# El efecto del enojo en los procesos automatizados de identificación forense de personas locutoras basados en espectros del habla a largo plazo

**Ortega-Rodríguez, M.**[1]; **Solís-Sánchez, H.**[1]; **Valverde-Méndez, D.**[1,2]; **Venegas-Li, A.**[1,3]

[1] Escuela de Física y Centro de Investigaciones Geofísicas, Universidad de Costa Rica, San José, Costa Rica, {manuel.ortega, hugo.solis}@ucr.ac.cr
[2] Department of Physics, Princeton University, Princeton, Estados Unidos, dsmendez@princeton.edu
[3] Physics Department, University of California at Davis, Davis, Estados Unidos, avenegasli@ucdavis.edu

## Resumen

La identificación forense de locutores/locutoras ha considerado tradicionalmente acercamientos al problema basados en el análisis de espectros a largo plazo (varias decenas de segundos de duración). Estos acercamientos han demostrado ser especialmente robustos, en el sentido que siguen funcionando bien incluso si las grabaciones son cortas; además, el método no es sensible a cambios en la intensidad sonora de la muestra, y sigue funcionando bien en la presencia de ruido y de ancho de banda limitado. Por todo esto, constituye una de las técnicas preferidas para la identificación forense, junto con el análisis de formantes, la velocidad del habla y la determinación de la frecuencia fundamental. Se halla, sin embargo, que el estado de enojo produce una distorsión importante en la señal acústica para efectos del análisis de espectros del habla a largo plazo. Incluso si el nivel de enojo es solamente moderado, hay un desvío de los resultados cuantitativos de la identificación forense de personas locutoras que representa el 33 % de la distancia (en el espacio de correlación entre muestras) hacia una persona locutora totalmente distinta. Por tanto, se concluye que es importante tener cautela en el momento de aplicar este método.

**Palabras clave:** identificación forense de locutor y locutora, espectros a largo plazo, acústica forense, distorsiones emocionales, enojo.

**PACS:** 43.72.Uv, 43.72.Ar, 43.72.Fx.

## The effects of anger on automated long-term-spectra based speaker-identification

## Abstract

Forensic speaker identification has traditionally considered approaches based on long-term (a few tens of seconds) spectra analysis as especially robust. This is because they work well for short recordings, are not sensitive to changes in the intensity of the sample, and continue to function in the presence of noise and limited passband. Because of this, the long-term spectra approach is one of the preferred tools for forensic speaker identification, in addition to formant analysis, speed of speech, and determination of the fundamental frequency. However, we find that anger induces a significant distortion of the acoustic signal for long-term spectra analysis purposes. Even moderate anger offsets speaker identification results by 33% in the direction of a different speaker altogether (in the space of sample correlations). Therefore, caution should be exercised when applying this tool.

**Keywords:** automated speaker identification, long term spectra, forensic acoustics, emotional distortions, anger.

DOI: 10.55753/aev.v37e54.28

## 1. INTRODUCCIÓN

El propósito del presente artículo es estudiar cómo determinar cuantitativamente el efecto de las distorsiones causadas por estados emocionales (en particular, el de enojo) en el análisis de espectro medio del habla a largo plazo (conocido como LTS por sus siglas en inglés: *long term spectra*) para efectos de identificación forense de personas locutoras (IFL). (Ejemplos de artículos de expertos en el tema de la IFL vienen dados por Hollien [1] y Hollien [2].) El proceder de la presente investigación se hace mediante una metodología cuidadosa y reproducible, explicada más adelante. El objetivo del proceso de la IFL es el de identificar una persona hablante por medio del análisis de su voz, usualmente bajo condiciones que no son ideales. Uno de los desafíos fundamentales con los cuales la IFL tiene que lidiar es el de determinar si la variabilidad intralocutor es menor que la variabilidad interlocutor (lo cual es claramente deseable), y cómo esta relación se mantiene para diferentes condiciones (Hollien [3]). Entre las condiciones más comunes, se puede mencionar las distorsiones tecnológicas debidas al equipo usado para hacer las grabaciones, así como las distorsiones ambientales causadas por ruido o sonidos ásperos de fondo.

