Revista Logos, Ciencia & Tecnología
ISSN 2145-549X (impreso)
ISSN 2422-4200 (en línea)

Artículo de revisión

# Influencia de las propiedades de los registros de audio en sistemas de verificación de hablantes en el contexto forense: una revisión del estado del arte

## Influence of the properties of audio records in speaker verification systems in a forensic context: a review of the state of the art

## Influência das propriedades dos registros de áudio em sistemas de verificação de falantes no contexto forense: uma revisão do estado da arte

Alexander Sepúlveda*
https://orcid.org/0000-0002-9643-5193 Universidad Industrial de Santander, Bucaramanga, Colombia

- **Fecha de recepción:** 2019-07-25
- **Fecha concepto de evaluación:** 2019-09-30
- **Fecha de aprobación:** 2019-10-15
http://dx.doi.org/10.22335/rlct.v11i3.982

**Para citar este artículo / To reference this article / Para citar este artigo:** Sepúlveda, A. (2019). Influencia de las propiedades de los registros de audio en sistemas de verificación de hablantes en el contexto forense: una revisión del estado del arte. *Revista Logos Ciencia & Tecnología*, *11*(3), 181-194. http://dx.doi.org/10.22335/rlct.v11i3.982

## RESUMEN

El procedimiento de verificación de hablantes (VH) en el campo forense ha de ser confiable. Sin embargo, su desempeño se ve afectado por propiedades intrínsecas de los registros de audio. En tal sentido, es importante analizar la afectación sobre los métodos de VH encontrados en el campo forense, a fin de estar en capacidad de llevar a cabo procedimientos más confiables en las diligencias forenses. En el presente artículo, el análisis se hace con base en trabajos reportados en el estado del arte, a partir del cual se encuentra que el desempeño del proceso de verificación depende de propiedades tales como tipo de codificación, longitud de audio, contenido de ruido, presencia de saturaciones y transitorios; donde el grado de afectación de estas propiedades depende del método de verificación que se utiliza. Aunque existen otros elementos que afectan el desempeño, en el presente trabajo se abordan los previamente mencionados. Según la revisión realizada, se nota una falencia de reportes acerca del grado de afectación en el caso de métodos diferentes al método automático, especialmente. Además, en cuanto a la influencia de la saturación del rango dinámico y de transitorios se encontró poca información reportada, lo cual dificulta establecer la influencia de las mismas.

**Palabras clave:** métodos de comparación de voces, acústica forense, codificación, verificación de hablantes, relación señal-ruido

* Autor para correspondencia. Correo electrónico: alexander.sepulveda@gmail.com

## SUMMARY

The procedure for verifying speakers (SV) in the forensic field must be reliable. However, the performance is affected by the intrinsic properties of audio records. In this regard, it is important to analyze the impact on the SV methods found in the forensic field in order to be able to carry out more reliable procedures in forensic proceedings. In this article, the analysis is based on studies reported in the state of the art, from which it is found that the performance of the verification process depends on properties such as type of coding, audio length, noise content, presence of saturations and transients, where the degree of affectation of these properties depends on the verification method that is used. Although there are other elements that affect performance, this research addresses the aforementioned. According to the review carried out, there is a lack of reports about the degree of affectation, especially in the case of methods other than the automatic method. In addition, regarding the influence of the saturation of the dynamic and transient range, not much reported information was found, which makes it difficult to establish their influence.


**Keywords:** voice comparison methods, forensic acoustics, coding, speaker verification, signal-to-noise ratio


## SUMÁRIO

O procedimento de verificação de falantes (VF) no campo forense deve ser confiável. Contudo, seu desempenho é afetado por propriedades intrínsecas dos registros de áudio. Neste sentido, é importante analisar a afetação sobre os métodos de VF encontrados no campo forense, com o fim de estar capacitado para realizar procedimentos mais confiáveis nas diligências forenses. Neste artigo, a análise é feita com base em trabalhos reportados no estado da arte, a partir do qual se constata que o desempenho do processo de verificação depende de propriedades tais como tipo de codificação, longitude de áudio, conteúdo de ruído, presença de saturações e transitórios; onde o grau de afetação destas propriedades depende do método de verificação utilizado. Ainda que existam outros elementos que afetam o desempenho, neste trabalho abordam-se os previamente mencionados. Segundo a revisão realizada, percebe-se uma falência de relatos sobre o grau de afetação em caso de métodos diferentes ao método automático, especialmente. Além disso, no que se refere à influência da saturação do rango dinâmico e de transitórios encontrou-se pouca informação reportada, o que dificulta estabelecer a influência das mesmas.


**Palavras-chave:** métodos de comparação de vozes, acústica forense, codificação, verificação de falantes, relação sinal-ruído

El rápido desarrollo de la comunicación facilita que la voz forme parte de actos criminales, al tiempo que facilita la grabación de la misma para su posterior inclusión en las diligencias judiciales. Esto provoca que el reconocimiento de hablantes sea una importante herramienta de apoyo en procedimientos forenses a la hora de reconocer el posible autor de determinado crimen. La tarea de reconocimiento del hablante abarca dos subáreas: identificación y verificación. La verificación del hablante busca, como su nombre lo indica, verificar si una voz cuestionada corresponde o no con una voz cuya fuente se conoce. Al proceso de verificación del hablante en Colombia también se le conoce como cotejo de voces y en Perú es común usar el nombre de homologación de voces. Gran parte de las grabaciones de voz provienen de interceptaciones telefónicas, las cuales están afectadas por el ruido aditivo, posibles distorsiones en el canal y fenómenos inherentes al proceso. Debido a estas propiedades que podrían llegar a afectar considerablemente la confiabilidad de los resultados de la tarea de verificación del hablante, resulta importante indagar en qué grado las afectaciones recién mencionadas podrían llegar a afectar el proceso de verificación de hablantes. Además, es común encontrar en los estrados judiciales preguntas relacionadas con la idoneidad de los registros antes de procesar una prueba que podría inculpar a una persona. De hecho, en Colombia se documenta la realización del análisis preliminar de

audios con el fin de determinar la idoneidad del material[1]. Sin embargo, también se reporta que el proceso de análisis preliminar de audios puede afectarse por elementos subjetivos[2] y, por ende, este procedimiento requiere de mayor objetividad.

En tal sentido, resulta de importancia práctica el contar con un protocolo adecuado que les permita a los practicantes de la acústica forense establecer la idoneidad de un registro de audio para propósitos de comparación forense de voces. La publicación de protocolos cuyo seguimiento permita determinar la idoneidad de los audios para el desarrollo de procesos de verificación del hablante de manera confiable es escasa. En particular, en Romito y Galatà (2004) se expone la importancia de contar con un protocolo para el análisis preliminar de audios, y además, menciona los pasos básicos que conforman el proceso de verificación de hablantes. En Barinov (2010) se presentan aquellas propiedades de los audios que afectan el análisis de las señales y su consecuente uso en la verificación. Además, Barinov (2010) propone valores de las propiedades catalogados como aceptables, obtenidos a partir de experimentos con métodos basados en expertos y métodos automáticos.

