FONÉTICA CLÍNICA, FONÉTICA FORENSE
FONÉTICA APLICADA - TECNOLOGÍA DEL HABLA

LA CUALIDAD INDIVIDUAL DE LA VOZ Y LA IDENTIFICACIÓN DEL LOCUTOR: EL PROYECTO CIVIL*

HELENA ALVES/JUANA GIL/CAROLINA PÉREZ/EUGENIA SAN SEGUNDO**
Consejo Superior de Investigaciones Científicas

RESUMEN

En este trabajo se exponen los objetivos, la metodología y las contribuciones ya realizadas en el proyecto CIVIL: Cualidad individual de la voz e identificación del locutor. El propósito principal de este proyecto no es solo estudiar la cualidad de voz como tal y su rendimiento como parámetro identificador del hablante, sino también comprobar el grado en que puede verse alterada a voluntad del locutor hasta el punto de enmascarar su personalidad e impedir su reconocimiento. El equipo investigador se ha centrado únicamente en los rasgos laríngeos que determinan dicha cualidad, puesto que finalmente son estos los que los hablantes modifican con más frecuencia cuando pretenden disimular su voz. Asimismo, y como primera actuación para poder llevar acabo con garantía los análisis posteriores, se ha elaborado un protocolo de recogida de muestras indubitadas, tanto disimuladas como no disimuladas.

PALABRAS CLAVE: fonética judicial, cualidad de voz, modos de fonación, disimulo de la voz**.**

ABSTRACT

In this article, the aims, methodology and first contributions of the CIVIL project (CSIC) are presented. The primary objectives pursued by this research are both to study voice quality and its performance as a useful parameter for speaker identification, and to assess the degree in which

* Este trabajo se ha realizado gracias a la subvención recibida del Ministerio de Economía y Competitividad (Plan Nacional de I+D+i, Ref. FFI2010-21690-C02-01).

** Eugenia San Segundo Fernández es beneficiaria de una beca-contrato del Programa Nacional de Formación de Profesorado Universitario (FPU) concedida por el Ministerio de Educación, con resolución del BOE del 11-07-2009.

it can be deliberately altered by speakers trying to disguise their personality. The research team has only focused on the laryngeal features determining vocal profile, since they are undoubtedly the most frequently modified by speakers with criminal purposes. Furthermore, as a first contribution to the field, a protocol has been elaborated to elicit and record both known and unknown (modal and disguised) voice samples, so that posterior comparisons and analyses be fully guaranteed.

KEYWORDS: Forensic Phonetics, voice quality, phonation modes, voice disguise.

1. ANTECEDENTES BIBLIOGRÁFICOS

De la revisión de la bibliografía existente sobre cualidad de voz en relación con la identificación de locutores, pese a ser escasa, se pueden extraer algunas conclusiones interesantes[1]. Los títulos analizados para el presente proyecto se han agrupado en tres apartados de acuerdo con un criterio temático: (a) el primer apartado, centrado en los estudios propiamente judiciales, recoge las opiniones expresadas en diversos trabajos por destacados fonetistas forenses con respecto al uso de los parámetros relacionados con la cualidad de voz, (b) el segundo apartado, al margen del ámbito forense, agrupa las numerosas referencias al concepto de cualidad vocal –como índice que marca la individualidad del hablante– que encontramos en muy diversos estudios sobre la voz, y (c) la tercera sección reúne aquellos estudios del área de las patologías de la voz de los cuales se puede extraer información que resulta de interés para la aplicación judicial del estudio de la cualidad de voz.

a) Dentro de las investigaciones fonéticas estrictamente forenses, algunos autores (cf., por ejemplo, Nolan, 2007) señalan las dificultades que conlleva emplear este parámetro en fonética judicial, debido en buena medida a la falta de unidad entre los fonetistas en cuanto a la concepción y el uso del término "cualidad de voz" en sus infor-

[1] En San Segundo (2012) se puede encontrar una revisión crítica más extensa de la bibliografía relevante sobre cualidad de voz aplicada a la fonética judicial.

mes periciales. En general, se insiste en que esta noción debe entenderse en un sentido amplio como un concepto que engloba tanto las características supralaríngeas como las laríngeas que emergen en el habla de una persona de forma recurrente, y se subraya la idea de que su estudio debería objetivarse a partir de los hallazgos acústicos cuantitativos, apoyados además por el análisis perceptivo (Nolan, 2005).

Como es sabido, cabe distinguir dos ámbitos principales de aplicación de la fonética judicial (cf. Gil, este mismo volumen): el reconocimiento perceptivo de hablantes por testigos (*earwitness evidence*) y la comparación forense de muestras de habla desde un punto de vista acústico. Los estudios que aquí nos interesan, esto es, los inscritos en el segundo ámbito, algunos de ellos muy recientes (Nolan *et ál.*, 2011), inciden en el creciente interés por la aplicación de análisis sistemáticos de la cualidad de voz en la práctica forense, y llevan a cabo estudios experimentales para tratar de encontrar una correlación entre ciertas dimensiones pseudoperceptivas y determinados parámetros acústicos que capturen las características relacionadas con la cualidad de voz. En resumen, casi todos los fonetistas forenses más eminentes, dedicados a la práctica pericial, a la investigación básica en este campo, o a ambos aspectos, recogen en algún apartado de sus obras capitales (French, 1994; Jessen, 1997; Künzel, 1987; Nolan, 1983; Rose, 2002) alguna referencia, con mayor o menor extensión, al concepto de "cualidad de voz" o a algún parámetro específico relacionado con ella, destacando su importancia como posible clave identificativa del hablante. Debido a que los diversos autores que utilizan el concepto de "cualidad de voz" lo hacen, en ocasiones y como ya se ha dicho, de manera muy diferente, es necesario un análisis más detallado de cada uno de los estudios anteriormente mencionados, que constituyen obras de referencia en fonética judicial, para matizar la posición de cada uno de ellos (véase San Segundo, 2012).