En particular, las personas hablantes pueden ser ellas mismas la fuente de la distorsión, pues puede haber una variedad de sentimientos tales como miedo, enojo y ansiedad (una situación probable, por ejemplo, en la IFL cuando la persona hablante podría estar cometiendo un crimen). Estas emociones desencadenan una modificación en la producción del habla que se manifiesta como un cambio en los valores de los parámetros de señal (tales como las frecuencias y la velocidad del habla) (Williams y Stevens [4]; Banse y Scherer [5]; Johnstone [6]). La producción de la voz consiste en pulsos de aire causados por la vibración de las cuerdas vocales (que son luego modificados por el tracto vocal supralaríngeo), de forma que los factores dominantes de vocalización son los patrones de respiración y la tensión variante de los músculos involucrados en el proceso. Como estos factores tienen una correlación alta con las emociones, es por tanto muy probable que dichos cambios de los factores vayan a ser detectables en la onda acústica (Scherer [7]).

Este tema, sin embargo, no ha sido estudiado de manera extensiva para ningún idioma, en gran parte porque hay restricciones éticas y metodológicas que hacen complicada la producción controlada de emociones fuertes. El consenso (Johnstone [6]) ha sido que condiciones controladas de laboratorio son viables solamente para estados emocionales de baja intensidad, para los cuales es más bien difícil notar un cambio en la efectividad de la IFL. Otra dificultad yace en cómo inducir la emoción buscada. Martin [8] brinda un panorama de algunas posibles técnicas usadas para generar emociones específicas, tales como música e imágenes emotivas, así como técnicas de auto generación tales como uso de la imaginación y memorias. Estos métodos se clasifican de acuerdo a cómo las emociones son producidas.

Para propósitos de la presente investigación, se seleccionó el método de recuerdos autobiográficos auto inducidos. En esta técnica, a los participantes (todos hombres, como se explica más abajo) se les pide que recuerden eventos emotivos con el fin de generar la emoción deseada. Si bien este no es el único método aplicable al problema en cuestión, fue seleccionado por su simplicidad y porque le permitía a los participantes tener privacidad mientras hacían las grabaciones. Además, la cualidad de experimento ciego se aseguraba haciendo que los participantes mismos determinaran su nivel de enojo (en una escala numérica), como se describe más adelante.

## 2. FUNDAMENTOS

En esta sección se discutirá la lógica del funcionamiento del proceso de identificación forense de personas locutoras mediante el empleo del análisis LTS, proceso que se designará de ahora en adelante con la abreviación IFLLTS. A pesar de que existen muchos marcadores (también conocidos como "vectores") que ayudan a distinguir las grabaciones de distintas personas hablantes, desde el trabajo pionero de Hollien y

Majewski [9] uno de los métodos más comunes empleados en la IFL es el análisis LTS (Kinnunen *et al.* [10]; Ortega-Rodríguez *et al.* [11]). El análisis LTS revela cuantitativamente el promedio temporal del timbre de la voz, la propiedad acústica que le permite a un oyente distinguir entre un clarinete y un violín (por citar un ejemplo) que estén tocando la misma nota musical (frecuencia) con la misma intensidad sonora. La distribución se obtiene computando la Transformada de Fourier de la señal sobre un período largo de tiempo, por ejemplo, unos 30 segundos. La idea es que la persona hablante tenga suficiente tiempo para moverse a través de todo el espacio de fase sonoro, es decir, que tenga oportunidad de pronunciar varias veces todos (o casi todos) los sonidos del español.

Este vector ha sido ampliamente estudiado en términos de qué tan eficiente es para identificar a la persona locutora, y se ha hallado que es uno de los más confiables, principalmente porque continúa funcionando incluso en la presencia de ruido y de ancho de banda limitado (Hollien [3]).

Uno de los desafíos a la hora de usar este vector es cómo definir la correlación entre dos espectros. Existen varias formas de atacar este problema. Entre las más comunes, se tiene el asignar un número específico a cada conjunto de datos de LTS (de acuerdo con algún algoritmo), o incluso inspección visual de las gráficas. Para propósitos de este artículo, sin embargo, hace falta una forma más sofisticada. Dos coeficientes de correlación fueron considerados para estos fines: la Desviación Estándar de las Diferencias de Distribución (o SDDD, or sus siglas en inglés: *Standard Deviation of the Differences Distribution*) (Harmegnies [12]) y la el coeficiente de correlación cruzada de Bravais-Pearson, $R$ (Stanton [13]). Varios experimentos de carácter exploratorio llevados a cabo por nuestro grupo sin el componente de enojo (Ortega-Rodríguez *et al.* [11]) mostraron que el coeficiente de correlación de Bravais-Pearson es el que da el mejores resultados para esta línea de investigación, y fue por ende seleccionado. (El criterio empleado en esta decisión fue el siguiente: se considera un método como superior a otro cuando la correlación entre muestras del mismo hablante se acerca más a 1, y la correlación cruzada (distintos hablantes) se aleja de 1.)

