (*PLDA*), dando nacimiento al enfoque *i-vector/PLDA*. Para calcular el *LR*, dichos enfoques requieren emplear un modelo de la población relevante, generalmente representado por el modelo universal (*UBM*).

Siguiendo la metodología habitual en el ámbito del reconocimiento automático de hablantes, que emplea *LR* como medida final de las comparaciones, hemos utilizado gaussianas simples para modelar a los hablantes, mientras que para el modelado de la población relevante se ha seguido el método de densidad de función de núcleo multivariada (*MVKD*), propuesta originalmente por Aitken y Leese (1995).

3.4. Cotejo de voces y cálculo de *LRs*

La comparación de hablantes empleada en el ámbito forense parte de la grabación de una voz relacionada con un hecho delictivo (p. ej. grabación dubitada, prueba o evidencia), que se compara con otros registros atribuidos a una persona, normalmente conocida (p. ej. grabación indubitada o plana de voz del imputado). La grabación dubitada generalmente se obtiene de registros telefónicos, mientras que la indubitada se realiza durante la toma de declaración del imputado en la sede policial o el laboratorio forense. En otros casos la grabación indubitada puede provenir de un registro telefónico (p. ej. escucha telefónica) y la dubitada de grabaciones en vivo (p. ej. *whatsapp* o grabadora del teléfono móvil).

3.4.1. Método de medida: parámetros de corto plazo

Para el entrenamiento de todas las *GMM* y modelos de variabilidad total se utilizó el paquete Alize (véase § 3.2.1). Se entrenó un modelo *GMM* de 1024 mezclas independientes del género a partir de la base de datos desarrolladas por *NIST*

(Instituto Nacional de Estándares y Tecnología de USA) (https://www.nist.gov/itl/iad/mig/speaker-recognition) en la evaluación de sistemas de reconocimiento de hablantes *SRE04*. Después de esto, se entrenó un espacio de variabilidad total con las bases de datos *SRE08 short2* y *short3* para producir *i-vectors* de 400 dimensiones. Antes de aplicar la técnica de compensación de canal denominada análisis de discriminación lineal (*LDA*) sin reducción de dimensionalidad, cada *i-vector* fue transformado utilizando la normalización esférica de atributos perjudiciales (*sphNorm – Spherical Nuisance Normalization*). Este post-procesamiento también se realizó con la base de datos SRE08. La compensación de canal y hablantes empleó el análisis discriminativo lineal probabilístico (*PLDA*) utilizando 300 dimensiones para el espacio de los hablantes y 100 dimensiones para el del canal. Para ello se utilizaron las bases de datos *SRE08 short2* y *short3* y la base de datos de desarrollo del *SITW*. No se aplicó ninguna técnica de normalización de medida.

Para la calibración de los *LR* de salida del sistema se empleó el algoritmo de regresión logística implementado en *MATLAB* por medio de la herramienta *BOSARIS* (Brümmer y de Villiers, 2011). Además de *LR* se incorporaron en la calibración las duraciones de las emisiones de entrenamiento y prueba, con el objetivo de obtener una función de costo logarítmica (C_{llr}). El entrenamiento se realizó con la base de datos de desarrollo *SITW*.

En este estudio, el sistema base (Martínez Soler *et al.*, 2018) emplea el enfoque *i-vector/PLDA*, conformado por parámetros de corto plazo (*MFCC*), y utiliza el "método directo" para el cálculo de *LR*. Esta metodología requiere un modelo estadístico que permita medir directamente un valor de verosimilitud cuando los vectores de características son comparados con respecto a dicho modelo.

Habiendo unificado como medida el *LR*, tanto para los parámetros de corto plazo como para los parámetros de largo plazo (véase § 3.4.2), y siendo el cociente de verosimilitud final el producto de los cocientes de verosimilitud de cada uno de los parámetros empleados, podemos expresar la medida final del sistema como en la fórmula (1), donde LR_{final} es el cociente de verosimilitudes final, $LR_{ivector}$ el cociente de verosimilitudes de los parámetros de corto plazo empleando el enfoque *i-vector/PLDA*, LR_j el cociente de verosimilitudes del parámetro de largo plazo *j*, y *n* la cantidad de parámetros de largo plazo empleados.

$$(1) \qquad LR_{final} = LR_{ivector} \prod_{j=1}^{n} LR_j$$

Los valores de *LR* correspondientes a los parámetros de largo plazo, resultantes del método de medida, están generalmente bien calibrados y son poco dependientes del canal y de la duración de las emisiones. En cambio, el *LR* que surge del método directo suele estar descalibrado y requiere, por tanto, una calibración previa antes de introducirlo en la ecuación final. Un método de calibración generalmente empleado es la regresión logística, en la que se fusiona el valor de *LR* y la duración de las emisiones, con el objetivo de obtener la función de costo logarítmica (C_{llr}) (Brümmer, 2004), que sirve para medir el rendimiento de un sistema forense.

Debido a la magnitud del cociente de verosimilitud, normalmente se emplea el logaritmo del *LR*, denominado *LLR* (*log Likelihood Ratio*), con lo cual la ecuación 1 queda expresada como en (2).

$$LLR_{final} = LLR_{ivector} + \sum_{j=1}^{n} LLR_j \qquad (2)$$

3.4.2. Método de medida: parámetros de largo plazo

El modelado a partir de los parámetros de largo plazo emplea una variante del "método de medida" (*scoring method*), conformado por tres etapas. En la primera se calcula la medida de similitud (es decir, distancia euclidiana normalizada) entre pares de emisiones de prueba para cada parámetro, según la ecuación (3), donde x_i^j es la medida de similitud *i* entre los valores *p* del parámetro *j* para las emisiones *a* y *b*.

$$x_i^j = \frac{\left| p_a^j - p_b^j \right|}{\left| p_a^j + p_b^j \right|} \qquad (3)$$