El grado de afectación depende principalmente del tipo de fenómeno en sí mismo, pero además del método y de los rasgos o características que se utilizan para representar la señal acústica de la voz. A su vez, estos rasgos se ven afectados en mayor o menor medida por fenómenos como el tipo de codificación, longitud del audio, contenido de ruido y distorsiones de la señal. Aunque puedan existir además otros fenómenos que afectan el desempeño de sistemas VH (verificación del hablante), en el presente trabajo nos enfocamos en aquellas propiedades que puedan ser medibles.

Asimismo, se presenta un análisis de los efectos de diferentes propiedades relacionadas con la etapa de registro de las señales de audio en sistemas de verificación del hablante. A modo de insumo, se utilizan reportes científicos previos. Este trabajo se desarrolla mediante las siguientes secciones: primero, se exponen los diferentes tipos de métodos utilizados para representar la información útil contenida en la señal de voz; segundo, se exponen los distintos tipos de métodos utilizados para la verificación de hablantes desde el punto de vista forense; tercero, se presentan aquellos parámetros que influyen en el desempeño de este tipo de sistemas; finalmente, se dan conclusiones generales.

## Características utilizadas en los métodos de verificación del hablante

Los parámetros utilizados en voz, idealmente, deberían cumplir con las siguientes condiciones (Rose, 2002): (a) gran variabilidad entre-hablantes y baja variabilidad intra-hablante; (b) ser robustos ante condiciones de ruido y distorsión por efectos del canal; (c) que ocurran de manera frecuente y natural en el habla; (d) fáciles de medir; (e) que sean difíciles de imitar por otras personas, y (f) que no se vean afectados por cambios en la salud y la edad. Una forma de clasificar las características es como sigue (Kinnunen & Li, 2010): 1) características de tiempo corto, y relacionadas con la fuente de voz; 2) características espectro-temporales, y 3) características prosódicas y de información de alto nivel.

## Características espectrales de tiempo corto y características relacionadas con la fuente

Las características de tiempo corto, como su nombre lo indica, son estimadas en segmentos cortos entre 20 y 30 ms. Aunque son fáciles de estimar y requieren de una menor longitud total de audio, estas características tienden a ser más afectadas por ruido y desacople en el canal (Kinnunen & Li, 2010). Dentro de las características de tiempo corto utilizadas en sistemas de verificación y reconocimiento del hablante tenemos: MFCC (*mel-frequency cepstrum coefficients*) (González-Rodríguez, Drygajlo, Ramos-Castro, García-Gomar & Ortega-García, 2006), los cuales, aun con el pasar de los años, es difícil encontrarles reemplazo. Los parámetros MFCC están relacionados con la envolvente espectral, la cual entrega información acerca de la forma tracto vocal y ha mostrado ser de bastante utilidad en propósitos de reconocimiento del hablante. Aunque no solo los MFCC entregan información de la forma del tracto vocal, están además los PLP (*perceptual linear prediction*), LPC (*linear predictive coding*) y el *cepstrum*, entre otros. Los MFCC son comúnmente utilizados en aplicaciones de software automáticas de tipo comercial.

1 En el proceso 34232, Corte Suprema de Justicia, Sala de Casación Penal, magistrado ponente Sigifredo Espinosa Pérez, 01/02/2012, se menciona la realización de un proceso que busca determinar si el material de audio es apto o no para cotejo de voces.

2 En el proceso 33120, Corte Suprema de Justicia, Sala de Casación Penal, magistrados ponentes Sigifredo Espinosa Pérez y Alfredo Gómez Quintero, Acta 374, del 3 de diciembre de 2009, se menciona que la fonoaudióloga perito se contradice.

Por otra parte, las características relacionadas con la fuente caracterizan el comportamiento de la señal de excitación glotal de sonidos del tipo sonoro (Kinnunen & Li, 2010). En experimentos previos se ha encontrado que el valor discriminante de las características asociadas a la fuente es menor que los asociados al tracto vocal; sin embargo, los primeros contribuyen a mejorar el desempeño del sistema como un todo (Zheng, Lee & Ching, 2007).

## Características espectro-temporales

Dentro de este tipo de características se destacan los formantes. Los formantes han sido preferiblemente utilizados en el método acústico-fonético, aunque también han sido eventualmente utilizados en métodos del tipo automático. En particular, la bondad de las trayectorias de los formantes en tareas de verificación del hablante ha sido probada bajo el nuevo paradigma de la razón de verosimilitud en Morrison (2009a). En general, el segundo formante tiende a verse menos afectado que los formantes primero y tercero debido a su ubicación dentro del ancho de banda de la voz. Los formantes cuarto y quinto simplemente no se utilizan debido a las consideraciones de ancho de banda del canal. En contraste, la estimación de la frecuencia fundamental se ve poco afectada por efectos del canal. Otro tipo de parámetros que han sido probados con éxito corresponde a los componentes principales de tiempo-frecuencia (Magrin-Chagnolleau, Durou & Bimbot, 2002), que consisten en una representación reducida en parámetros obtenida mediante la aplicación de análisis de componentes principales sobre los valores de energía de los átomos de tiempo-frecuencia calculado entre un tiempo $t_a$ y $t_b$.

## Características de alto nivel

Estas características son más robustas contra ruido y desacople del canal, pero son más difíciles de estimar, requieren de una mayor longitud de audio y son más fáciles de imitar (Kinnunen & Li, 2010). El tipo de palabras y los sonidos que los hablantes utilizan en su conversación pueden también ayudar a determinar su identidad. En Campbell, Campbell, Gleason, Reynolds y Shen (2007), fonemas y secuencias de fonemas se agregaron a modo de entrada a un sistema de verificación del hablante, junto con características cepstrales, con el fin de mejorar la rata de clasificación. A este grupo también pertenecen las características prosódicas, las cuales están relacionadas con el estrés sobre las silabas, patrones de entonación, rata de habla y ritmo. El parámetro prosódico más importante es la frecuencia fundamental (*Fo*). La combinación de parámetros relacionados con la *Fo* y los espectrales ha mostrado ser relevantes para el mejoramiento del desempeño de los sistemas de reconocimiento del hablante (Kinnunen & Li, 2010). En Leung, Mak, Siu y Kung (2006) se usan los patrones de pronunciación de las personas para diseñar un sistema VH, representados estos en un modelo que relaciona características articulatorias (manera y punto de articulación) con los fonemas. Aunque características de alto nivel han sido ya probadas en sistemas de VH, aún permanece abierta la pregunta acerca de exactamente cuáles características usar de la señal de voz con fines de incrementar la robustez (Kinnunen & Li, 2010); además, en Fazel y Chakrabartty (2011) se plantea el uso de parámetros de alto nivel para mejorar la robustez de estos sistemas. Por otra parte, en Univaso, Ale y Gurlekian (2015) se muestra que además de características de tono, aquellas relacionadas con la calidad y duración de las emisiones de la voz tienen también buena capacidad para tareas de discriminación de hablantes.