En el método de Bravais-Pearson, el espectro del análisis LTS se considera como un vector de dimensión $k$, con un total de $k$ canales de frecuencia. El espectro puede definirse entonces así:

$$S \equiv (S_1, S_2, \ldots, S_i, \ldots, S_k)\,, \qquad (1)$$

en donde $S_i$ es el nivel de la $i$-ésima componente de frecuencia (Harmegnies [12]). En este contexto, el coeficiente $R$ mide qué tan relacionadas están las dos muestras LTS. $R$ se define así:

$$R_{SS'} \equiv \frac{1}{k} \frac{\sum_{i=1}^{k} (S_i - M_S)\,(S_{i'} - M_{S'})}{\sigma_S\, \sigma_{S'}}\,, \qquad (2)$$

en donde $M_S$ y $M_{S'}$ se refieren a las medias de cada espectro, en tanto que $\sigma_S$ y $\sigma_{S'}$ son las respectivas desviaciones estándar. El coeficiente de Bravais-Pearson posee varias ventajas, pues no solamente tiene una gran capacidad discriminatoria, sino que es independiente de las diferencias en intensidad relativa entre los dos espectros (Harmegnies [12]). Esto permite comparar grabaciones que fueron realizadas bajo distintas condiciones ambientales o de posicionamiento del micrófono.

La mayor parte de los estudios sobre la relación entre el habla y las emociones se ha concentrado en la capacidad de distinguir entre distintos estados emocionales del hablante por medio del análisis de la señal acústica (Williams y Stevens [4]; Fuller [14]; Scherer [7]; Johnstone [6]; Harnsberger *et al.* [15]). En particular, el análisis LTS ha sido usado para tratar de identificar estados emocionales y de depresión (Pittam [16]), aunque algunas veces se ha empleado filtrado humano en el proceso (Banse y Scherer [5]).

Menos común es la investigación de cómo las emociones o la intención deliberada de fingir la voz afectan la IFL. Rodman y Powell [17] recomiendan y planean investigación para estudiar los efectos de enmascaramiento en la IFL, aunque no la llevan a cabo. Más en relación com el presente artículo, Hollien y Majewski [9] estu-

dian el efecto que tiene el estrés (inducido mediante electrochoques en los sujetos) en el análisis LTS de la señal, pero ellos no encuentran impactos importantes para efectos de realizar una IFL adecuada. Exceptuando lo mencionado, no existen, hasta donde los autores y autoras del presente artículo saben, estudios del efecto de las emociones en la IFLLTS.

## 3. MATERIALES Y MÉTODOS

En esta sección se describirá la manera en la cual se definió la población de estudio y la forma de ejecución de las respectivas grabaciones.

### 3.1 Grabaciones de los sujetos

El problema de la IFL tiene muchas variables. Considérese, por ejemplo, el género, el origen geográfico y la edad de los sujetos (Hertrich y Ziegelmayer [18]; Linville [19]; Hollien y Majewski [9]; Pittam [16]; Yüksel y Gündüz [20]). Por este motivo, se escogieron sujetos con características similares. De esta manera se buscó poner en primer plano el efecto de la emoción por encima de otras variables, y reducir así la complejidad general del problema. Los sujetos cumplían los siguientes criterios: todos eran hombres en el intervalo de edad 18–25, y su origen geográfico es la Gran Área Metropolitana del Valle Central de Costa Rica (esta región está compuesta por las cuatro ciudades más grandes del país, localizadas en la región central del país, que es la que tiene la densidad de población más alta). Todos los sujetos hicieron su escuela primaria en esta región.