En la segunda etapa se transforman las medidas de similitud en dos distribuciones de densidad de probabilidad univariadas; una representa las verosimilitudes de las medidas de similitud entre emisiones que pertenecen a los mismos hablantes (hipótesis H0) y la otra a hablantes diferentes (hipótesis H1) (figura 1.a). Dichas distribuciones pueden estimarse con la fórmula de densidad de función núcleo (*KDE – Kernel Density Estimation*) desarrollada por Silverman (1986) y expresada por la ecuación (4), donde x_i es uno de los *k* elementos correspondiente a la medida

de similitud del parámetro *j*, *s* es la varianza de la muestra, θ el valor medio del elemento y λ el parámetro de suavización elegido.

$$K(\theta \mid x_i, \lambda) = \frac{1}{\lambda s \sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x_i - \theta)^2}{\lambda^2 s^2} \right) \quad (4)$$

Aitken y Taroni (2004) sugieren que este parámetro puede elegirse subjetivamente basándose en la experiencia del investigador. En nuestro caso se consideró un valor de λ=1. Por tanto, la estimación de la distribución de densidad de probabilidades quedó expresada como en la ecuación (5).

$$f_j(\theta \mid D, \theta) = \frac{1}{k} \sum_{i=1}^{k} K(\theta \mid x_i, \lambda) \quad (5)$$

En la tercera etapa se determina el *LR* entre ambas hipótesis en función de la medida de similitud (figura 1.b) y se aproxima dicha ecuación por un polinomio de segundo orden. Esta última ecuación (6) permite calcular el aporte, expresado en *LR*, de cada parámetro al sistema de comparación de hablantes final (véase la ecuación (2)).

$$LR_j = \frac{f_j(x_i \mid H_0)}{f_j(x_i \mid H_1)} \cong a_j \cdot x_i^2 + b_j \cdot x_i + c_j \quad (6)$$

Para la determinación de las ecuaciones polinómicas de segundo orden, correspondientes a cada parámetro de largo plazo empleado, se utilizaron los datos de micrófono (*lapel*) de la base de datos *SITW*. Se seleccionaron 42 muestras de grabaciones de micrófono, correspondientes a 18 hablantes que conformaron 147 pares de grabaciones de diferentes hablantes y 41 pares de grabaciones no contemporáneas de los mismos hablantes.

La selección de los parámetros de largo plazo se realizó en dos etapas. En la primera se seleccionaron los parámetros acústicos que se consideró que representan mejor la información prosódica que no capturan los parámetros de corto plazo y que, además, pueden ser extraídos de forma automática (véase § 3.2, tabla 1). En una segunda etapa se determinó cuál de ellos posee capacidad para la comparación de hablantes, de manera que pudieran incorporarse al sistema base.

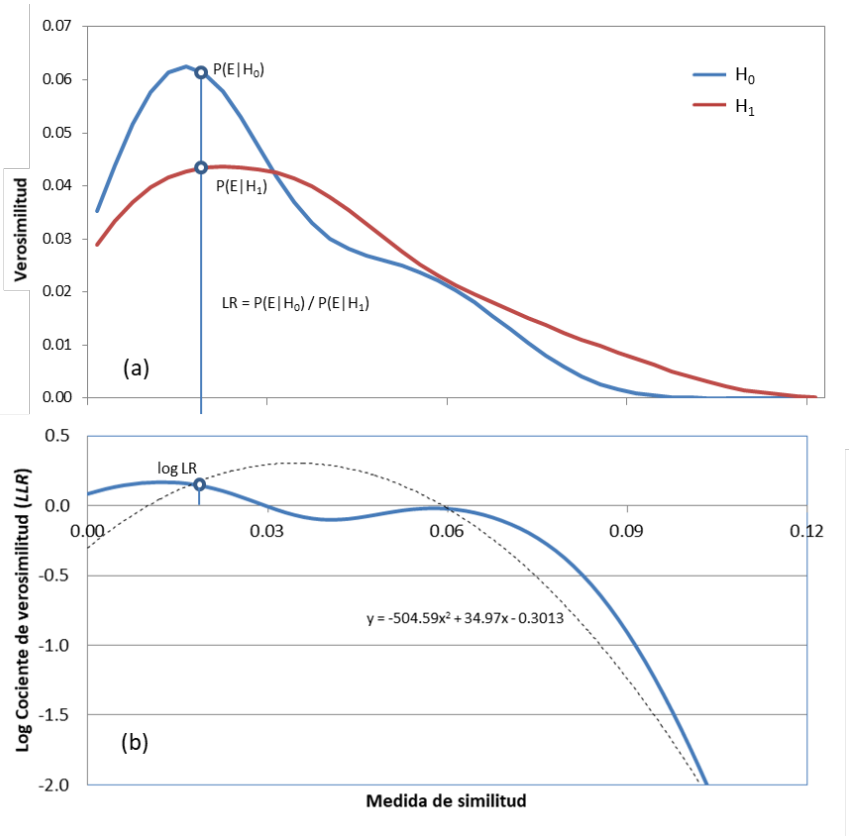

Figura 1. a) *Ejemplo de distribución de la medida de similitud correspondiente a emisiones del mismo hablante (H_0) y diferentes hablantes (H_1), y b) aproximación polinómica de segundo orden del logaritmo del cociente de verosimilitud (LLR) en función de la medida de similitud.*

En la tabla 2 pueden observarse los valores obtenidos para los coeficientes de la función polinómica de segundo orden (ecuación (6)) correspondientes a cada uno de los parámetros de largo plazo empleados. El error promedio porcentual (error%) entre la función de aproximación y la real se calculó como el promedio de los valores absolutos de los errores absolutos divido por el rango.