## Métodos de comparación forense de hablantes

Estos métodos, según Rose (2002) y Morrison (2010), pueden clasificarse en las siguientes cuatro categorías: auditivo, auditivo-espectrográfico, fonético-acústico y automático. De estos métodos, solo los dos últimos están basados en medidas objetivas de las propiedades acústicas de la señal de voz; sin embargo, los incluimos en la presente sección debido a que estos aún (año 2018) se mencionan en procedimientos realizados por entidades oficiales de algunos países: entre ellos Colombia y Perú. Adicional a los métodos mencionados anteriormente, en Univaso (2017) se agrega una nueva familia de métodos denominados semiautomáticos. En estos métodos existe una notable interacción entre el analista y la aplicación de software, buscando mezclar las ventajas de algunos métodos pertenecientes a los métodos automáticos con el conocimiento de expertos en fonética.

## Método auditivo

También conocido como auditivo-perceptual, se basa en las habilidades auditivas de profesionales del área de la

comunicación hablada, previo entrenamiento, para así realizar la identificación de personas a través de su voz. En este método las diferencias percibidas son usadas para estimar la similitud entre las voces. En general, los parámetros de voz utilizados en este método corresponden a parámetros de alto nivel. Aunque el humano está provisto de habilidades a la hora de reconocer hablantes, varios factores afectan la confiabilidad de este método (Bonastre et al., 2003): la familiaridad con el hablante, duración de la muestra, el contexto, la prosodia e imitación, y el grado de entrenamiento del experto. Por otra parte, se ha establecido que es posible encontrar voces que se escuchen igual, aunque su contenido acústico muestre diferencias notables (Rose, 2002).

Respecto a la capacidad de reconocimiento de personas por parte de humanos, en Van Lancker, Kreiman y Emmorey (1985) se reporta una tasa de reconocimiento de hablantes del 71% a partir de audios de personas famosas, y en Nielsen y Stern (1985) se obtiene un valor del 88% para el caso en el que se busca reconocer voces familiares a partir de audios sin distorsiones. Aunque se reportan éxitos respecto al uso del método auditivo-perceptual, para ello se requiere que sea cuidadosamente aplicado bajo condiciones específicas y controladas, y que los resultados sean cuidadosamente interpretados, lo cual limita su uso. Además, la alta intervención humana convierte al profesional practicante de la prueba en una posible fuente de error, lo cual limita aún más su utilidad en el campo forense (Hollien, Didla, Harnsberger & Hollien, 2016). A pesar de ello, la experiencia en este método podría llegar a ser útil para el desarrollo de nuevos métodos.

Por otra parte, con la aparición y posterior progreso de los métodos automáticos de verificación se han realizado experimentos en los que se busca comparar el desempeño del oído humano frente a las máquinas. En particular, en Schmidt-Nielsen y Crystal (2000) se obtiene que la capacidad de reconocimiento del oído humano es similar a la de los métodos automáticos de ese entonces, aunque el desempeño del humano fue más robusto en condiciones de degradación de los registros. Sin embargo, experimentos reportados en varios trabajos posteriores muestran que el desempeño de métodos automáticos recientes es superior al oído humano (Hansen & Hasan, 2015), incluso en voz imitada (Hautamäki, Kinnunen, Hautamäki & Laukkanen, 2014).

## Método auditivo-espectrográfico

Este método involucra la comparación de las voces dubitadas e indubitadas teniendo en cuenta como se escuchan los segmentos de voz, así también como se ven en el espectrograma (Rose, 2002). El examen auditivo se realiza con el propósito de buscar diferencias y similitudes entre las voces dubitada e indubitada. De manera complementaria, el examen visual busca comparar y analizar patrones acústicos en la voz a partir de los espectrogramas (Tosi, 1979). Este proceso es desarrollado por un experto entrenado para tal tarea. Como primer paso se le hace al sospechoso una sesión de recolección de audios, donde se le solicita que repita varias veces un conjunto de frases seleccionadas. Las frases se seleccionan de forma tal que coincidan tanto como sea posible con las frases pronunciadas en el audio a analizar (audio dubitado) (Tosi, 1979); sin embargo, oponentes al presente método exponen que al pedirle al sospechoso que trate de imitar aspectos prosódicos y temporales del audio dubitado, se podría caer en el riesgo de implicar a una persona inocente (Gruber & Poza, 1995; Rose, 2002).

A pesar de que la comunidad científica tiene varias reservas respecto a la confiabilidad y error estadístico de este método, que por naturaleza es subjetivo, aún se sigue referenciando su uso en países como Colombia[3] y Perú[4]. La entidad IAFPA (*International Association for Forensic Phonetics and Acoustics*) emitió una resolución en el 2007 en la cual la mencionada asociación descalifica el proceso de verificación del hablante basado en espectrogramas (http://www.iafpa.net/voicepintsres.htm) (Morrison, 2010). Similares medidas se tomaron en Francia (Add-Decker et al., 1999).

## Método acústico-fonético

Es practicado sobre todo por expertos debidamente entrenados, los cuales realizan mediciones de propiedades acústicas de la voz y luego hacen un análisis estadístico sobre estas. Como un primer paso se identifican unidades acústicas equiparables, tanto en la señal dubitada como

3 Caso número 11-001-60-00717-2011-00132; Laboratorio de Acústica Forense, Área de Policía Científica y Criminalística, Dirección de Investigación Criminal e Interpol, Policía Nacional, República de Colombia.

4 Informes periciales acústicos forenses identificados con números 0205-2017, 0207-2017, 0208-2017 y 0209-2017; año 2017, Instituto de Medicina Legal y Ciencias Forenses-Gerencia de Criminalística, Perú.

en la indubitada, y se procede a efectuar las mediciones acústicas sobre estos segmentos. Seguidamente, se buscan aquellas unidades acústicas para las cuales se cumple que la cantidad de estas es suficiente, desde el punto de vista estadístico, dentro de la muestra de voz dubitada, y luego para la voz indubitada. Ha de cumplirse que la cantidad de realizaciones (segmentos) de la unidad acústica bajo análisis sea la suficiente a fin de poder entregar resultados confiables (Rose, 2002). Para el análisis pueden utilizarse fonemas de variados tipos tales como vocales, sonidos fricativos (Cicres, 2011) y nasales (Amino & Arai, 2009), entre otros. Además, es posible utilizar segmentos de voz de tamaño superior a los de un fonema mediante el uso de propiedades acústicas consideradas como relevantes para la identificación de personas. A modo de ejemplo, se mencionan la frecuencia fundamental, los formantes (Morrison, 2009b), *jitter* y *shimmer* (Farrús & Ejarque, 2007), entre otros.

Debido a que este método está basado en mediciones acústicas, está habilitado para realizar análisis de tipo estadístico tendientes a entregar medidas de desempeño, tales como el grado de coincidencia y razón de verosimilitud entre las voces objeto de análisis; además, algunas de las medidas utilizadas en este método son robustas ante efectos de distorsión provocados por el canal y el ruido. La principal desventaja corresponde al hecho de requerir gran cantidad de horas/hombre durante el desarrollo del procedimiento. Aunque el método está basado en mediciones, aún se tiene un grado menor de subjetividad relacionado con la determinación de inicio y fin de las unidades fonéticas.