b) En el segundo apartado de nuestra clasificación temática se incluyen varios estudios que se centran tanto en las características glóticas de hablantes determinados (Hanson 1996) como en otros aspectos relacionados con la cualidad de voz (Pittam, 1987), y que, sin hacer especial mención a la fonética judicial, sí que incluyen múltiples referencias a la potencial capacidad identificativa de los parámetros que analizan; en otras palabras, aluden en algún momento a la posible aplicación forense de los resultados de sus estudios, sin ser este el principal objetivo de sus investigaciones. Por ejemplo, Pittam (1987), que se interesa por el uso del espectro medio (*Long Term Average Spectrum –LTAS*) como procedimiento para discriminar distintos tipos de cualidad de voz, sostiene que, como un rasgo vocal a largo plazo, la cualidad de voz puede funcionar como una marca de individualidad y como portadora de características de la personalidad. También en relación con el espectro medio, el estudio de Harmegnies y Landercy (1988) tiene repercusiones en el ámbito forense, ya que su objetivo es investigar la variación intralocutor del espectro medio. Partiendo de la base de que el espectro medio constituye una pista acústica fiable para caracterizar la cualidad de voz, comparan varias realizaciones orales de un mismo texto llevadas a cabo por un mismo hablante y llegan a la conclusión de que existen diferencias significativas entre locutores. Existirían lo que ellos denominan "hablantes coherentes", en el sentido de que sus distintas producciones de un mismo texto obtienen una alta correlación (sus espectros medios, por tanto, son muy similares), y "hablantes no coherentes", que serían aquellos cuyas producciones orales apenas exhiben correlación en lo que al espectro medio se refiere.

c) Finalmente, cabe mencionar en este tercer apartado a varios autores que señalan la importancia de estudiar en el ámbito judicial ciertos parámetros que tradicionalmente se han usado para la descripción de las voces patológicas, como el *jitter* y el *shimmer* (Farrús et ál. 2007). Su punto de

partida es el siguiente: ya que estos parámetros caracterizan algunos aspectos de ciertas voces, también podría esperarse encontrar diferencias inter-locutores en los valores de los mismos, por ejemplo en los parámetros de la onda glótica. Gutiérrez-Arriola et ál. (2003), a fin de obtener los rasgos más importantes en lo que concierne a la identidad del hablante, listan algunos parámetros de la fuente glótica, como el cociente de apertura, el coeficiente de velocidad, la asimetría del cierre glótico y el coeficiente de retorno. En un campo de investigación ubicado en la intersección del estudio de las patologías de la voz y de la aplicación forense de la fonética, Gómez-Vilda et ál. (2007) proponen una metodología no invasiva para la estimación de los parámetros biomecánicos de las cuerdas vocales, de modo que se separe la información relacionada con la fuente glótica y la información del tracto vocal. Dicha metodología, útil tanto para una detección efectiva de patologías como para la caracterización biométrica de los hablantes (Gómez-Vilda et ál. 2008 y 2009, véase asimismo *infra*) ofrecería mejores resultados que la metodología basada en la combinación de fuente y filtro (*full voice*).

2. La cualidad de voz y los modos de fonación

Como se dijo más arriba, la cualidad de voz se deriva de los ajustes laríngeos y supralaríngeos que caracterizan el habla de un locutor. En el proyecto *CIVIL* –cuyo objetivo no es solo estudiar la cualidad de voz como tal sino también comprobar el grado en que esta puede verse alterada a voluntad del hablante hasta el punto de enmascarar la personalidad e impedir el reconocimiento– el equipo investigador se ha centrado únicamente en los rasgos laríngeos que determinan dicha cualidad, puesto que finalmente son estos los que los hablantes modifican con más frecuencia cuando quieren disimular su voz (como sucede, por ejemplo, en el caso de las comunicaciones telefónicas con propósitos delictivos: extorsiones, petición de rescates, llamadas obscenas, etc.). Tales rasgos no están del todo bien es-

tudiados, al menos por lo que se refiere al mundo académico hispanohablante, de forma que una de las primeras tareas realizadas por el equipo ha sido tratar de aclarar, precisar y, en la medida de lo posible, sistematizar, las nociones relacionadas con la fonación y los términos empleados para denominarlas, tal y como se explica a continuación.