### 3.2 Condiciones de grabación

Las condiciones de grabación fueron homogeneizadas en la medida de lo posible con el fin de tratar de obtener los mejores resultados. Cada una de las muestras de habla fue grabada en un lugar privado en donde los sujetos se sintieran cómodos. Además, se solicitó que el habla fuera fluida y espontánea, y no recitada o aprendida. Se emplearon micrófonos de *smartphone* de la mejor calidad posible, tales como los de iPhone y Samsung Galaxy.

Según el estado emocional, hubo dos tipos de grabación: grabaciones normales, en las cuales a los sujetos se les pedía que hablaran de su vida cotidiana con el fin de tener la menor cantidad de respuesta emotiva posible; y grabaciones con enojo, en las cuales a los sujetos se les pedía que se provocaran un estado de enojo describiendo una situación personal que los haya enojado. La entrevistadora dejó al sujeto solo en esta parte con el fin de evitar que el sujeto se inhibiera.

La longitud de las muestras fue de 45 y 60 segundos para los casos normal y de enojo, respectivamente (el tiempo adicional para los casos de enojo permite que haya un tiempo de transición).

## 4. METODOLOGÍA

Un total de 32 sujetos (que satisfacían los criterios de selección mencionados anteriormente) fueron entrevistados. Este número fue escogido para tener una muestra estadísticamente significativa (National Institute of Standards and Technology [21]) de la Gran Área Metropolitana del Valle Central de Costa Rica. Cada entrevista consistió en tres grabaciones: Para 16 de los 32 entrevistados, se implementó el siguiente orden de grabación: normal-normal-enojado. Para los otros 16 entrevistados, se usó, en cambio, el orden normal-enojado-normal. La motivación detrás de hacer dos tipos de ordenamiento de las grabaciones es el de reducir los efectos de posibles errores sistemáticos debidos al orden de muestreo.

Los sujetos fueron conducidos a un lugar privado en donde se les leyó un conjunto de instrucciones; cada sujeto hizo su grabación de manera separada de los demás. La entrevistadora le explicó a cada sujeto que la experiencia formaba parte de un proyecto de investigación de la Universidad de Costa Rica y que los contenidos de las grabaciones no iban a ser usados ni escuchados. Se enfatizó que solamente las propiedades acústicas de las muestras eran de interés, y que el nivel de enojo sería determinado únicamente mediante su propia autoevaluación al final del proceso de grabación (esta característica vuelve el proceso en un experimento ciego). Además,

la naturaleza del proyecto fue descrita únicamente en términos muy generales para que la producción de voz fuera lo más natural posible.

En el caso de las grabaciones normales, a los hablantes se les dio la instrucción de hablar por 45 segundos sobre su vida (por ejemplo, su día, su mascota, un evento reciente, etc.). En el caso de las grabaciones con enojo, se les pidió hablar durante un minuto sobre algún tema que los hiciera enojar, por ejemplo, alguna persona con quien no se llevaran bien o algún evento que los haya irritado mucho. A los sujetos se les indicó que no fingieran la emoción (por ejemplo, forzándola levantando la voz). Se les dejó a solas con la grabadora y se les dijo que hablaran cuando estuvieran listos. Una vez concluida la grabación con enojo, se le pidió a cada participante que clasificara con un valor numérico en una escala de 1 a 5 su nivel de enojo, de tal manera que "1" significaba "no fui capaz de enojarme del todo", "3" significaba "moderadamente enojado", y "5" significaba "furioso".

Note que cuando se usó el primer ordenamiento de grabaciones (normal-normal-enojado), los sujetos no sabían sobre la parte de enojo hasta la última grabación, mientras que en el otro ordenamiento (normal-enojado-normal), el hablante ya sabía sobre este aspecto de la investigación durante su última grabación. Como se mencionó anteriormente, ambos ordenamientos fueron usados y promediados para reducir cualquier posible sesgo sistemático.

## 5. PROCESAMIENTO DE DATOS

El procesamiento de datos en el cual la presente investigación está basada fue desarrollado por nuestro grupo (Ortega-Rodríguez *et al.* [11]) estudiando la IFL (sin el componente de enojo) y ha sido extensivamente puesto a prueba y optimizado para asegurar los mejores resultados de identificación. Por optimización nos referimos a probar distintos tipos de ventanas temporales (por ejemplo, rectangular, gaussiana, Welch, Hanning, Hamming, etc.) y distintos tipos de tasa de muestreo para determinar cuáles funcionan mejor según el criterio mencionado en la Sección 2.