## Método combinado

En Colombia, durante el análisis comparativo de hablantes, los peritos del Cuerpo Técnico de Investigación (CTI) utilizan el método combinado[5], el cual involucra tres tipos de análisis: 1) análisis perceptual-auditivo; 2) análisis lingüístico; 3) análisis acústico. Este método combinado podría ser visto como una combinación entre los métodos auditivo, auditivo-espectrográfico y acústico-fonético, arriba mencionados. Un procedimiento que guarda similitudes es el realizado en Perú, donde se sigue una metodología que denominan "integrada", y que consiste en la aplicación de las siguientes fases[6]: (a) fase auditiva, reproducción de la señal de audio e identificación de los rasgos fono-articulatorios lingüísticos; (b) fase espectrográfica y evaluación estadística de los parámetros físicos en las muestras analizadas.

El método combinado involucra procedimientos que utilizan observaciones y mediciones que dan lugar a la subjetividad, especialmente aquellos procedimientos relacionados con el método auditivo-perceptivo y auditivo-espectrográfico, lo cual afecta la objetividad. Además, dificulta la repetibilidad y la verificación de los experimentos por parte de la defensa o la comunidad científica, en caso de ser requerida.

## Método automático

Una práctica muy extendida es usar un modelo de referencia UBM (*Universal background Model*) (Hasan & Hansen, 2011), el cual se usó por primera vez en Reynolds (1997) y Reynolds, Quatieri y Dunn (2000). El elemento UBM es, en esencia, una función de densidad de probabilidad que representa las propiedades de la voz de la población que se utiliza a modo de referencia. En tal sentido, los modelos de los registros de voz dubitados e indubitados se comparan respecto al modelo de referencia UBM. En tal caso, se tienen dos modelos: modelo del hablante $\lambda_s$, y el modelo de referencia UBM-GMM $\lambda_o$. Al pasar las observaciones correspondientes a la señal interceptada $\chi$ se obtienen dos valores de probabilidad $p(\chi|\lambda_s)$ y $p(\chi|\lambda_o)$, respectivamente; con los cuales se construye la razón de verosimilitud (LR, *Likelihood Ratio*). Pero usualmente se usa el valor logarítmico del LR,

$$\mathcal{L}(\chi) = log\ p(\chi|\lambda_s) - log\ p(\chi|\lambda_o) \qquad (1)$$

A medida que este valor $\mathcal{L}(\chi)$ aumenta, la evidencia de que los registros dubitado e indubitado corresponden se hace más fuerte. Es decir, se fortalece la hipótesis de que el indiciado es la fuente de la voz dubitada. En la tabla 1 se muestra los valores de equivalentes verbales propuestos en Rose (2002).

**Tabla 1. Equivalentes verbales de los valores logarítmicos de la razón de verosimilitud**

| Log-LR | Interpretación | |
|---|---|---|
| > 4 | Muy fuerte… | Evidencia que |
| 3 a 4 | Fuerte… | soporta que |
| 2 a 3 | Moderadamente fuerte… | ambos audios |
| 1 a 2 | Moderada… | provienen de la |
| 0 a 1 | Pobre… | misma fuente |

5 Manual de Procedimientos de Fiscalía en el Sistema Penal Acusatorio Colombiano, Fiscalía General de la Nación, https://www.fiscalia.gov.co/colombia/wp-content/uploads/2012/03/spoa.pdf.

6 Informes periciales acústicos forenses identificados con números 0205-2017, 0207-2017, 0208-2017 y 0209-2017; año 2017, Instituto de Medicina Legal y Ciencias Forenses-Gerencia de Criminalística, Perú.

Aunque el desarrollo del método GMM-UBM fue un evento muy importante, se han creado nuevos métodos que mejoran aún más el desempeño de la tarea VH. Después de los GMM se desarrolló el método de súper-vectores, que corresponde a un vector de valores obtenido a partir de concatenar los parámetros de los modelos GMM (Kenny, Mihoubi & Dumouchel, 2003), por ejemplo, sus valores esperados. Aunque estos súper-vectores también pueden obtenerse sobre otros modelos tales como redes neuronales autoasociativas (Garimella & Hermansky, 2013). Los súper-vectores GMM se utilizaron en Campbell, Sturim y Reynolds (2006) a modo de características para un clasificador basado en SVM (*support vector machines*). Con esta estrategia se combina la habilidad del modelado de los GMM con la habilidad para la clasificación de las SVM. El uso de súper-vectores ayudó al desarrollo de técnicas que mejoraron la robustez, por ejemplo el caso de proyección de atributos no convenientes (NAP, *nuisance attribute projection*) (Castaldo, Colibro, Dalmasso, Laface & Vair, 2007) y análisis conjunto de factores (JFA, *join factor analysis*) (Dehak, Kenny, Dehak, Dumouchel & Ouellet, 2011; Kenny, Ouellet, Dehak, Gupta & Dumouchel, 2008). Sin embargo, estas técnicas también son aplicables sobre los parámetros acústicos, como se muestra en Hasan y Hansen (2013).

Posteriormente, aparecen los *i-vectors*, los cuales en esencia corresponden a versiones reducidas en dimensión de los súper-vectores. Un súper-vector GMM $m_h$ del hablante *h* puede representarse mediante (Dehak et al., 2011)

$$m_h = m + T \cdot w_h \qquad (2)$$

donde $m$ corresponde al súper-vector GMM-UBM que representa aquella componente independiente del hablante y del canal; y *T*, es una matrix de rango reducido, llamada matrix de variabilidad total, que contiene aquellas direcciones más relevantes obtenidas a partir de una cantidad suficiente de datos (Kanagasundaram, 2014). Por otra parte, $w_h \mathcal{N}(0, I)$ corresponde a un vector de variables ocultas estimadas, llamadas *i-vectors*, con la capacidad de encapsular la información de los registros de audio en unos pocos valores, lo que permite que posteriormente sean aplicados métodos de compensación del canal (Hansen & Hasan, 2015). Los *i-vectors* son utilizados en una buena cantidad de sistemas de verificación del hablante del estado del arte. El desempeño de estos, probados siguiendo los lineamientos propuestos por los protocolos de evaluación NIST SRE 2012[7], se muestra en Saedi et al. (2013), en donde se reportan valores de hasta 4% en el valor EER. Finalmente, con el advenimiento del aprendizaje profundo las mejoras en desempeño podrían ser aún mayores (Li, Chen, Shi, Tang & Wang, 2017; Snyder, García-Romero, Povey & Khudanpur, 2017).

## Propiedades de los registros de señales de voz y su influencia en el desempeño de los sistemas de verificación de hablantes

Se ha mostrado que el desempeño de los procedimientos de comparación forense de voces se ve afectado por propiedades tales como codificación, longitud de los audios, contenido de ruido, presencia de artefactos en la señal, entre otros. Por otra parte, este efecto es diferente dependiendo del método, ya sea el método automático o el método acústico-fonético. Para el caso del método acústico-fonético se muestra la influencia de las propiedades del audio sobre parámetros acústicos comúnmente utilizados en estos procedimientos: frecuencia fundamental y formantes.