Los ajustes laríngeos de los que depende la cualidad de voz tienen que ver fundamentalmente con los *modos* o *tipos de fonación*[2]. Como es lógico y bien sabido, el tamaño de los pulmones de los distintos locutores, la longitud de su tráquea, la posición de la laringe en su cuello y el tamaño de sus pliegues vocales influyen en las características acústicas del sonido producido en cada caso[3]. Además de estas características anatómicas, el comportamiento fonador influye en el resultado acústico, es decir, en el tipo de fonación (Catford 1964; Laver 1980; Ladefoged *et ál.* 1988, entre otros). La voz de una misma persona no suena igual cuando susurra, cuando grita, o cuando canta una canción de cuna. Tres factores de comportamiento vocal condicionan la fonación: el grado de tensión longitudinal, el grado de compresión de los pliegues vocales en la línea media, y el grado de cierre de la válvula laríngea. Veamos a continuación la implicación de cada uno de ellos en la fonación.

a) Grado de tensión longitudinal. La elongación del pliegue vocal hace que aumente su tensión longitudinal. Esto incrementa la rigidez de sus capas más externas y provoca que el pliegue vocal vuelva antes a su posición original. Los músculos intrínsecos laríngeos que controlan la F0 son el Cricotiroideo (CT) y el Tiroaritenoideo (TA): a medida que el primero se contrae y el segundo se relaja, los pliegues vocales se elongan, aumenta la rigidez de

[2] Recuérdese que la fonación es la conversión en sonido del aire que proviene de los pulmones por la acción de los pliegues vocales, situados en la laringe: cuando el flujo aéreo tiene una determinada velocidad de salida, los pliegues vocales se ven atraídos hacia el centro. Se juntan (*fase de aducción*) y se separan (*fase de abducción*) interrumpiendo y permitiendo el flujo aéreo periódicamente. Son estos cambios de presión aérea los que generan el sonido.

todas sus capas y la frecuencia fundamental aumenta, debido a la "mayor fuerza de recuperación efectiva", en palabras de Titze (2000). Por el contrario, si el TA se contrae y el CT se relaja, los pliegues vocales se acortan, y la rigidez de las capas más externas disminuye. La frecuencia fundamental baja debido a la menor fuerza de recuperación: el pliegue vocal tarda más en volver a su posición original. También se puede dar una condición isométrica, en la que ambos músculos están activos pero no hay cambios en la longitud de los pliegues vocales (véanse Titze 2000 y Whalen *et ál.* 1999 para más detalles).

b) Grado de compresión. Producir una voz poco intensa obliga a expulsar el aire despacio; mientras que para producir una voz muy intensa necesitamos que el aire sea expulsado a gran velocidad. La fonación resultante depende, así, de la respuesta que den los pliegues vocales a los diferentes tipos de flujo aéreo: si es rápido, la aducción será fuerte y por tanto la voz tendrá una intensidad alta; si es lento, la aducción será floja y la voz tendrá una intensidad baja.
c) Grado de cierre. El hecho de que la glotis desaparezca totalmente o no durante la fase de aducción es uno de los signos más audibles en la fonación: si existe escape de aire, que se percibe como un ruido de fricción más o menos fuerte añadido al sonido vocal, el sonido vocal tendrá menos presencia y unas características acústicas determinadas: mayor declive espectral, con los armónicos graves mucho más amplificados que los agudos. En un espectrograma de banda estrecha, una voz soplada o aérea (lo que en el mundo anglosajón se conoce como *breathy voice*) se verá con zonas grises en el centro, sin las líneas horizontales que reflejan los armónicos. En una voz que no tiene escape aéreo, por el contrario, se reflejan con nitidez todos los armónicos.

Estos tres parámetros de comportamiento laríngeo definirán, pues, cómo se produce la fonación, independientemente de cuáles sean las características anatómicas del individuo.

a) El primer parámetro, es decir, la modificación de la frecuencia fundamental, nos sitúa ante el fenómeno ampliamente estudiado de los *registros de fonación* (Hirano, 1981; Hollien, 1974, entre otros) también llamados *mecanismos de fonación* (Roubeau *et al.*, 2009), cuyas características principales resumimos en los siguientes subapartados.

1. Voz "pulsada"

Justo en el momento en que los pliegues vocales se separan se produce el estallido acústico. Esa energía sonora va decayendo a lo largo del tiempo, y la subsiguiente apertura produce un nuevo estallido acústico. Tales estallidos acústicos se perciben de forma continuada si ocurren a una frecuencia superior a 70- 80 Hz: los cierres glóticos ocurren antes de que el sonido haya desaparecido por completo. Por eso cuando escuchamos una voz percibimos un sonido continuo, solo interrumpido por consonantes no sonoras o por pausas inspiratorias. Pero por debajo de esa frecuencia de 70-80 Hz, los momentos de contacto glótico llegan cuando el sonido ya ha desaparecido, y por eso se percibe por separado cada uno de esos estallidos de energía acústica, seguido de un intervalo de silencio. Así, la voz que ocurre a una frecuencia inferior a 70-80 Hz es percibida a impulsos, de ahí que se la denomine "voz pulsada", o "frito vocal" –en América se suele usar el término "vocal fry" (Hollien y Mitchell, 1968), mientras que en Europa, sobre todo en Reino Unido, se usan los términos "creak", y "creaky voice" (Laver, 1980), que se podrían traducir por *crepitación* y *voz crepitante*–. Veremos que, sin embargo, sí se puede producir –y percibir– crepitación en frecuencias más altas. En ese caso, se entenderá la crepitación como un ajuste añadido al registro básico en el que se produzca la fonación, y no como un registro en sí mismo[3].

[3] Por tanto, el equipo *CIVIL* llama convencionalmente "voz pulsada" al registro de voz más grave, por debajo de los 70 – 80 Hz, y "crepitación" o "voz crepitante" al producto de un ajuste laríngeo específico que puede darse en combinación con el registro *modal* o el *falsetto* (véase más adelante).