El procesamiento de datos que corresponde al presente artículo puede resumirse así. Segmentos de 30 segundos son obtenidos a partir de las grabaciones originales. En el caso de las grabaciones con estado de enojo, este procedimiento es particularmente importante para eliminar la parte de transición de estado normal a estado de enojo. Seguidamente se empleó el *software* de procesamiento de audio Audacity 2.1.0 (Audacity Team [22]) para obtener la Transformada Rápida de Fourier (FFT por sus siglas en inglés) para cada grabación. Una ventana de Hanning fue usada con un muestreo de 4096 valores. La utilización de este parámetro obedece al hecho de que demostró dar los mejores resultados en nuestros estudios previos exploratorios. Cada FFT fue guardada como un archivo de texto para seguir siendo procesada. En las Figuras 1 y 2 pueden verse muestras de dichos espectros para los casos normal y enojado, respectivamente.

A continuación, un programa en C++ calculó el coeficiente de correlación de Bravais-Pearson para las muestras. Finalmente, se procedió a obtener el promedio, la desviación estándar (SD) de la muestra y el error estándar de la media (SE) para las mediciones de los coeficientes de correlación. Se efectuaron los dos casos: normal-normal-enojado y normal-enojado-normal.

## 6. RESULTADOS Y DISCUSIÓN

La Tabla 1 muestra los resultados de los cálculos descritos en la sección anterior.

**Tabla 1:** Resultados estadísticos del coeficiente de correlación de Bravais-Pearson *R* para el caso intrahablante. Un total de 32 sujetos participaron para el contraste entre los casos normal-normal y normal-enojado.

| | Coeficiente de correlación, caso normal-normal | Coeficiente de correlación, caso normal-enojado |
|---|---|---|
| Promedio | 0.950 | 0.934 |
| Desviación estándar | 0.028 | 0.037 |
| Error estándar de la media | 0.005 | 0.005 |

Después que los 32 sujetos fueron entrevistados,

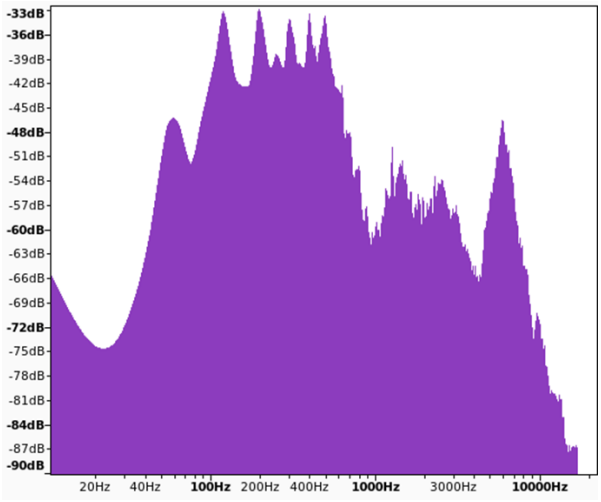


**Figura 1:** Espectro de una de las grabaciones correspondientes a habla normal (es decir, carente de enojo). La duración de la muestra es de 30 segundos, y el espectro fue producido empleando una ventana de Hanning con 4096 valores.

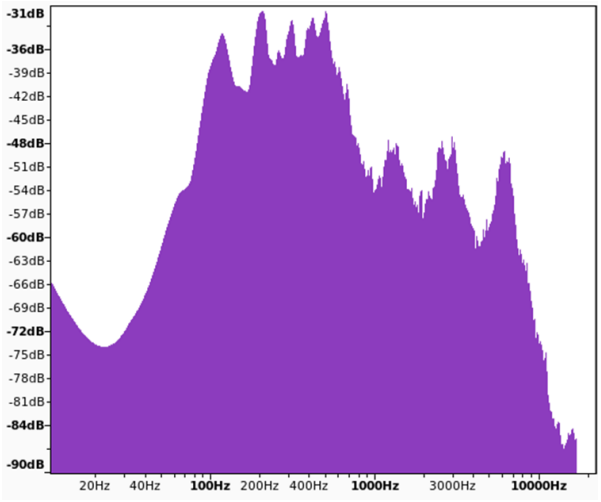


**Figura 2:** Espectro de una de las grabaciones correspondientes a habla con enojo de nivel 4 (ver la Tabla 2). La duración de la muestra es de 30 segundos, y el espectro fue producido empleando una ventana de Hanning con 4096 valores. El hablante es el mismo que para el caso de la Figura 1.