## Codificación

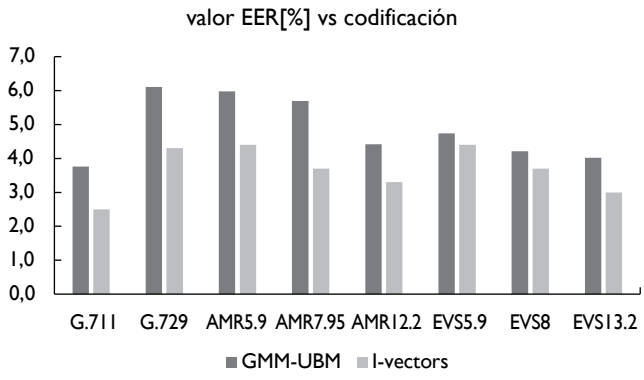
La codificación corresponde al proceso de encontrar una representación que permita transmitir o almacenar la voz de tal manera que se pueda transmitir más eficientemente por los diferentes canales (e.g., canales inalámbricos). La gran mayoría de las redes de comunicación son de tipo digital y todas las señales enviadas a través de la misma red son codificadas en paquetes de bits a variadas ratas de bits, que dependen, entre otras cosas, del tipo de codificación; y dado que buena parte de los tipos de codificación son del tipo con pérdida, es inevitable que se introduzcan distorsiones (Moreno-Daniel, 2004).

En Polacky, Jarina y Chmulik (2016) se reporta un sistema del tipo automático en el que el EER (*equal error rate*) se reduce (equivale a decir que su desempeño aumenta) al incrementar la rata de bits. Además, en Polacky *et al.* (2016) se menciona que, dentro de los sistemas de codificación, el G.711 es uno de los que ofrece menos degradación en el desempeño en sistemas de verificación del hablante basados en métodos automáticos del tipo

7 https://www.nist.gov/multimodal-information-group/speaker-recognition-evaluation-2012.

GMM-UBM. En contraste, el formato AMR5.9 resulta ser de los menos convenientes al utilizar GMM-UBM. Este resultado se confirma con lo hallado en Jarina, Polacký, Počta y Chmulík (2017), donde además se muestra que la técnica de *i-vectors* es más robusta ante cambios en la codificación respecto a GMM. Parte de los resultados reportados en estos trabajos se transcriben en la figura 1, donde puede observarse el desempeño medido en EER para diferentes tipos de codificación. Se observa que el mejor desempeño es para la codificación G.711, seguido por el EVS13.2. En contraste, el peor desempeño es para los tipos de codificación AMR5.9 y G.729.

**Figura 1.** Valor EER [%] aproximado para el caso en el que se tiene la señal de voz dubitada en diferentes estándares de codificación si se utilizan los métodos de GMM-UBM e i-vectors




Fuente: los datos de la presente figura se toman de los trabajos Polacky et al. (2016) y Jarina et al. (2017), quienes usan el conjunto de datos TIMIT.

El estándar G.711, operando a 64 Kb/s, ha sido comúnmente utilizado en telefonía fija; y para el caso de telefonía móvil, versiones de ancho de banda angosta son la selección por defecto; por ejemplo, el AMR-NB, que operan usualmente a tasas de 12 Kb/s. Un estándar más reciente es el EVS, el cual no tiene el problema de tener un reducido ancho de banda, puede representar señales acústicas diferentes a la voz a diferencia del AMR, y opera con ratas de bits similares a los de telefonía celular. El estándar EVS se está difundiendo ampliamente en comunicaciones sobre redes LTE (la sucesora indiscutible de la tecnología 3G) (Jarina *et al.*, 2017). Por tanto, se espera que a futuro buena parte de las comunicaciones de voz estén codificadas en EVS.

Por otra parte, los esquemas de codificación pueden ser del tipo de banda angosta (NB, *narrow band*) y de banda ancha (WB, *wide band*). En los de banda angosta el muestreo se realiza a 8.000 Hz y cubre el rango de frecuencias de la voz de entre los 300 Hz y 3.400 Hz del espectro de voz, aproximadamente; y en los de banda ancha se cubre el rango de frecuencias entre los 50 Hz y 7.000 Hz, con una frecuencia de muestreo de 16.000 Hz. En Jarina et al. (2017) se muestra que con la codificación del tipo WB se obtienen desempeños superiores que con los del tipo NB (de 1% a 3% en términos absolutos en un sistema del tipo automático), lo cual sugiere que existe información por arriba de los 3.400 Hz de la señal de voz que resulta ser útil para propósitos de reconocimiento de personas.

Referente al método acústico-fonético de verificación de hablantes, la codificación también afecta la estimación de los parámetros en sí misma. En comunicaciones móviles, la codificación más ampliamente utilizada es la codificación multitasa adaptiva (AMR, *adaptive multi-rate*). Para esta codificación en particular se tiene la opción de banda ancha y de banda angosta (AMR-NB) que codifica el rango de frecuencias 200-3.400 Hz a tasas variables, y la opción de banda ancha (AMB-WB) que codifica la banda 50-6.400 Hz. La codificación AMR-NB genera archivos de muy bajo peso, pero de baja calidad. En Ireland, Knuepffer y Mcbride (2015) se estima la influencia de la codificación AMR sobre parámetros acústicos medidos en vocales, en el cual se reporta que la influencia sobre los parámetros $F_0$ (frecuencia fundamental) y HNR (relación armónico-ruido) es despreciable tanto para la codificación AMR-NB como para AMR-WB. En contraste, la distorsión es alta para el caso de los parámetros acústicos *jitter* y *shimmer*. Para el caso de los formantes y los coeficientes MFCC, la distorsión es considerable para la codificación de banda angosta AMR-NB, pero muy poca para la de banda ancha AMR-WB. Según el mismo trabajo, para todos los casos se observa una tendencia en la cual la distorsión aumenta a medida que se decrementa la rata de bits por segundo en la codificación. En la tabla 2 se muestran la diferencia en la medición en porcentaje entre las señales de voz sin codificación y señales con codificación en AMR de banda angosta para diferentes valores de rata de bits.

**Tabla 2.** Diferencia en la medición en porcentaje entre las señales de voz sin codificación y señales con codificación en AMR de banda angosta para diferentes valores de rata de bits. Si el valor es positivo indica que hay sobreestimación. De lo contrario, si es negativo indica que se tiene una subestimación.

| codificación | sexo | Fo | jitter | shimmer | HNR | F1 | F2 |
|---|---|---|---|---|---|---|---|
| AMR_5.9 | M | 0 | **-23** | **-42** | **2** | 13 | **17** |
| | F | 0 | **-26** | **-56** | 4 | **31** | **29** |
| AMR_7.95 | M | 0 | -16 | **-27** | **2** | 13 | **18** |
| | F | 0 | -18 | **-38** | 4 | **32** | **28** |
| AMR_12.2 | M | 0 | -18 | **-18** | **2** | 11 | **16** |
| | F | 0 | -27 | **-27** | 4 | **29** | **28** |

Fuente: estos valores se tomaron del trabajo Ireland et al. (2015). El valor en negrilla indica que este corresponde a una diferencia estadísticamente significativa.