2. Voz modal

El término "modal" (que ha de interpretarse en su sentido estadístico, es decir, como derivado de la "moda") es acuñado por Hollien (1974) para evitar atribuir la condición de "normal" al tipo de fonación que se usa de forma espontánea en el modo de habla neutro; hablar de voz normal implicaría que los otros registros o tipos de fonación son "anormales". Este registro se caracteriza, de acuerdo con van den Berg (1968) por vibraciones amplias de unos pliegues vocales gruesos y cortos en tonos graves, con poca tensión longitudinal. En el marco teórico de *cuerpo-cubierta* (Titze, 2000), se explican estas amplias vibraciones porque la parte muscular del *cuerpo* –músculo tiroaritenoideo (TA)– provee la tensión longitudinal, mientras que todas las capas de la Lamina Propria y el epitelio quedan lo bastante laxas como para propagar la onda superficial de la mucosa. Este registro comporta además un contacto glótico completo en sentido longitudinal, que se produce con las inserciones de los pliegues vocales de ambos aritenoides en contacto antes de iniciarse la fonación. En esta configuración, no hay ruido turbulento audible, y la fuerza de aducción y compresión medial son moderadas, lo que se debe también a la configuración diferenciada entre el *cuerpo* y la *cubierta* de los pliegues vocales.

3. Voz falsetto

Cuando la F0 sube más allá de un determinado punto –que depende de cada locutor–, el músculo Tiroaritenoideo (TA) no puede seguir activándose en un pliegue vocal tan elongado, y se relaja. Si en la voz modal el TA está activo y crea un engrosamiento de la parte media del pliegue vocal, al retraerse en el registro *falsetto*, deja el borde libre del pliegue vocal más delgado. La tensión longitudinal es soportada por todas las capas de la Lamina Propria, mientras que la mucosa está relativamente laxa.

Entre la voz modal y la voz *falsetto* se produce una transi-

ción que, en el nivel perceptivo tiene que ver con el declive espectral. La voz modal es rica en armónicos, y su espectro es más bien plano. Por su parte, la voz *falsetto* es aflautada, tiene pocos armónicos y muy separados, y su espectro tiene un fuerte declive. Así, cuando un cantante o un hablante pasan de un registro modal a un registro *falsetto* se percibe un empobrecimiento del timbre de la voz, una pérdida de energía sonora.

b) El grado de compresión, por su parte, afectará en el plano acústico a la amplificación de los armónicos agudos. La rapidez con la que se interrumpe el flujo aéreo en la fase de aducción es lo que se supone que determina la intensidad de la voz (Rothenberg, 1973). Si la interrupción es rápida, el flujo se interrumpe bruscamente, luego el cambio de presión es también muy veloz: esto genera unos armónicos agudos muy amplificados. Si por el contrario la velocidad de la aducción es lenta, la lenta interrupción del flujo aéreo no favorece la amplificación de los armónicos más agudos. Los resultados acústicos posibles, según el grado de compresión, son los siguientes:

1. Voz apretada

La aducción es muy firme y duradera, y se suele producir "en bloque": los bordes inferior y superior de los pliegues vocales se juntan en la línea media simultáneamente. Su característica acústica es un espectro plano: armónicos agudos con casi el mismo nivel que los graves. El declive espectral de esta voz es de unos 6 dB por octava (es decir, cada armónico más agudo tiene una amplitud de 6 dB menos que el armónico anterior).

2. Voz modal

La aducción es moderada; primero se aduce el borde inferior y después el borde superior, por la configuración cuadrada de los bordes libres de los pliegues. El declive espectral generado por esta voz es de unos 12 dB por octava.

3. *Voz suave*
La aducción es leve, posiblemente solo contacta el borde superior de los pliegues, porque el músculo TA esté inactivo y retraído. Es un tipo de fonación que genera un fuerte declive espectral, de unos 18 dB por octava.

c) Por último, veamos cómo afecta a la fonación el grado de contacto en el plano longitudinal de los pliegues vocales (Södersten y Lindestad, 1990).

1. Susurro (whisper)
No existe contacto entre los pliegues vocales, y el aire sale emitiendo un fuerte ruido turbulento. No hay voz.

2. Voz aérea o soplada (breathy voice)
El flujo transglótico no se interrumpe por completo, debido a que los pliegues vocales permanecen abiertos por algún punto; son muy frecuentes dos tipos de abertura: el hiato posterior, que deja los aritenoides abducidos, y el hiato llamado "en ojal", que deja el tercio medio sin contacto. Este tipo de fonación está generalmente relacionado con cierta hipotonía laríngea. No hay demasiada tensión longitudinal.

3. Voz susurrada (whispery voice, *Solomon* et ál., *1989*)
De la misma forma que en la voz aérea o soplada, la voz susurrada proviene de una aducción incompleta de los pliegues vocales; la diferencia es que en este tipo de fonación, la tensión longitudinal de los pliegues vocales incrementa el ruido de fricción provocado por el flujo transglótico.
Otros productos fónicos, que en el modelo de Laver (1980) se consideran "tipos de fonación combinados", y se supone que provienen de configuraciones laríngeas complejas, son la voz *apretada*, la voz *crepitante* (de la que ya se habló más arriba, v. nota 3), la voz *ronca* o la voz *áspera*. En un análisis preliminar con el programa Glottex, desarrollado en la Universidad Politécnica de Madrid, se ha visto que la configuración laríngea cambia poco, por lo

que probablemente estos resultados acústicos provienen de ajustes supralaríngeos (Pérez Sanz y Gómez Vilda, en preparación).