se procedió a calcular el coeficiente de correlación de Bravais-Pearson $R$ con el fin de obtener la correlación entre habla normal y normal, y la correlación entre habla normal y enojada (intra-hablante). Como se nota de la tabla, existe una diferencia notable entre los dos valores promediados. Para darse una idea de cuán significativa es esta diferencia, es útil compararla con los resultados del mencionado trabajo de nuestro grupo (Ortega-Rodríguez *et al.* [11]), el cual es una versión simplificada del proceso descrito en el presente artículo (pues el elemento de enojo no estaba presente), aunque las condiciones experimentales eran las mismas. En dicho trabajo, el coeficiente de correlación promediado $R$ entre dos hablantes distintos se midió en 0.890 con un SE de 0.010, mientras que la correlación entre mediciones del mismo hablante tuvo un $R$ promedio de 0.955 con un SE de 0.005.

Haciendo la comparación entre estas tres correlaciones: 0.950 (normal-normal, mismo hablante), 0.934 (normal-enojado, mismo hablante), y 0.890 (diferentes hablantes, ambos en modo normal), se llega a la conclusión de que los efectos de enojo desvían la señal un significativo 33 % de lo esperado en la dirección de otro hablante distinto. Esto es muy notable, puesto que el promedio de enojo autorreportado fue de tan solo 2.9 en la escala de 1 a 5 descrita anteriormente, lo cual significa que en promedio los sujetos estaban solo moderadamente enojados. La distribución del nivel de enojo para los sujetos puede apreciarse en la Tabla 2, y los respectivos resultados estadísticos se hallan en la Tabla 3.

**Tabla 2:** Enojo autorreportado por los sujetos; 1 significa “no fui capaz de enojarme del todo”, 3 significa “moderadamente enojado”, y 5 significaba “furioso”.

| Cantidad de sujetos | Nivel de enojo autorreportado |
|---|---|
| 5 | 4 |
| 18 | 3 |
| 9 | 2 |

**Tabla 3:** Resultados estadísticos para los datos de la Tabla 2.

| | Nivel de enojo autorreportado |
|---|---|
| Promedio | 2.90 |
| Desviación estándar | 0.65 |
| Error estándar de la media | 0.12 |

Es de esperar que mayores niveles de enojo generen mayores efectos en la IFLLTS. Para poner a prueba esta hipótesis, se procedió a filtrar los resultados obtenidos, dejando solamente los 5 individuos más enojados (aquellos con nivel 4 en la escala de 1 a 5). La Tabla 4 muestra los resultados obtenidos con esta condición. Para enojo fuerte, la desviación con respecto a la nor-

malidad es cercana al 50 % del camino a otro hablante.

**Tabla 4:** Resultados estadísticos intrahablante para el coeficiente de correlación Bravais-Pearson para el caso de las cinco grabaciones con mayor enojo. Como se esperaba, el enojo fuerte tiene un mayor efecto en la IFLLTS que el enojo moderado.

| | Coeficiente de correlación, caso normal-normal | Coeficiente de correlación, caso normal-enojado |
|---|---|---|
| Promedio | 0.950 | 0.922 |
| Desviación estándar | 0.030 | 0.047 |
| Error estándar de la media | 0.013 | 0.015 |

Esto muestra que los efectos de enojo en la IFLLTS sí crecen cuando el enojo aumenta.

## 7. CONCLUSIONES

Aunque algunos autores elogian el método de la IFLLTS por ser robusto ante estrés del hablante (Hollien y Majewski [9]), se ha encontrado en la presente investigación que existe una distorsión significativa en la voz humana debido al enojo para efectos de la IFLLTS. Incluso cuando la respuesta emocional de los participantes se quedó en lo moderado, se halla una diferencia apreciable en los coeficientes de correlación entre los casos de grabaciones normal-normal y normal-enojado. El enojo moderado desvía los resultados de la IFL por un 33 % en la dirección de otro hablante. Cabe recalcar que estos resultados fueron obtenidos con un método que es totalmente automatizable, brindando un acercamiento objetivo independiente de los errores humanos de percepción. El método también evita evaluar la sinceridad de los participantes, y es por tanto acorde con el código de práctica de la Asociación Internacional de Acústica y Fonética Forenses (The International Association for Forensic Phonetics and Acoustics [23]).