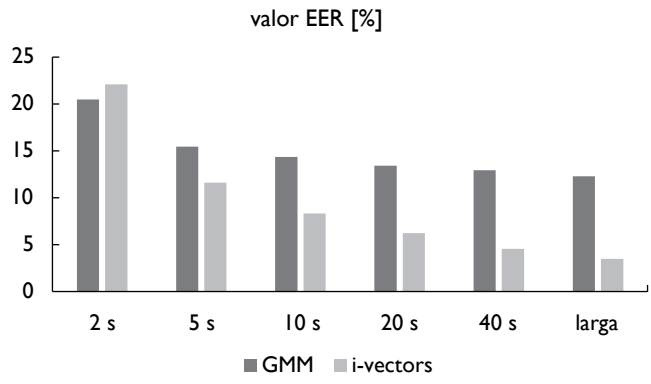
## Longitud de audio

Longitudes de audio relativamente cortas afectan en mayor manera al método acústico-fonético debido a que debe contar con una cantidad estadísticamente suficiente de realizaciones de aquellos fonemas a analizar, las cuales no siempre ocurren de manera frecuente en la señal de voz. Además, el grado de afectación dependería de qué tan común son esos rasgos que se utilizan para el proceso de comparación forense. Al respecto, en Poddar, Sahidullah y Saha (2018) se muestra que la cantidad promedio de fonemas encontrados una única vez dentro de un registro de audio se reduce de manera exponencial con la longitud de este mismo audio. En cuanto a la frecuencia de ocurrencia de los fonemas, en Hasan et al. (2013) se muestra el histograma esperado de ocurrencia de fonemas del idioma inglés para varias longitudes de audio. Para el caso del castellano colombiano, en González y Mejía (2011) se muestra cuáles son aquellos fonemas más frecuentes.

Por otra parte, es importante tener en cuenta que en el método acústico-fonético comúnmente se hace uso de medidas de tiempo largo, tales como LTFo (*long term fundamental frecuency*) y LTAS (*long term average spectrum*), las cuales requieren de un tiempo de análisis mínimo para que estas medidas para ese hablante en particular sean confiables. En Arantes y Eriksson (2014) se estima cuánto tiempo se requiere a fin de obtener una medición estable del valor promedio y la mediana de la frecuencia fundamental, para 26 idiomas. Se encuentra que estas medidas se estabilizan en, a lo mucho, 30 segundos, y que el umbral a partir del cual estas medidas se inician a estabilizar es de alrededor de 10 segundos. Es decir, para obtener medidas confiables del promedio de la frecuencia fundamental, las longitudes de audio han de ser de mínimo 10 segundos, con un valor recomendado 20% superior, es decir, 12 segundos.

En cuanto a métodos del tipo automático, se ha encontrado que el desempeño en EER se reduce notablemente cuando la longitud del audio se reduce. En la figura 2 se muestra el gráfico de desempeño frente a la longitud de audio dubitado (2, 5, 10, 20 y 40 segundos). Para el caso del entrenamiento del modelo de hablante a cuestionar se utilizan varios minutos de audio, lo mismo que para el caso del valor de etiquetado como "larga". Se puede observar para el caso de la técnica de *i-vectors*, que se pasa de tener un EER de 22,1% a 4,6% al pasar de contar con 2 segundos a 40 segundos de longitud en el registro de audio. De contar con varios minutos tanto para el audio dubitado como indubitado el valor EER estimado es de 3,5%.

**Figura 2.** Desempeño en EER para diferentes valores de longitud de señal de audio dubitada, si se utilizan los métodos de GMM-UBM e i-vectors




Fuente: se construyó la figura a partir de los datos reportados en Poddar, Sahidullah y Saha (2015), en el cual se utilizaron datos del corpus NIST 2008 para entrenamiento y validación de hablantes.

Aunque en los estudios previos se analiza la longitud de los audios, hay que tener en cuenta que en las interceptaciones telefónicas los audios podrían incluir segmentos de silencio que no proveen información útil para el proceso. Al medir la longitud efectiva del audio se descartan aquellas zonas de silencio de información irrelevante. Al contar con audios de mayor tamaño en su longitud efectiva se agrega mayor información al proceso de verificación de hablantes, y, por tanto, el desempeño de estos sistemas mejora (Hautamäki, Cheng, Rajan & Lee, 2013; Sarkar, Driss, Bousquet & Bonastre, 2012). Sin embargo, la relación entre longitud efectiva de audio y desempeño en EER es un tema aún por investigar.

## Relación señal-ruido

El nivel de contenido de ruido se mide mediante la relación señal-ruido (SNR, *signal-to-noise ratio*). La SNR es una medida que compara la potencia de la señal de interés con la del ruido de fondo y se expresa en $dB$. Un valor SNR mayor que 0 $dB$ indica que hay más señal que ruido; y si valiese 0 dB el nivel de ruido sería igual al nivel de la señal. El ruido de fondo suele presentarse en las grabaciones y afecta la comparación directa entre las señales dubitada e indubitada. A modo de ejemplo, en el ámbito forense de Chile se ha comprobado que las condiciones de canal y ruido tienen un alto impacto en el análisis de casos (Rosas & Sommerhoff, 2009).

Para el caso del método acústico-fonético, es importante tener en cuenta que el ruido afecta las estimaciones de parámetros acústicos de voz en mayor o menor grado,

dependiendo del parámetro acústico, del método usado y del nivel de ruido (SNR).

Nakatani y Irino (2004) muestran el desempeño del método de estimación de la $F_O$ que ellos desarrollan (DASH, *dominance spectrum based harmonics extraction*) respecto a otros métodos, tales como el basado en *cepstrum* y el método *YIN* reportado en Cheveigné y Kawahara (2002). La comparación se realiza en presencia tanto de ruido blanco gaussiano como de ruido tipo *babble*. Según los datos reportados, el método basado en *cepstrum* es el de menor desempeño. Para este método, si el nivel de SNR fuese de 5 dB, se esperarían errores de estimación superiores al ± 5% para el 22% de los valores correspondientes a las ventanas de análisis, en caso de que el ruido fuese del tipo blanco gaussiano. Por otra parte, si el ruido fuese de tipo *babble*, para ese mismo nivel de SNR, se esperaría que 48% de las estimaciones tuviesen errores de estimación superiores al 5%. En la gráfica 4 del trabajo presentado por Nakatani y Irino (2004) se muestra el desempeño de la estimación de la frecuencia fundamental utilizando varios métodos respecto al nivel de ruido, tanto para ruido blanco como para ruido tipo *babble*.