Una vez definidos y precisados estos conceptos, el equipo de *CIVIL* ha concentrado su atención en dos de los registros mencionados, la crepitación (o *creak*) y el *falsetto*, porque estos son a los que más frecuentemente recurren los delincuentes cuando quieren disimular su voz, según se explica en el siguiente apartado.

3. El disimulo de la cualidad de voz

El disimulo voluntario de la cualidad de voz propia ha sido y es un auténtico quebradero de cabeza para los especialistas en fonética judicial. Es un aspecto muy relevante para la investigación en esta área porque se ha comprobado que puede estar presente hasta en un 20% de los casos de delincuencia, y que afecta en un alto grado el proceso de comparación y posible identificación del locutor (hasta en un 60%), tanto si se trata de un reconocimiento de la voz llevado a cabo por informantes legos en la materia como si son fonetistas expertos los que escuchan los mensajes, si bien es cierto que estos últimos suelen alcanzar mejores resultados.

En la actualidad, la alteración de la voz puede conseguirse electrónicamente, empleando *software* que por lo general o bien modifica la frecuencia de los segmentos sin cambiar el tiempo o bien altera la duración segmental sin cambiar al frecuencia. Como se adelantó arriba, en el proyecto CIVIL no se aborda este tipo de transformación de la voz y la investigación se circunscribe a procedimientos mecánicos, en concreto a aquellos que se relacionan con el tipo de fonación usado por el locutor. De cualquier forma, estos no son los únicos posibles ni los únicos documentados en los registros de que disponen los diferentes departamentos de criminalística nacionales e internacionales (véase en la Figura 1 los más frecuentemente utilizados), si bien se encuentran entre los más empleados de acuerdo con los datos estadísticos.

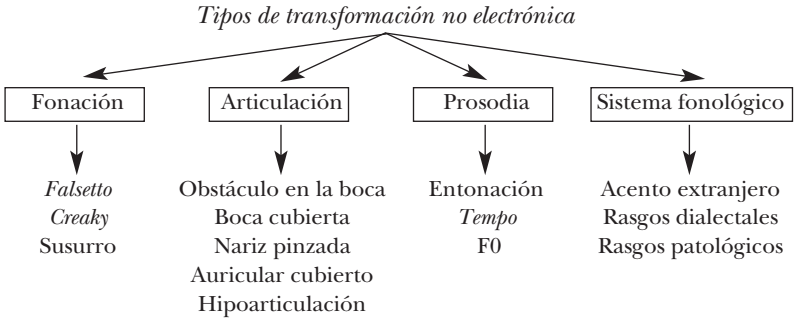

FIGURA 1. *Principales procedimientos utilizados con el fin de disimular la voz.*

Normalmente, los delincuentes solo se sirven de una única modalidad de transformación, y la estabilidad de su disimulo depende en última instancia de su habilidad, aunque es obvio que mantener la continuidad es el mayor problema con el que se encuentran, porque los rasgos alterados tienden a desaparecer con el transcurso de los minutos para dejar paso a las características auténticas de la voz.

En todos los casos, pues, el reto para el analista consiste en detectar en primer lugar el tipo de mecanismo de disimulo utilizado y, en segundo lugar, descubrir indicadores permanentes en la voz individual que sean lo suficientemente resistentes como para sobrevivir a los intentos de disimulo. Esta es precisamente la hipótesis de la que se parte en CIVIL: existen **marcadores individuales robustos** que toleran los intentos deliberados de distorsión de la cualidad de voz a través de la variación en el modo de fonación, en concreto a través de la fonación *creak* y *falsetto*, que son, como ya quedó dicho, las alteraciones que con más frecuencia se emplean en el disimulo con fines delictivos.

En relación con esta hipótesis inicial, el equipo de investigación aspira también a precisar el grado en que la distorsión producida por el canal de transmisión telefónico interfiere en la identificación perceptiva de las diversas cualidades de voz. Es sabido que el filtro del teléfono elimina información im-

portante[4], pero falta por dilucidar en qué medida puede o no afectar a esos supuestos rasgos permanentes de la voz, si se comprueba que existen realmente.

¿Qué se sabe hasta el momento sobre las consecuencias del empleo intencionado de la *creaky voice* y del *falsetto* para ocultar la propia personalidad? En primer lugar, conviene distinguir entre los estudios realizados mediante procedimientos de reconocimiento convencionales y los basados en técnicas automáticas y algoritmos estadísticos. En el primer grupo se inscriben obras como las clásicas de Masthoff (1996), Hirson y Duckworth (1993), Künzel (2000) o Rodman (2003). En el segundo, los más recientes trabajos de Künzel, González y Ortega (2004), Perrot, Aversano y Chollet (2007) o Perrot y Chollet (2008), entre otros. Circunscribiéndonos al primer enfoque, en el que se enmarca el proyecto CIVIL, los datos apuntan a que la proporción en la que el reconocimiento de la voz se ve alterado por estos modos de fonación, aun siendo siempre considerable, depende en menor o mayor medida de factores colaterales, como el grado de familiaridad existente entre hablante y oyente, el grado de estrés del oyente, el conocimiento compartido de la lengua, etc. (cf. Hollien, Majewski y Doherty, 1982). En todos los casos se han observado algunos cambios -no solo los esperables a nivel laríngeo sino también en la estructura formántica (Moosmüller 2007)– que explicarían el bajo nivel de reconocimiento de las voces disimuladas. Lo que no se ha investigado hasta el momento es la posibilidad de que exista alguna "huella biométrica" –en expresión del profesor Gómez Vilda– localizada en la onda glotal. El equipo de CIVIL aspira a descubrir esa peculiaridad fonatoria mediante el análisis electroglotográfico y sirviéndose del programa Glottex arriba mencionado.