Los resultados del presente artículo son relevantes para la investigación forense, ya que el análisis LTS ha sido tradicionalmente considerado un vector robusto en la IFL, especialmente porque no es sensible a cambios en la intensidad sonora del habla, funciona bien para grabaciones cortas, y sigue funcionando en presencia de ruido y anchos de banda limitados. Los resultados del presente artículo, sin embargo, indican que se debe ser cuidadoso al usar la IFLLTS a la hora de calcular cocientes de verosimilitud (*likelihood ratios*) en un contexto de enojo, incluso si este enojo no es intenso. En aplicaciones forenses, se recomienda entonces registrar siempre el grado de enojo de la persona hablante, teniendo siempre presentes los valores de distorsión obtenidos en este artículo como referencia.

Puesto que otras emociones podrían también afectar significativamente la efectividad de la IFLLTS, su estudio en procesos de automatización es ampliamente justificado y recomendado. El estudio podría extenderse también a mujeres, o a hablantes de otros idiomas.

## AGRADECIMIENTOS

Este trabajo recibió el apoyo del proyecto 805-B2-175 de la Vicerrectoría de Investigación de la Universidad de Costa Rica, así como del Centro de Investigaciones Geofísicas de la misma universidad.

## CONTRIBUCIÓN DE LOS AUTORES Y LAS AUTORAS

Todas las personas autoras trabajaron conjuntamente en la generación del concepto general del presente artículo, y para definir su metodología. Todas ellas discutieron y aprobaron los resultados.

En particular, D. Valverde-Méndez y A. Venegas-Li se encargaron de efectuar las grabaciones, D. Valverde-Méndez realizó el procesamiento de datos e hizo la primera redacción del artículo (incluyendo la búsqueda de antecedentes), en tanto que M. Ortega-Rodríguez se encargó de la redacción final.

## CONFLICTO DE INTERESES

Los autores y las autoras declaran no tener conflicto de intereses con respecto al contenido de este artículo.

## REFERENCIAS

1. HOLLIEN, Harry. Barriers to Progress in Speaker Identification with Comments on the Trayvon Martin Case. *Linguistic Evidence in Security, Law and Intelligence*, University Library System, University of Pittsburgh, v. 1, n. 1, p. 76–98, dic. 2013. ISSN 2327-5596. doi: 10.5195/lesli.2013.3.

2. HOLLIEN, Harry. An Approach to Speaker Identification. *Journal of Forensic Sciences*, Wiley, v. 61, n. 2, p. 334–344, feb. 2016. doi: 10.1111/1556-4029.13034, pMID: 27404606.

3. HOLLIEN, Harry Francis. *Forensic Voice Identification*. Londres, Inglaterra: Academic Press, 2002. ISBN 0123526213.

4. WILLIAMS, Carl E.; STEVENS, Kenneth N. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, Acoustical Society of America (ASA), v. 52, n. 4B, p. 1238–1250, oct. 1972. doi: 10.1121/1.1913238.

5. BANSE, Rainer; SCHERER, Klaus R. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, American Psychological Association (APA), v. 70, n. 3, p. 614–636, 1996. doi: 10.1037/0022-3514.70.3.614.

6. JOHNSTONE, Tom. *The effect of emotion on voice production and speech acoustics*. Tesis (PhD) — University of Western Australia & University of Geneva, Perth, Australia, 2001. doi: https://doi.org/10.31237/osf.io/qd6hz.

7. SCHERER, Klaus R. Voice, Stress, and Emotion. In: ____. *Dynamics of Stress: Physiological, Psychological and Social Perspectives*. 1. ed. [S.l.]: Springer US, 1986. p. 157–179. ISBN 978-1-4684-5122-1. doi: 10.1007/978-1-4684-5122-1_9.

8. MARTIN, Maryanne. On the induction of mood. *Clinical Psychology Review*, Elsevier BV, v. 10, n. 6, p. 669–697, ene. 1990. ISSN 1873-7811. doi: 10.1016/0272-7358(90)90075-l.