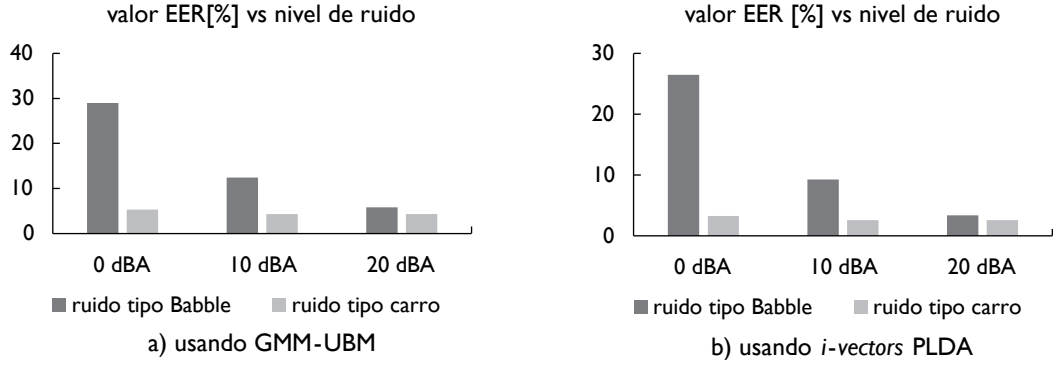
De manera similar, el contenido de ruido afecta la estimación de los formantes, tal como se muestra en Jameel, Fattah, Goswami, Zhu y Ahmad (2017) para el caso de ruido blanco gaussiano. En Jameel et al. (2017) puede observarse que para un nivel de ruido de aproximadamente 5 dB se espera encontrar errores de estimación del 16% a 18% dependiendo del método que se usó para la estimación de los formantes. Dentro de los métodos analizados están LPC (*linear predictive coding*), *WaveSurfer* y AFB (*adaptive filter bank* [Mustafá & Bruce, 2006]). Las gráficas 7 y 8 del trabajo presentado en Jameel *et al.* (2017) muestran los porcentajes de error promedio para diferentes niveles de ruido y tipos de ruido.

Referente a los métodos automáticos, varios estudios han demostrado que el ruido afecta el desempeño de sistemas de verificación de hablantes; sin embargo, estos efectos pueden ser disminuidos aplicando métodos desde el punto de las características de representación de la voz y desde el punto del modelo (Li & Mak, 2015). El primero busca encontrar características que sean más robustas que las MFCC convencionales utilizadas en métodos automáticos, mientras que el segundo busca entrenar los modelos de manera que estos sean más resistentes al ruido. En Mandasari, McLaren y Van Leeuwen (2012) se evalúa el desempeño ante condiciones de ruido de dos métodos ampliamente conocidos: GMM-UBM e *i-vectors* con PLDA. En el mencionado trabajo se muestra que, en presencia de ruido de automóvil, la caída relativa de EER de los sistemas de *i-vectors* fue de entre 10%-20% por cada degradación del ruido de magnitud de 5 dB. Para el caso del ruido tipo *coktail*, la caída reportada en desempeño fue de entre 40%-60%. En general, los resultados muestran que el método de *i-vectors* tiene mejor desempeño que el tradicional GMM-UBM. En la figura 3 se muestran algunos valores de EER para algunas configuraciones de nivel de ruido y tipo de ruido.

## Saturación del rango dinámico

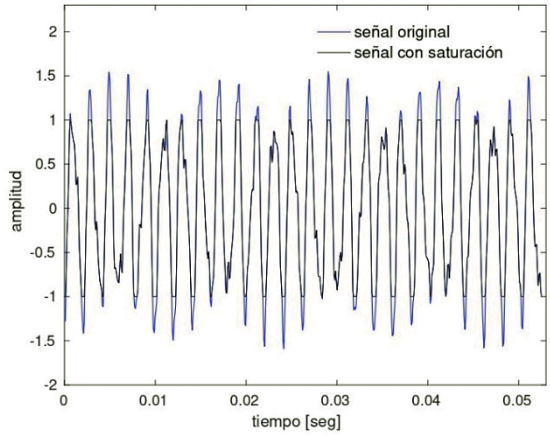
El efecto de saturación del rango dinámico ocurre cuando la señal de entrada a un dispositivo de grabación excede el rango dinámico disponible del dispositivo (Bie, Wang, Wang & Zheng, 2015; Eaton & Naylor, 2013), lo que provoca un achatamiento de la señal en los límites del rango dinámico.

**Figura 3.** Valor EER [%] para diferentes niveles de ruido para los tipos de ruido bubble y tipo carro. En a) cuando se utiliza el método GMM-UBM, y en b) cuando se utiliza la técnica de i-*vectors* con PLDA

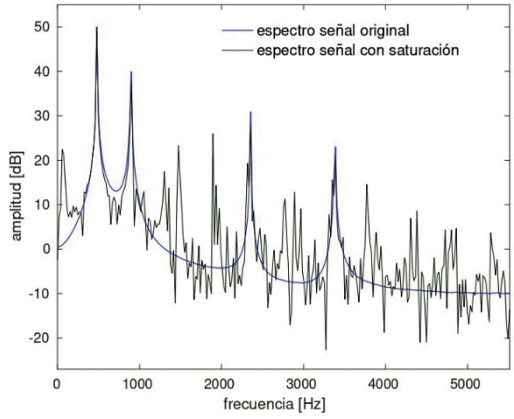



Fuente: los valores utilizados para la presente figura se tomaron de Mandasari et al. (2012).

**Figura 4.** En a), señal en el tiempo con el efecto de saturación; y en b) el efecto de la saturación en el dominio de la frecuencia. En este caso la saturación es del 20%




a) señal en el tiempo




b) espectro de la señal

Un ejemplo del efecto del *clipping* sobre la señal de audio se observa en la figura 4, en la cual puede apreciarse la aparición de componentes adicionales de frecuencia que podrían afectar los algoritmos de estimación de los formantes. En contraste, en Bie et al. (2015) se muestra que el efecto de *clipping* es relativamente bajo en sistemas de reconocimiento de hablantes basados en *i-vectors*, lo cual, según los mismos autores, podría deberse al aumento en robustez entregado por la técnica de *i-vectors*.

## Presencia de transitorios

Este tipo de perturbación, al que se le denomina comúnmente *clicks*, corresponde a un tipo particular de ruido impulsivo que degrada una pequeña porción de la señal de audio y cuyo tiempo de duración es de alrededor de 1 ms (Ávila y Biscainho, 2012). Este fenómeno se manifiesta como cambio abrupto de corta duración en el espectrograma con presencia de energías en un rango amplio de frecuencias (Manikandan, Yadav & Ghosh, 2017; Nongpiur, 2008; Wan, Ma & Li, 2018). En el estado del arte, hasta donde nuestro conocimiento llega, no se tienen reportes en los que se evalúe el efecto de estas afectaciones en sistemas de verificación de hablantes.

## Conclusiones

Se analizó la influencia de las propiedades de los audios sobre los métodos de verificación de hablantes que utilizan parámetros de voz medibles de manera objetiva; es decir, los métodos de verificación de hablantes del tipo acústico-fonético y automático. El utilizar medidas y procedimientos objetivos permite que los experimentos y resultados asociados a diligencias judiciales sean repetibles. Se encontró que las propiedades de los audios analizadas afectan las mediciones de parámetros acústicos, al tiempo que también afectan el desempeño de sistemas del tipo automático. Respecto a los otros métodos (auditivo y auditivo-espectrográfico), se encontró que varios estudios previos recomiendan no utilizar estos dos métodos.