[4] Por ejemplo, en la *breathy voice* el primer armónico aporta mucha información, pero en el caso de los hombres queda fuera de la banda de frecuencias que el filtro deja pasar. Asimismo, la presencia de ruido puede interferir en la percepción de esta cualidad de voz.

4. El protocolo de grabación

Uno de los problemas importantes con los que se tropiezan las personas que trabajan en fonética judicial es la carencia de directrices de actuación consensuadas por los especialistas que regulen el proceso de recogida de muestras de habla indubitada. Por ello, en el proyecto CIVIL se planteó como primer objetivo elaborar un protocolo al que pudieran ceñirse los supervisores de la toma de tales muestras (a menudo funcionarios judiciales) cuando no disponen del asesoramiento directo de un experto. Muchas de esas indicaciones son además extrapolables a la grabación de habla en laboratorio, con el propósito de constituir bases de datos y archivos sonoros en el curso de investigaciones puntuales, como la del propio proyecto CIVIL.

A continuación se resumen las recomendaciones más importantes que se incluyeron en el protocolo elaborado, cuyo desarrollo completo y fundamentación teórica se recogen en Gil, Alves y Hierro (en prensa):

a) Se recomienda obtener varias muestras de la voz indubitada del imputado, a ser posible separadas en el tiempo y tomadas en diferentes momentos del día.

b) La(s) grabación(es) indubitada(s) se debe(n) preparar con cierta anticipación, bajo asesoramiento de un experto, y tomando siempre como punto de partida la especificación de los parámetros potencialmente comparables presentes en las muestras de habla dubitada.

c) Ha de intentarse que en los registros obtenidos se localicen un buen número de observaciones (ocurrencias) de cada parámetro cotejable (lo ideal sería un número en torno a 30).

d) Se debe disponer de al menos 30 segundos (o 100 palabras) de habla dubitada e indubitada para que la fiabilidad de la comparación de voces no sea cuestionable. Esto implica grabar como mínimo 1 minuto de habla espontánea, 1 minuto de lectura de un texto preparado en fun-

ción de las muestras dubitadas disponibles, y 1 minuto de lectura de un texto fonéticamente equilibrado.

e) El equipamiento usado para todas las tomas de voz indubitada que se vayan a comparar debe ser siempre el mismo.

f) Se debería utilizar un micrófono de condensador con una respuesta en frecuencia plana de hasta 20kHz y directividad cardioide o hipercardioide. El mínimo aconsejado para la sensibilidad es 10mV/Pa y para el rango dinámico 100dB.

g) Una tarjeta de sonido externa da mejor resultado, siempre y cuando tenga una respuesta en frecuencia lo más lineal posible y permita escoger una frecuencia de muestreo adecuada para la grabación (44100Hz o 48000Hz). Es imprescindible que posea alimentación tipo Phantom para poder conectar un micrófono de condensador.

h) El software utilizado debería permitir la grabación de audio en formato no comprimido (wav con conversión tipo PCM), con una frecuencia de muestreo y una resolución mínimas de 44100Hz y 16 bits por muestra respectivamente.

i) Durante la grabación es muy importante poder realizar una escucha en tiempo real para comprobar que no hay ningún ruido indeseado ni se están produciendo saturaciones. Para ello se necesitarán unos cascos de tipo profesional.

j) La sala de grabación debe tener un tiempo de reverberación menor de 1 segundo y un nivel de ruido de fondo inferior a 35dB, con lo que se conseguirá una relación señal a ruido en torno a 30dB.

k) Tanto el micrófono como la tarjeta de sonido deberían estar lo más lejos posible de cualquier fuente de ruido, teniendo en cuenta que el primero debe estar colocado cerca del locutor (la distancia variará con el tipo de micrófono) y prestando especial atención a los posibles ruidos de respiración, golpes en la mesa o el suelo y movimiento de papeles.

l) La(s) grabación(es) indubitada(s) deben tratar de acentuar su grado de comparabilidad con las muestras dubitadas. Si en estas hay voz gritada, por ejemplo, es aconsejable conseguir que los sujetos implicados griten también en la toma controlada de voz espontánea, lo cual puede lograrse utilizando en la entrevista unos auriculares a través de los cuales los interlocutores –entrevistador y entrevistado- oigan música o ruido blanco a un nivel que no les cause daño ni les impida la retroalimentación auditiva, pero que les requiera gritar para hacerse oír.

m) Siempre que sea posible, es conveniente reunir muestras de habla de dos tipos, leídas y espontáneas.

n) La muestra leída puede conformarse a partir de fragmentos de la transcripción del habla dubitada, que se incorporarán al texto específicamente diseñado para la prueba, y con otro texto fonéticamente equilibrado. La muestra espontánea deberá prepararse siguiendo un procedimiento preparado de antemano y teniendo en cuenta todos los factores susceptibles de incrementar el grado de comparabilidad.

o) Si la muestra dubitada corresponde a habla telefónica interceptada, la indubitada debería grabarse en forma de una conversación teléfonica también filtrada entre el sospechoso y los funcionarios de turno situados en dos salas distintas de las dependencias policiales o judiciales.

p) El entrevistador que desee obtener muestras de habla lo más naturales posibles deberá plantear temas de conversación alejados por completo de los hechos delictivos por los que los imputados están siendo analizado, aunque provocando a la vez la aparición de determinados vocablos o rasgos auditiva o espectralmente interesantes.

q) Es conveniente crear un clima de relativa confianza con los entrevistados eligiendo asuntos sobre los que ellos tengan opiniones formadas fácilmente expresables (aficiones, experiencias, etc.).