	$a.\ x^2 + bx + c$			
Parámetro	*a*	*b*	*c*	error%
$F0_{mean}$	23.70	-4.90	0.13	5.0%
$F0_{min}$	-50.00	2.02	0.04	2.7%
$F0_{max}$	0.80	-0.02	-0.06	6.3%
$F0_{sd}$	-1.50	0.41	0.07	3.2%
$F0_{sd/mean}$	-1.64	0.33	0.09	1.4%
Jitter	12.60	-2.50	0.01	5.1%
Shimmer	-4.70	-0.20	0.13	1.0%
HNR	-14.10	1.20	0.13	1.9%
Grado	-2.40	0.07	0.13	0.9%
Ritmo	-6.00	-0.47	0.35	1.1%
VA	-11.00	1.38	0.01	4.2%

Tabla 2. *Valores de los coeficientes de la aproximación polinómica de segundo orden obtenidos a partir de la base de datos SITW, correspondientes a cada parámetro de largo plazo y el error promedio porcentual entre la función de aproximación y la real.*

En una segunda etapa se seleccionaron solo aquellos parámetros que poseían capacidad discriminativa, para lo cual se consideraron únicamente aquellos que lograban minimizar el C_{llr}. Con este fin, se empleó una metodología iterativa partiendo de la selección de todos los parámetros y la eliminación consecutiva de aquellos parámetros que incrementaban el C_{llr}, hasta cubrir todos los parámetros.

Para eliminar valores fuera de rango (*outliers*) se acotaron los valores de dicha función entre [-1;+1], con el fin de no desvirtuar el aporte de los parámetros de largo plazo en el LLR_{final} (ecuación (2)).

4. RESULTADOS

En la tabla 3 pueden verse los resultados de aplicar la metodología propuesta, que emplea parámetros de corto y largo plazo. Al sistema resultante lo denominaremos nuevo sistema y al enfoque basado únicamente en parámetros de corto plazo (*i-vector/PLDA*) lo llamaremos sistema base. En la tabla los resultados aparecen disgregados en función de las distintas bases de datos empleadas.

En la tabla 4 se muestran los parámetros de largo plazo finalmente seleccionados por su poder discriminativo en el reconocimiento de hablantes para las siguientes condiciones experimentales: grabaciones "salvajes" (SITW), gemelos (GEMELOS) y enmascaramiento por pinzamiento de nariz (DISIMULO).

		EER%	$C_{llr\ min}$
SITW	nuevo sistema	12.9%	0.540
	sistema base	14.6%	0.619
GEMELOS	nuevo sistema	8.3%	0.237
	sistema base	8.3%	0.265
DISIMULO	nuevo sistema	9.4%	0.369
	sistema base	11.1%	0.430

Tabla 3. *Tasa de igual error (EER) y función de costo logarítmica mínima ($C_{llr\ min}$) del nuevo sistema y del sistema base para las condiciones experimentales: SITW, GEMELOS y DISIMULO.*

Parámetro	*SITW*	*GEMELOS*	*DISIMULO*	*Tipo de parámetro*
$F0_{mean}$	X		X	Tono
$F0_{min}$	X		X	
$F0_{max}$			X	
$F0_{sd}$		X		
$F0_{sd/mean}$	X	X	X	
Jitter	X		X	Cualidad de voz
Shimmer	X		X	
HNR		X	X	
Grado	X	X	X	
Ritmo	X	X	X	Duración
VA		X	X	

Tabla 4. *Parámetros de largo plazo finalmente seleccionados por su poder discriminativo para cada base de datos empleada, agrupados por tipo de parámetro según la clasificación de Lehiste (1970).*

Para poder analizar la influencia de los parámetros de largo y corto plazo, definimos los coeficientes (7) y (8) a partir de la ecuación (6).

$$(7) \qquad LLR_{cp}\% = \frac{LLR_{ivector}}{LLR_{final}} \times 100$$

$$(8) \qquad LLR_{lp}\% = \frac{\sum_{j=1}^{n} LLR_j}{LLR_{final}} \times 100$$

Los coeficientes (7) y (8) representan el porcentaje de LLR_{final} que corresponde a los parámetros de corto y largo plazo respectivamente en un cotejo en particular. El valor promedio que se muestra en la tabla 5 es sobre el total de emisiones de cada base de datos empleada.

	$\overline{LLR}_{lp}\%$	$\overline{LLR}_{cp}\%$
SITW	4.10%	95.90%
GEMELOS	5.30%	94.70%
DISIMULO	10.00%	90.00%

Tabla 5. *Coeficientes promedio de largo plazo ($\overline{LLR}_{lp}\%$) y corto plazo ($\overline{LLR}_{cp}\%$) para cada base de datos empleada.*

5. DISCUSIÓN Y CONCLUSIONES

Los resultados obtenidos con el nuevo sistema de comparación de hablantes, que incluye rasgos distintivos del hablante de corto y largo plazo, muestran una mejora de rendimiento con respecto al sistema base, y también con respecto al estado del arte. Por ejemplo, el mismo corpus de gemelos (San Segundo 2013, 2014) se utilizó para probar el sistema de reconocimiento automático de hablantes comercial BATVOX (San Segundo y Künzel, 2015) y el EER obtenido fue de 9.9%. En el presente estudio se obtuvo un 8.3%, si bien es cierto que la versión del sistema BATVOX utilizada en San Segundo y Künzel (2015) estaba basada en el enfoque GMM-UBM (recordemos que la metodología propuesta en el presente estudio es la de *i-vectors*). Por otro lado, la investigación de 2015 usó grabaciones

espontáneas del corpus de gemelos mientras que el estudio presente se basa en una tarea de lectura.

