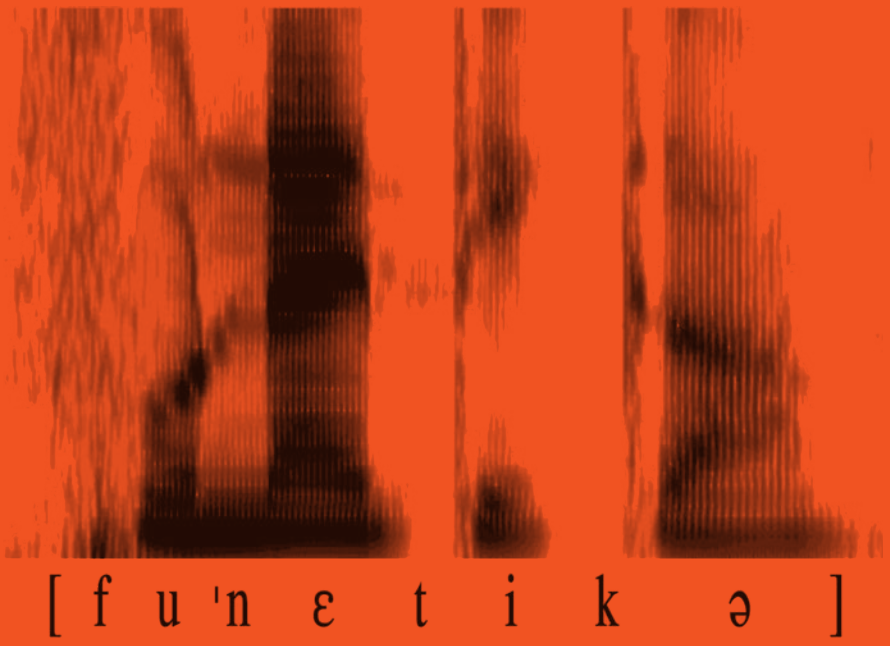
Estudios de Fonética Experimental

XXVIII



LF
Laboratori de Fonètica
UNIVERSITAT DE BARCELONA

Barcelona, 2019

ÍNDICE

Artículos

Miscelánea

Notas y reseñas

«Estudios de Fonética Experimental» informa

Anuncios

ARTÍCULOS

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