5. El estado actual del proyecto

No existen en español bases de datos estandarizadas de habla disimulada. Por consiguiente, la primera labor está consistiendo en recolectar muestras de enunciados y procesarlos de acuerdo con las recomendaciones de registro apuntadas en el apartado anterior y con los procedimientos habituales y homologados para constituir una base de datos. En particular, se tomarán en cuenta los estándares avanzados en Boves, Bogaart y Bos (1994) y en di Carlo, Falcone y Paoloni (1994), así como las especificaciones publicadas por el Linguistic Data Consortium (LDC). El objetivo es contar con datos de 40 hablantes masculinos y 40 femeninos grabados en varias sesiones con diferentes métodos de disimulo mecánico del habla.

En la etapa actual de la investigación, y a partir de las muestras ya recogidas, se está llevando a cabo un estudio piloto inicial de tipo perceptivo, con pocos sujetos (6 mujeres de edades comprendidas entre los 25 y los 35 años), para tratar de validar las siguientes hipótesis de partida: a) El disimulo mediante el *falsetto* y el *creak* sí afecta al reconocimiento de voces (en nuestro caso, no familiares); b) aun así, las personas sin especial entrenamiento en fonética son capaces, mediante una tarea de discriminación[5], de identificar las voces disimuladas y asociarlas con los modelos originales no disimulados, porque hay rasgos de la voz resistentes a la variación que permanecen; y c) en todo caso, la habilidad para llevar a cabo el reconocimiento es mayor en el caso de los expertos en fonética que en el de los profanos. Se espera obtener los primeros resultados en muy breve plazo.

Con posterioridad, se procederá al análisis electroglotográfico y acústico de las voces que han superado mejor la distor-

---

[5] Para realizar este experimento perceptivo se ha diseñado una tarea con tripletes de tipo XBA, de modo que cada oyente escucha tres estímulos (tres voces) en cada serie y establece cuál de las dos últimas muestras de voz, A o B, es la misma o se parece más a la X. Esta técnica, denominada en inglés *matching-to-sample*, tiene muchas ventajas para un estudio como el que ahora está en curso. Para los detalles, cf. Alves et ál. (en preparación).

sión creada por el disimulo voluntario, con la intención de localizar los rasgos robustos que han ayudado a mantener su individualidad.

Referencias Bibliográficas

Boves, Louis/Bogaart, Tineke/Bos, Leonie (1994): "Design and recording of large data bases for use in speaker verification and identification", en: *ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*. Martigny, Suiza.

Catford, Ian (1964): "Phonation types: the classification of some laryngeal components of speech production", en: Abercrombie, David/Fry, Denis B./MacCarthy, Peter A. D. *et al.* (eds.): *In honour of Daniel Jones*. London: Longmans, Green and Co., 26-37.

Di Carlo, Andrea/Falcone, Mauro/Paoloni, Andrea (1994): "Corpus design for speaker recognition assessment", en: *ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*. Martigny, Suiza.

Farrús, Mireia/Hernando, Javier/Ejarque, Pascual (2007): "Jitter and shimmer measurements for speaker recognition", en: *European Conference on Speech Communication and Technology*. Antwerp, 778-781.

French, Peter (1994): "An overview of forensic phonetics with particular reference to speaker identification", en: *Forensic Linguistics* 1, 2, 169-181.

Gil, Juana/Alves, Helena/Hierro, José Antonio (en prensa): « Proposition raisonnée de protocole de capture de voix connue à des fins judiciaires », en: *Revue Internationale de Criminologie et de Police Technique et Scientifique*. Lausanne, Suiza.

Gobl, Christer/Ni Chasaide, Ailbhe (1992): "Acoustic characteristics of voice quality", en: *Speech Communication* 11, 4-5, 481-490.

Gómez-Vilda, Pedro/Álvarez-Marquina, Agustín/Mazaira-Fernández, Luis M. *et al.* (2008): "Decoupling vocal tract from glottal source estimates in speaker's identification", en: Pamies, Antonio/Melguizo, Elizabeth (eds.): *New Trends in Experimental Phonetics. Language Design,* Special Issue, 111-118.

Gómez-Vilda, Pedro/Fernández-Baillo, Roberto/Nieto, Alberto *et al.* (2007): "Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters", en: *Journal of Voice* 21, 4, 450-476.

Gómez-Vilda, Pedro/Fernández-Baillo, Roberto Rodellar-Biarge, Victoria *et al.* (2009): "Glottal source biometrical signature for voice pathology detection", en: *Speech Communication* 51, 9, 759-781.

Gutiérrez-Arriola, Juana/Montero, Juan Manuel/Córdoba, Ricardo *et al.* (2003): "Analysis of Parameter Importance in Speaker Identity", en:

ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis. Geneva.

HANSON, HELEN M. (1997): "Glottal characteristics of female speakers: acoustic correlates", en: *Journal of the Acoustical Society of America* 101, 1, 466-481.