El empleo de grabaciones pertenecientes a casos complejos –como lo son las grabaciones en condiciones extremas (*SITW*), la comparación de gemelos y el disimulo de la voz mediante pinzamiento de nariz– nos permite concluir que el nuevo sistema puede ser de gran utilidad en ámbitos como el forense donde en ocasiones se trabaja con casos de similar complejidad. De todas formas, dichos casos complejos nos han servido para poner a prueba el sistema multiparamétrico propuesto, como era nuestro objetivo. Varios estudios consideran que los test realizados con voces de gemelos son una prueba de estrés muy útil para validar el rendimiento de un sistema en situaciones extremas de similitud entre hablantes (véase § 2). En este estudio proponemos validar los sistemas también con los otros dos supuestos complejos contemplados.

En todas las condiciones experimentales, los parámetros de largo plazo seleccionados cubren las tres categorías de parámetros suprasegmentales: tono, cualidad de voz y duración. Los parámetros de mayor relevancia, o que mayor peso tienen en la configuración del nuevo sistema propuesto, son el coeficiente de variación del F0 (tono), el Grado (cualidad de voz) y el ritmo (duración). Es decir, esos parámetros son los que juegan un papel más importante en el *LLR* resultante de añadir parámetros de largo plazo al sistema base, teniendo en cuenta las tres condiciones experimentales.

Para el caso de las grabaciones en condiciones extremas, con el nuevo sistema se obtuvieron mejoras del 11.6% en la *EER* y del 12.8% en la $C_{llr\ min}$. La comparación de gemelos logró mejoras del 16.2% en la *EER* con respecto al sistema comercial *BATVOX* que emplea la metodología *GMM-UBM*, mientras que el nuevo sistema no mostró mejoras en la *EER* con respecto al sistema base. Este resultado puede deberse a la cantidad limitada de casos que incluye la base de datos de gemelos. No obstante, sí se evidenció una mejoría del 10.6% en la $C_{llr\ min}$, lo que implicaría que los parámetros de largo plazo analizados pueden aportar una información valiosa para distinguir a un gemelo de otro. De todos los sistemas (véase la tabla 3) el mejor $C_{llr\ min}$ se obtiene con el nuevo sistema y para la comparación de gemelos, con un valor de 0.237. San Segundo y Yang (2019) obtienen valores $C_{llr\ min}$ muy parecidos (entre 0.15 y 0.30) cuando comparan gemelos usando un sistema semi-automático basado en trayectorias formánticas de secuencias vocálicas. Finalmente, en el caso de enmascaramiento por pinzamiento de nariz también se logró reducir la EER en un 15.3% y la $C_{llr\ min}$ en un 14.2%.

En definitiva, la magnitud media de la tasa de igual error –promediada para los tres casos complejos considerados– fue del 10.2%, valor de referencia a tener en cuenta para este tipo de cotejos en casos forenses. En cotejos de menor complejidad cabría esperar un rendimiento del sistema aún mejor, como se puede deducir del planteamiento anteriormente expuesto. Es decir, un *EER* en torno al 10% ocurriría de media en casos como los contemplados en nuestros tres supuestos experimentales extremos.

En cuanto al aporte de un tipo y otro de parámetros (de corto y de largo alcance), conviene recordar la división que remarcan varios autores (Laver, 1976; Nolan, 1983) entre factores intrínsecos –aquellos que están fuera del control del hablante y que vienen determinados por el tamaño y la forma de sus articuladores– y los factores extrínsecos, que serían aquellos que el hablante puede controlar con distintos fines (p. ej. afectivos, sociolingüísticos o pragmáticos). Ambos factores convergen en la mayoría de los parámetros fonéticos estudiados en la comparación forense de hablantes y son difícilmente separables. Como destaca Nolan (1983), las restricciones físicas del hablante no determinan valores acústicos absolutos en el plano acústico, sino que marcan los límites de la variación intralocutor. Con todo, es ampliamente aceptado que los parámetros de corto plazo usados en este estudio (*MFCC*) están estrechamente relacionados con la geometría vocal del hablante mientras que los rasgos de largo plazo seleccionados (suprasegmentales) no sólo capturan características físicas de cada individuo –rasgos intrínsecos– sino que también tienden a contener información de su manera característica de hablar (p. ej. prosodia, cualidad de voz, o timbre), revelando así no solo aspectos físicos sino aprendidos o extrínsecos. Por todo esto, la información que brinda el nuevo sistema permite mejorar la caracterización del hablante y, de este modo, la información detallada con la que contará el perito que lo utilice. En el caso de emplear adicionalmente métodos acústico-perceptivos, el sistema le permitirá corroborar su análisis o le ayudará a re-evaluar alguna característica a la que no haya prestado suficiente atención.

Por otra parte, el sistema multiparamétrico que hemos diseñado permite el desglose de información fonética, con el fin de presentar los resultados de la comparación entre hablantes de una manera más intuitiva que con un único resultado numérico. Así como Evett *et al.* (2000) sugieren presentar la fuerza de la evidencia en un formato textual, con este nuevo sistema se puede desarrollar dicha explicación al juez para su mejor comprensión. Por ejemplo, tomando uno de los cotejos de prueba del corpus de gemelos, con dos muestras de habla pertenecientes al mismo hablante (MZ01_1_1 vs MZ01_1_2), podemos expresar la fuerza de la evidencia como se muestra en la tabla 6. Si el perito utilizara, además, otras

metodologías como la perceptiva (p. ej. Hollien, 2002 o San Segundo y Mompeán, 2017), podría adicionar sus resultados al sistema paramétrico que acabamos de describir, lo cual proporcionaría aún mayor información sobre los hablantes comparados, para una redacción más completa del informe forense correspondiente.