9. HOLLIEN, Harry; MAJEWSKI, Wojciech. Speaker identification by long-term spectra under normal and distorted speech conditions. *The Journal of the Acoustical Society of America*, Acoustical Society of America (ASA), v. 62, n. 4, p. 975–980, oct. 1977. ISSN 1520-8524. doi: 10.1121/1.381592.

10. KINNUNEN, Tomi; HAUTAMAKI, Ville; FRANTI, Pasi. On the Use of Long-Term Average Spectrum in Automatic Speaker Recognition. In: *Proc. International Symposium on Chinese Spoken Language Processing*. [s.n.], 2006. p. 559–567. Disponible en: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d3b4740466aeb1d25831b6329599b615a5bab9b1.

11. ORTEGA-RODRIGUEZ, Manuel. *Informe Final: Articulación de un sistema de identificación de locutor con fines forenses*. [S.l.], 2016. Accedida en noviembre de 2021. Disponible en: https://hdl.handle.net/10669/85190.

12. HARMEGNIES, Bernard. SDDD: A new dissimilarity index for the comparison of speech spectra. *Pattern Recognition Letters*, Elsevier BV, v. 8, n. 3, p. 153–158, oct. 1988. ISSN 1872-7344. doi: 10.1016/0167-8655(88)90093-1.

13. STANTON, Jeffrey M. Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education*, Informa UK Limited, v. 9, n. 3, ene. 2001. ISSN 1069-1898. doi: 10.1080/10691898.2001.11910537.

14. FULLER, Fred H. *Detection of emotional stress by voice analysis final report*. Bethesda, Maryland, USA, 1972. Disponible en: https://www.ojp.gov/ncjrs/virtual-library/abstracts/detection-emotional-stress-voice-analysis-final-report.

15. HARNSBERGER, James D.; HOLLIEN, Harry; MARTIN, Camilo A.; HOLLIEN, Kevin A. Stress and Deception in Speech: Evaluating Layered Voice Analysis. *Journal of Forensic Sciences*, Wiley, v. 54, n. 3, p. 642–650, mayo 2009. ISSN 1556-4029. doi: 10.1111/j.1556-4029.2009.01026.x.

16. PITTAM, Jeffery. The Long-Term Spectral Measurement of Voice Quality as a Social and Personality Marker: A Review. *Language and Speech*, SAGE Publications, v. 30, n. 1, p. 1–12, ene. 1987. ISSN 1756-6053. doi: 10.1177/002383098703000101.

17. RODMAN, Robert D.; POWELL, Michael S. Computer Recognition of Speakers Who Disguise Their Voice. In: *The International Conference on Signal Processing Applications and Technology (ICSPAT 2000)*. [s.n.], 2000. Disponible en: https://api.semanticscholar.org/CorpusID:16980245.

18. HERTRICH, I.; ZIEGELMAYER, G. Sexual dimorphism in the long term speech spectrum. *Human Evolution*, Springer Science and Business Media LLC, v. 2, n. 3, p. 255–262, mayo 1987. doi: 10.1007/bf03016110.

19. LINVILLE, Sue Ellen. Source Characteristics of Aged Voice Assessed from Long-Term Average Spectra. *Journal of Voice*, Elsevier BV, v. 16, n. 4, p. 472–479, dic. 2002. doi: 10.1016/s0892-1997(02)00122-4.

20. YÜKSEL, Mustafa; GÜNDÜZ, Bülent. Long term average speech spectra of Turkish. *Logopedics Phoniatrics Vocology*, Informa UK Limited, v. 43, n. 3, p. 101–105, sep. 2017. doi: 10.1080/14015439.2017.1377286.

21. National Institute of Standards and Technology. *NIST/SEMATECH e-Handbook of Statistical Methods*. [s.n.], 2012. Accedida en octubre de 2021. Disponible en: https://www.itl.nist.gov/div898/handbook/prc/section2/prc222.htm.

22. Audacity Team. *Audacity (versión 2.1.0), editor y grabador de audio*. 2015. Disponible en: https://www.audacityteam.org/.

23. The International Association for Forensic Phonetics and Acoustics. *Code of Practice*. [S.l.], 2004. Accedida en enero de 2018. Disponible en: https://www.iafpa.net/the-association/code-of-practice/.