HARMEGNIES, BERNARD/LANDERCY, ALBERT (1988): "Intra-speaker variability of the long term speech spectrum", en: *Speech Communication* 7, 81-86.

HIRANO, MINORU (1981): *Clinical Examination of Voice*. New York: Springer-Verlag.

HIRSON, ALLEN/DUCKWORTH, MARTIN (1993): "Glottal fry and voice disguise: a case study in forensic phonetics", en: *Journal of Biomedical Engineering* 15, 193-200.

HOLLIEN, HARRY (1974): "On vocal registers", en: Journal of Phonetics 2, 125-143.

HOLLIEN, HARRY/MICHEL, JOHN F. (1968): "Vocal fry as a phonational register", en: *Journal of Speech and Hearing Research* 11, 3, 600.

HOLLIEN, HARRY/MAJEWSKI, WOJCIECH/DOCHERTY, E. THOMAS (1982): "Perceptual identification of voices under normal, stress and disguise speaking conditions", en: *Journal of Phonetics* 10, 139-148.

JESSEN, MICHAEL (1997): "Speaker-specific information in voice quality parameters", en: *Forensic Linguistics* 4, 1, 84-103.

KÜNZEL, HERMANN (1987): *Sprechererkennung: Grundzüge forensischer Sprachverarbeitung*. Heidelberg: Kriminalist Verlag.

KÜNZEL, HERMANN (2000): "Effects of voice disguise on speaking fundamental frequency", en: *Forensic Linguistics* 7, 149–179.

KÜNZEL, HERMANN/GONZÁLEZ, JOAQUÍN/ORTEGA, JAVIER (2004): "Effect of voice disguise on the performance of a forensic automatic speaker recognition system", en: *ODYSSEY 2004 - The Speaker and Language Recognition Workshop*. Toledo.

LADEFOGED, PETER/MADDIESON, IAN/JACKSON, MICHEL (1988): "Investigating phonation types in different languages", en: Fujimura, Osamu (ed.): *Vocal physiology: Voice production, mechanisms and functions*. London: Lippincott Williams & Wilkins, 297–317.

LAVER, JOHN (1980): *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

MASTHOFF, HERBERT R. (1996): "A report on a voice disguise experiment", en: *Forensic Linguistics* 3, 160-167.

MOOSMÜLLER, SYLVIA (2007): "The influence of creaky voice on formant frequency changes", en: *International Journal of Speech Language and the Law* 17, 1, 87-93.

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

NOLAN, FRANCIS (2007): "Voice quality and Forensic Speaker Identification", en*:* GOVOR: *Casopis za fonetiku* 24, 2, 111-128.

NOLAN, FRANCIS/FRENCH, PETER/MCDOUGALL, KIRSTY *et al.* (2011): "The role of voice quality 'settings' in perceived voice similarity", en: *International Association of Forensic Phonetics and Acoustics Conference* (*IAFPA 2011*). Viena.

PERROT, PATRICK/AVERSANO, GUIDO/CHOLLET, GÉRARD (2007): "Voice disguise and automatic detection: Review and perspectives", en: Stylianou, Yannis/Faundez, Marcos/Esposito, Anna (eds.): *Progress in Non-Linear Speech Processing. Lecture Notes in Computer Science*. Berlín: Springer Verlag, 101-117.

PITTAM, JEFFRY (1987): "The long-term spectral measurement of voice quality as a social and personality marker: a review", en: *Language and Speech* 30, 1.

RODMAN, ROBERT D. (2003): *Speaker Recognition of Disguised Voices: A Program for Research*. Raleigh: North Carolina State University.

ROSE, PHILIP (2002): *Forensic Speaker Identification*. London: Taylor and Francis.

ROTHENBERG, MARC (1973): "A new inverse filtering technique for deriving the glottal air flow waveform during voicing", en: *The Journal of the Acoustical Society of America* 53, 1632.

ROUBEAU, BERNARD/HENRICH, NATALIE/CASTELLENGO, MICHÈLE (2009): "Laryngeal vibratory mechanisms: the notion of vocal register revisited", en: *Journal of Voice* 23, 4, 425-438.

SAN SEGUNDO, EUGENIA (2012): "Fonética judicial y cualidad de voz: análisis crítico de la bibliografía relevante para la comparación forense de voces" (manuscrito).

SÖDERSTEN, MARIA/LINDESTAD, PER-AKE (1990): "Glottal closure and perceived breathiness during phonation in normally speaking subjects", en: *Journal of Speech and Hearing Reasearch* 33, 3, 601-611.

SOLOMON, NANCY P./MCCALL, GERALD N./TROSSET, MICHAEL W. *et al.* (1989): "Laryngeal configuration and constriction during two types of whispering", en: *Journal of Speech and Hearing Research* 32, 1, 161-174.

TITZE, INGO R. (2000): *Principles of Voice Production*. Iowa City: National Center for Voice and Speech.

VAN DEN BERG, JAN WILLEM (1968): "Mechanism of the larynx and the laryngeal vibrations", en: Malmberg, Bertil (ed.): *Manual of Phonetics*. London: North-Holland, 278-308.

WHALEN, DOUGLAS H./GICK, BRYAN/KUMADA, MASANOBU *et al.* (1999): "Cricothyroid activity in high and low vowels: Exploring the automaticity of intrinsic F0", en: *Journal of Phonetics* 27, 2, 125-142.