En el apéndice se muestran los resultados de todas las comparaciones intralocutor (tabla 7) e interlocutor (tabla 8). La *LLR* media para las comparaciones de un mismo hablante, con un valor de 2.33, indica que, de media, es $10^{2.33}$ veces (unas 214 veces) más probable encontrar la evidencia forense (esto es, las características vocales del cotejo en cuestión) si es cierta la hipótesis de que pertenecen al mismo hablante que si es cierta la hipótesis de que pertenecen a hablantes distintos. La *LLR* de -2.30 que obtienen de media las comparaciones entre distintos hablantes indica que, en promedio, es $1/10^{-2.30}$ veces (unas 199 veces) más probable encontrar la evidencia forense si es cierta la hipótesis de que pertenecen a distintos hablantes que si es cierta la hipótesis de que pertenecen al mismo hablante. En cualquiera de los dos casos, estaríamos hablando de apoyo "moderadamente fuerte" a esas hipótesis (véanse los equivalentes verbales de *LR* en la guía del ENFSI; ENFSI, 2015).

Cotejo de voz	*Enfoques individuales*				*Enfoque Multi-para-métrico*	
	MFCCs	*Tono*	*Cualidad de voz*	*Duración*		
	1.20	0.18	0.27	0.38	2.03	*LLR*
MZ01_1_1 vs. MZ01_1_2	apoyo *moderado*, hipótesis mismo hablante	apoyo *débil*, hipótesis mismo hablante	apoyo *débil*, hipótesis mismo hablante	apoyo *débil*, hipótesis mismo hablante	apoyo *moderadamente fuerte*, hipótesis mismo hablante	Escala verbal

Tabla 6. *Presentación de la fuerza de la evidencia como valores de LLR y con una interpretación verbal aclaratoria. MZ0X_Y_Z se refiere a la pareja X (01-12) de gemelos, Y es el miembro de esa pareja (1 o 2) y Z es la sesión de grabación (1 o 2). En este caso se trata de una comparación intralocutor (mismo hablante).*

Finalmente, podemos concluir que cada caso complejo debe analizarse de manera independiente, empleando bases de datos ad hoc para el entrenamiento del sistema,

su calibración y la selección de parámetros de largo plazo. El aporte promedio de los parámetros de largo plazo al LLR_{final} fue de un 6.5%, siendo el restante 93.5% responsabilidad de los parámetros de corto plazo. Es decir, indudablemente el mayor aporte para la comparación forense de hablantes proviene de las características del tracto vocal y de los articuladores, mientras que los parámetros suprasegmentales cumplen una función secundaria, que –no obstante– puede resultar especialmente útil en casos complejos como los que hemos presentado en este trabajo.

6. REFERENCIAS

ADAMI, A.; L. BURGET, S. DUPONT, H. GARUDADRI, F. GREZL, H. HERMANSKY, P. JAIN, S. KAJAREKAR, N. MORGAN y S. SIVADAS (2002): «Qualcomm-ICSI-OGI features for ASR», en J. H. L. Hansen y B. Pellom (eds.), *Proceedings of the 7th International Conference on Spoken Language Processing* (ICSLP2002 - INTERSPEECH 2002), Denver, Colorado, USA, September 16-20, 2002, ISCA Archive, Vol. 1, pp. 4-7, http://www.isca-speech.org/archive/icslp02, [25/02/2019].

AITKEN, C. G. y M. LEESE (1995): *Statistics and the Evaluation of Evidence for Forensic Scientists*, Chichester, John Wiley & Sons.

AITKEN, C. G. y F. TARONI (2004): *Statistics and the Evaluation of Evidence for Forensic Scientists*, Chichester, John Wiley & Sons.

BERGER, C.; B. ROBERTSON y G. VIGNAUX (2010): «Interpreting scientific evidence», en I. Freckelton y H. Selby (eds.): *Expert Evidence*, Sydney, Thomson Reuters, pp. 9-28.

BHUTA, T.; L. PATRICK y J. D. GARNETT (2004): «Perceptual evaluation of voice quality and its correlation with acoustic measurements», *Journal of Voice*, 18(3), pp. 299-304.

BOERSMA, P. and D. WEENINK (2005): *Praat v.5.2.01*, www.praat.org [11/11/2012].

BONASTRE, J. F.; F. WILS y S. MEIGNIER (2005): «ALIZE, a free toolkit for speaker recognition», en *Proceedings of (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, 2005*, vol. 1, pp. 737-740.

BRÜMMER, N. (2004): «Application-independent evaluation of speaker detection», en *Proceedings of Odyssey-04: The ISCA Speaker and Language Recognition Workshop, Toledo, 2004*, pp. 33-40.

BRÜMMER, N. y E. DE VILLIERS (2011): «The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing», Documentation of BOSARIS toolkit.

BRÜMMER, N. y J. DU PREEZ (2006): «Application-independent evaluation of speaker detection», *Computer Speech & Language*, 20(2), pp. 230-275.

CHAMPOD, C. y D. MEUWLY (2000): «The inference of identity in forensic speaker recognition», *Speech Communication*, 31(2-3), pp.193-203.

CHAMPOD, C.; F. TARONI, A. BIEDERMANN y T. HICKS (2018): «Challenging Forensic Science: How Science should speak to Court?», Coursera Massive Open Online Course (MOOC), School of Criminal Justice, ESC, University of Lausanne, https://www.coursera.org/learn/challenging-forensic-science [25/02/2019].

DA COSTA FERNANDES, V. S. (2018): *Alterações acústicas e percetivas introduzidas nas vozes de indivíduos gémeos e devidas ao canal telefónico - Uma discussão de impacto na análise forense*, Tesis doctoral, Universidade do Porto.

DAVIS, S. y P. MERMELSTEIN (1980): «Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences», *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357-366.

DEHAK, N.; P. J. KENNY, R. DEHAK, P. DUMOUCHEL y P. OUELLET (2011): «Front-end factor analysis for speaker verification», *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 19(4), pp. 788-798.

DEJONCKERE, P. H.; M. REMACLE, E. FRESNEL-ELBAZ, V. WOISARD, L. CREVIER-BUCHMAN y B. MILLET (1996): «Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements», *Revue de Laryngologie-otologie-rhinologie*, 117(3), p. 219.

DE JONG, N. H. y T. WEMPE (2008): *Praat Script Syllable Nuclei (Praat script)*, https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2 [25/02/2019].

ENFSI, European Network of Forensic Science Institutes (2015): *ENFSI Guideline for Evaluative Reporting in Forensic Science*, http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf [19/02/2019].

EVETT, I. W. (1995): «Avoiding the transposed conditional», *Science & Justice*, 35(2), pp. 127-131.

EVETT, I. W.; G. JACKSON, J. A. LAMBERT y S. MCCROSSAN (2000): «The impact of the principles of evidence interpretation on the structure and content of statements», *Science & Justice*, 40(4), pp. 233-239.

GIL, J. y E. SAN SEGUNDO (2013): «El disimulo de la cualidad de voz en fonética judicial: un estudio perceptivo para un caso de hiponasalidad», en A. Penas (ed.): *Panorama de la fonética española actual*, Madrid, Arco / Libros, pp. 321-366.

GOLD, E. (2018): «Articulation rate as a speaker discriminant in British English», en *Proceedings of Interspeech 2018, 98th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India*, pp.1828-1832.

GONZÁLEZ-RODRÍGUEZ, J.; P. ROSE, D. RAMOS, D. TOLEDANO y J. ORTEGA-GARCÍA (2007): «Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition», *IEEE Transactions on Audio, Speech and Language Processing*, 15(7), pp. 2104-2115.

GURLEKIAN J. A. y N. MOLINA (2012): «Índice de perturbación, de precisión vocal y de grado de aprovechamiento de energía para la evaluación del riesgo vocal», *Revista de Logopedia, Foniatría y Audiología*, 32(4), pp. 156-163.

HANSEN J. H. y T. HASAN (2015): «Speaker recognition by machines and humans: A tutorial review», *Signal Processing Magazine, IEEE*, 32(6), pp. 74-99.

HAUTAMÄKI, R. G.; M. SAHIDULLAH, V. HAUTAMÄKI y T. KINNUNEN (2017): «Acoustical and perceptual study of voice disguise by age modification in speaker verification», *Speech Communication*, 95, pp. 1-15.

HERMANSKY, H. (1990): «Perceptual Linear Predictive (PLP) Analysis of Speech», *Foundations and Trends in Signal Processing*, 87(4), pp. 1738-1752.

HINTON, G.; L. DENG, D. YU, G. E. DAHL, A. R. MOHAMED, N. JAITLY, A. SENIOR, V. VANHOUCKE, P. NGUYEN, T. N. SAINATH y B. KINGSBURY (2012): «Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups», *Signal Processing Magazine, IEEE*, 29(6), pp. 82-97.

HIRANO, M. (1981): *Clinical Examination of Voice*, Vienna / New York: Springer.

HOLLIEN, H. F. (2002): *Forensic Voice Identification*, Academic Press, London / San Diego.

HUGHES, V.; P. HARRISON, P. FOULKES, P. FRENCH, C. KAVANAGH y E. SAN SEGUNDO (2017): «Mapping across feature spaces in forensic voice comparison: The contribution of auditory-based voice quality to (semi-) automatic system testing», en *Proceedings of the Annual Conference of the International Association of Speech Communication (Interspeech)*, August 2017, Stockholm, Sweden, pp. 3892-3896.

KÜNZEL, H. J. (2000): «Effects of voice disguise on speaking fundamental frequency», *International Journal of Speech Language and the Law*, 7(2), pp. 150-179.

LAVER, J. (1976): «The semiotic nature of phonetic data», *York Papers in Linguistics*, 6, pp. 55-62.

LEE, K. ; V. HAUTAMÄKI, T. KINNUNEN, A. LARCHER, C. ZHANG,A. NAUTSCH, T. STAFYLAKIS, G. LIU, M. ROUVIER, W. RAO, F. ALEGRE1, J. MA, M. W. MAK, A. K. SARKAR, H. DELGADO, R. SAEIDI, H. ARONOWITZ, A. SIZOV, H. SUN, T. H. NGUYEN, G. WANG, B. MA, V. VESTMAN, M. SAHIDULLAH, M. HALONEN, A. KANERVISTO, G. LE LAN, F. BAHMANINEZHAD, S. ISADSKIY, C. RATHGEB, C. BUSCH, G. TZIMIROPOULOS, Q. QIAN, Z. WANG, Q. ZHAO, T. WANG, H. LI, J. XUE, S. ZHU, R. JIN, T. ZHAO, P.-M. BOUSQUET, M. AJILI, W. B. KHEDER, D. MATROUF, Z. H. LIM, C. XU, H. XU, X. XIAO, E. S. CHNG, B. FAUVE, K. SRISKANDARAJA, V. SETHU, W. W. LIN, D. A. L. THOMSEN, Z.-H. TAN, M. TODISCO, N. EVANS, H. LI, J. H. L. HANSEN, J.-F. BONASTRE, E. AMBIKAIRAJAH (2017): «The I4U Mega Fusion and Collaboration for NIST Speaker Recognition Evaluation 2016», en *Proceedings of the Annual Conference of the*

International Association of Speech Communication (Interspeech), August 2017, Stockholm, Sweden.

LEHISTE, I. (1970): *Suprasegmentals*, Cambridge, MIT Press.

MAKHOUL, J. (1975): «Linear prediction: A tutorial review», *Proceedings of the IEEE*, 63(4), pp. 561-580.

MARTIN, D.; J. FITCH y V. WOLFE (1995): «Pathologic voice type and the acoustic prediction of severity», *Journal of Speech, Language, and Hearing Research*, 38(4), pp. 765-771.

MARTÍNEZ SOLER, M.; P. UNIVASO y J. GURLEKIAN (2018): «FORENSIA- Technical Specifications», DOI 10.13140/RG.2.2.36718.92488.

MASTHOFF, H. (1996): «A report on a voice disguise experiment», *Forensic Linguistics*, 3, pp. 160-167.

MCLAREN, M.; L. FERRER, D. CASTAN y A. LAWSON (2016): «The Speakers in the Wild (SITW) Speaker Recognition Database», en *Proceedings of Interspeech 2016*, San Francisco, ISCA, pp. 818-822.

MEUWLY, D. (2006): «Forensic individualisation from biometric data», *Science and Justice*, 46(4), pp. 205-213.

MORRISON, G. S. (2009a): «Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs», *The Journal of the Acoustical Society of America*, 125(4), pp. 2387-2397.

MORRISON, G. S. (2009b): «Forensic voice comparison and the paradigm shift», *Science and Justice*, 49(4), pp. 298-308.

MORRISON, G. S.; P. ROSE y C. ZHANG (2012): «Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice», *Australian Journal of Forensic Sciences*, 44(2), pp. 155-167.

NOLAN, F. (1983): *The Phonetic Bases of Speaker Recognition*, Cambridge, Cambridge University Press.

RAMOS-CASTRO, D. (2007): *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems*, Tesis doctoral, Universidad Autónoma de Madrid.

ROSE, P. (2002). *Forensic Speaker Identification*, London, Taylor & Francis.

SABATIER, S. B.; M. R. TRESTER y J. M. DAWSON (2019): «Measurement of the impact of identical twin voices on automatic speaker recognition», *Measurement*, 134, pp. 385-389.

SAN SEGUNDO, E. (2013): «A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application», *Procedia-Social and Behavioral Sciences*, 95, pp. 59-67.

SAN SEGUNDO, E.; H. ALVES y M. FERNÁNDEZ TRINIDAD (2013): «CIVIL Corpus: Voice Quality for Speaker Forensic Comparison», *Procedia-Social and Behavioral Sciences*, 95, pp. 587-593.

SAN SEGUNDO, E. (2014): *Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics*, Tesis doctoral, Universidad Internacional Menéndez Pelayo.

SAN SEGUNDO, E. y H. KÜNZEL (2015): «Automatic speaker recognition of Spanish siblings: (monozygotic and dizygotic) twins and non-twin brothers», *Loquens*, 2(2), 021.

SAN SEGUNDO E. y J. A. MOMPEÁN (2017): «A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity», *Journal of Voice*, 31(5), 644-e11.

SAN SEGUNDO, E.; P. FOULKES, P. FRENCH, P. HARRISON, V. HUGHES y C. KAVANAGH (2018): «The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals», *Journal of the International Phonetic Association*, pp. 1-28.

SAN SEGUNDO, E. y J. YANG (2019): «Formant dynamics of Spanish vocalic sequences in related speakers: A forensic-voice-comparison investigation», *Journal of Phonetics*, 75, pp. 1-26.

SCHWEITZER, N. J. y M. J. SAKS (2007): «The CSI effect: Popular fiction about forensic science affects the public's expectations about real forensic science», *Jurimetrics*, 47, p. 357.

SILVERMAN, B. (1986): *Density Estimation for Statistics and Data Analysis*, London, Chapman and Hall.

THOMPSON, W. C. y E. L. SCHUMANN (1987): «Interpretation of statistical evidence in criminal trials», *Law and Human Behavior*, 11(3), pp. 167-187.

VAN LEEUWEN, D. y N. BRÜMMER (2007): «An introduction to application-independent evaluation of speaker recognition systems», en C. Müller (ed.), *Speaker Classification I: Fundamentals, Features, and Methods*, Heidelberg, Springer, pp. 330-353.

YOUNG, S.; G. EVERMANN, M. GALES, T. HAIN, D. KERSHAW, X. LIU, G. MOORE, J. ODELL, D. OLLASON, D. POVEY, V. VALTECH y P. WOOLAND (2006): *The HTK Book*, Cambridge, Cambridge University Press.

ZHOU, Z. H. (2012): *Ensemble Methods: Foundations and Algorithms*, London, Chapman and Hall.

APÉNDICE

Cotejo de voz		*Enfoques individuales*				*Enfoque multi-paramétrico*
		MFCC	*Tono*	*Cualidad de voz*	*Duración*	
MZ01_1_1	MZ01_1_2	1.20	0.18	0.27	0.38	2.03
MZ01_2_1	MZ01_2_2	2.10	0.19	0.27	0.40	2.95
MZ02_1_1	MZ02_1_2	1.80	0.17	0.27	0.32	2.56
MZ02_2_1	MZ02_2_2	0.30	0.18	0.27	0.35	1.10
MZ03_1_1	MZ03_1_2	2.30	0.17	0.27	0.37	3.11
MZ03_2_1	MZ03_2_2	2.30	0.19	0.28	0.32	3.09
MZ04_1_1	MZ04_1_2	2.60	0.20	0.27	0.38	3.45
MZ04_2_1	MZ04_2_2	0.40	0.17	0.27	0.02	0.86
MZ05_1_1	MZ05_1_2	2.20	0.18	0.26	0.38	3.02
MZ05_2_1	MZ05_2_2	1.30	0.16	0.27	0.35	2.09
MZ06_1_1	MZ06_1_2	0.90	0.19	0.28	0.36	1.74
MZ06_2_1	MZ06_2_2	-0.60	0.19	0.26	0.36	0.21
MZ07_1_1	MZ07_1_2	2.50	0.18	0.28	0.38	3.35
MZ07_2_1	MZ07_2_2	2.40	0.19	0.28	0.39	3.25
MZ08_1_1	MZ08_1_2	3.40	0.19	0.27	0.39	4.25
MZ08_2_1	MZ08_2_2	-0.40	0.17	0.27	-0.04	0.01
MZ09_1_1	MZ09_1_2	2.60	0.18	0.27	0.36	3.41
MZ09_2_1	MZ09_2_2	-1.10	0.19	0.27	0.23	-0.41
MZ10_1_1	MZ10_1_2	1.20	0.20	0.26	0.27	1.92
MZ10_2_1	MZ10_2_2	2.70	0.19	0.28	0.34	3.51
MZ11_1_1	MZ11_1_2	0.40	0.20	0.27	0.37	1.23
MZ11_2_1	MZ11_2_2	3.00	0.19	0.26	0.39	3.84
MZ12_1_1	MZ12_1_2	0.90	0.19	0.27	0.38	1.74
MZ12_2_1	MZ12_2_2	2.90	0.19	0.28	0.29	3.66
	Promedio	1.55	0.18	0.27	0.32	2.33

Tabla 7. *Comparación de los mismos hablantes en diferentes sesiones (valores en LLR).*

Cotejo de voz		*Enfoques individuales*				*Enfoque multi-paramétrico*
		MFCCs	*Tono*	*Cualidad de voz*	*Duración*	
MZ01_1_1	MZ01_2_1	-1.97	0.19	0.27	0.30	-1.22
MZ01_1_1	MZ01_2_2	-3.78	0.17	0.28	0.13	-3.20
MZ01_1_2	MZ01_2_1	-0.81	0.17	0.27	0.14	-0.22
MZ01_1_2	MZ01_2_2	-3.52	0.15	0.27	-0.11	-3.21
MZ02_1_1	MZ02_2_1	-1.56	0.19	0.28	0.37	-0.72
MZ02_1_1	MZ02_2_2	-1.92	0.18	0.28	0.33	-1.13
MZ02_1_2	MZ02_2_1	-2.67	0.18	0.28	0.37	-1.84
MZ02_1_2	MZ02_2_2	-3.61	0.17	0.28	0.40	-2.77
MZ03_1_1	MZ03_2_1	-1.84	0.17	0.27	0.38	-1.03
MZ03_1_1	MZ03_2_2	-2.66	0.19	0.27	0.30	-1.91
MZ03_1_2	MZ03_2_1	-1.14	0.17	0.27	0.37	-0.33
MZ03_1_2	MZ03_2_2	-2.60	0.19	0.27	0.31	-1.84
MZ04_1_1	MZ04_2_1	-3.88	0.19	0.22	0.38	-3.09
MZ04_1_1	MZ04_2_2	-2.91	0.20	0.23	0.25	-2.23
MZ04_1_2	MZ04_2_1	-3.50	0.19	0.26	0.26	-2.79
MZ04_1_2	MZ04_2_2	-3.05	0.18	0.27	0.38	-2.22
MZ05_1_1	MZ05_2_1	-2.77	0.19	0.27	0.01	-2.30
MZ05_1_1	MZ05_2_2	-4.98	0.19	0.27	0.00	-4.51
MZ05_1_2	MZ05_2_1	-3.62	0.20	0.27	0.08	-3.06
MZ05_1_2	MZ05_2_2	-5.74	0.20	0.28	0.06	-5.19
MZ06_1_1	MZ06_2_1	-2.86	0.18	0.27	-0.07	-2.48
MZ06_1_1	MZ06_2_2	-4.11	0.20	0.27	0.27	-3.37
MZ06_1_2	MZ06_2_1	-1.36	0.20	0.28	0.29	-0.59
MZ06_1_2	MZ06_2_2	-1.74	0.19	0.27	0.40	-0.87
MZ07_1_1	MZ07_2_1	-2.68	0.20	0.29	0.33	-1.86
MZ07_1_1	MZ07_2_2	-3.85	0.18	0.28	0.35	-3.04
MZ07_1_2	MZ07_2_1	-4.06	0.19	0.28	0.36	-3.24
MZ07_1_2	MZ07_2_2	-4.38	0.19	0.28	0.39	-3.52
MZ08_1_1	MZ08_2_1	-1.41	0.18	0.28	0.38	-0.58
MZ08_1_1	MZ08_2_2	1.34	0.15	0.28	-0.52	1.26
MZ08_1_2	MZ08_2_1	-1.79	0.18	0.27	0.37	-0.97
MZ08_1_2	MZ08_2_2	-0.66	0.18	0.28	-0.09	-0.29
MZ09_1_1	MZ09_2_1	-3.67	0.20	0.27	0.36	-2.84
MZ09_1_1	MZ09_2_2	-3.02	0.19	0.27	0.33	-2.23
MZ09_1_2	MZ09_2_1	-4.30	0.20	0.27	0.34	-3.50

MZ09_1_2	MZ09_2_2	-3.16	0.20	0.27	0.27	-2.42
MZ10_1_1	MZ10_2_1	-9.33	0.16	0.27	-0.52	-9.42
MZ10_1_1	MZ10_2_2	-6.76	0.19	0.24	0.01	-6.31
MZ10_1_2	MZ10_2_1	-7.78	0.20	0.25	-0.10	-7.43
MZ10_1_2	MZ10_2_2	-4.82	0.19	0.21	0.23	-4.18
MZ11_1_1	MZ11_2_1	-0.79	0.19	0.27	0.38	0.05
MZ11_1_1	MZ11_2_2	-1.17	0.20	0.27	0.25	-0.45
MZ11_1_2	MZ11_2_1	0.58	0.19	0.28	0.23	1.27
MZ11_1_2	MZ11_2_2	-0.03	0.16	0.28	-0.11	0.31
MZ12_1_1	MZ12_2_1	-3.09	0.20	0.22	0.35	-2.32
MZ12_1_1	MZ12_2_2	-3.17	0.15	0.24	0.27	-2.51
MZ12_1_2	MZ12_2_1	-2.86	0.19	0.23	0.40	-2.04
MZ12_1_2	MZ12_2_2	-2.68	0.18	0.25	0.35	-1.90
	Promedio	-2.96	0.18	0.27	0.21	-2.30

Tabla 8. *Comparación de diferentes hablantes (gemelos entre sí) para diferentes sesiones de grabación (valores en LLR).*