Según la revisión presentada, el método automático funciona mejor que el método acústico-fonético, para el caso en él se tienen audios cuya longitud efectiva es más corta. En cuanto a la influencia de la saturación del rango dinámico y de transitorios existe poca información reportada, lo cual dificulta establecer la influencia de las mismas; sin embargo, estos son fenómenos que suelen presentarse sobre segmentos cortos del audio; por tanto, estos segmentos se podrían descartar sin afectar demasiado la confiabilidad de los resultados totales.

Analizar todas las influencias posibles sobre la señal de voz que repercuten en los diferentes métodos de verificación de hablantes en el campo forense está más allá del alcance del presente trabajo. Dentro de estas influencias se podrían analizar: la reverberación, la edad, las afectaciones de salud, los estados emocionales, entre otros.

## Referencias

Add-Decker, M. *et al.* (1999). *Pétition pour l'arrêt des expertises vocales, tant qu'elles n'auront pas été validées scientifiquement: Pétition du GFCP de la SFA.* Association Francophone de la Communication Parl'ee. Recuperado de http://www.afcp-parole.org/doc/petition.pdf.

Amino, K., & Arai, T. (2009). Speaker-dependent characteristics of the nasals. *Forensic Science International*, *185*(1), 21-28. http://dx.doi.org/10.1016/j.forsciint.2008.11.018

Arantes, P., & Eriksson, A. (2014). *Temporal stability of long-term measures of fundamental frequency*. http://dx.doi.org/10.13140/2.1.4619.0089

Ávila, F. R., & Biscainho, L. W. P. (2012). Bayesian restoration of audio signals degraded by impulsive noise modeled as individual pulses. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(9), 2470-2481. http://dx.doi.org/10.1109/TASL.2012.2203811

Barinov, A. (2010). Voice samples recording and speech quality assessment for forensic and automatic speaker identification. En *Audio engineering society, convention paper* (vol. 129th Convention, pp. 366-373). https://speechpro.com/files/en/media/publications/voice_samples_recording_for_forensic_speaker_identification.pdf.

Bie, F., Wang, D., Wang, J., & Zheng, T. F. (2015). Detection and reconstruction of clipped speech for speaker recognition. *Speech Communication*, *72*, 218-231. http://dx.doi.org/10.1016/j.specom.2015.06.008

Bonastre, J.-F., Bimbot, F., Böe, L.-J., Campbell, J. P., Reynolds, D. A., & Magrin-Chagnolleau, I. (2003). Person authentication by voice: A need for caution. En *Interspeech.* ISCA. https://www.isca-speech.org/archive/eurospeech_2003/e03_0033.html.

Campbell, W. M., Campbell, J. P., Gleason, T. P., Reynolds, D. A., & Shen, W. (2007). Speaker verification using support vector machines and high-level features. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 2085-2094. http://dx.doi.org/10.1109/TASL.2007.902874

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308-311. http://dx.doi.org/10.1109/LSP.2006.870086

Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., & Vair, C. (2007). Compensation of nuisance factors for speaker and language recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 1969-1978. http://dx.doi.org/10.1109/TASL.2007.901823

Cheveigné, A., & Kawahara, H. (2002). YIN, A fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917-1930. http://dx.doi.org/10.1121/1.1458024

Cicres, J. (2011). Los sonidos fricativos sordos y sus implicaciones forenses. *Estudios Filológicos*, *48*, 33-48.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788-798. http://dx.doi.org/10.1109/TASL.2010.2064307

Eaton, J., & Naylor, P. A. (2013). Detection of clipping in coded speech signals. En *21st european signal processing conference (eusipco 2013)* (pp. 1-5). https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6811469.

Farrús, M., Hernando, J., & Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. En *8th annual conference of the international speech communication association, interspeech* (pp. 778-781). Antwerp (Belgium). http://dx.doi.org/10.15332/iteckne.v14i2.1767

Fazel, A., & Chakrabartty, S. (2011). An overview of statistical pattern recognition techniques for speaker verification. *IEEE Circuits and Systems Magazine*, *11*(2), 62-81. http://dx.doi.org/10.1109/MCAS.2011.941080

Garimella, S., & Hermansky, H. (2013). Factor analysis of auto-associative neural networks with application in speaker verification. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(4), 522-528. http://dx.doi.org/10.1109/TNNLS.2012.2236652

González-Rátiva, M. C., & Mejía-Escobar, J. A. (2011). Frecuencia fonemática del español de Colombia. *Forma y Función, 24*(2), 69-102.

González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., & Ortega-García, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech & Language*, *20*(2), 331-355. (Odyssey 2004: The speaker and Language Recognition Workshop). http://dx.doi.org/10.1016/j.csl.2005.08.005

Gruber, J., & Poza, F. (1995). *Voicegram identification evidence*. Lawyers Cooperative Pub.

Hansen, J. H. L., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, *32*(6), 74-99. http://dx.doi.org/10.1109/MSP.2015.2462851

Hasan, T., & Hansen, J. H. L. (2011). A study on universal background model training in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 1890-1899. http://dx.doi.org/10.1109/TASL.2010.2102753

Hasan, T., & Hansen, J. H. L. (2013). Acoustic factor analysis for robust speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(4), 842-853. http://dx.doi.org/10.1109/TASL.2012.2226161

Hasan, T., Saeidi, R., Hansen, J. H. L., & van Leeuwen, D. A. (2013). Duration mismatch compensation for i-vector based speaker recognition systems. *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, pp. 7663-7667. http://dx.doi.org/10.1109/ICASSP.2013.6639154

Hautamäki, R. G., Kinnunen, T., Hautamäki, V., & Laukkanen, A.-M. (2014). *Comparison of human listeners and speaker verification systems using voice mimicry data*. http://cs.uef.fi/~villeh/mimicry_odyssey2014.pdf.

Hautamäki, V., Cheng, Y.-C., Rajan, P., & Lee, C.-H. (2013). Minimax i-vector extractor for short duration speaker verification. *Proceedings of the Annual Conference of the International Speech Communication Association*, *Interspeech*, 3708-3712.

Hollien, H., Didla, G., Harnsberger, J. D., & Hollien, K. A. (2016). The case for aural perceptual speaker identification. *Forensic Science International*. http://dx.doi.org/10.1016/j.forsciint.2016.08.007

Ireland, D., Knuepffer, C., & Mcbride, S. (2015). Adaptive multirate compression effects on vowel analysis. *Frontiers in bioengineering and biotechnology*, *3*, 118. http://dx.doi.org/10.3389/fbioe.2015.00118

Jameel, A. S. M. M., Fattah, S. A., Goswami, R., Zhu, W. P., & Ahmad, M. O. (2017). Noise robust formant frequency estimation method based on spectral model of repeated autocorrelation of speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(6), 1357-1370. http://dx.doi.org/10.1109/TASLP.2016.2625423

Jarina, R., Polacký, J., Počta, P., & Chmulík, M. (2017). Automatic speaker verification on narrowband and wideband lossy coded clean speech. *IET Biometrics*, *6*(4), 276-281. http://dx.doi.org/10.1049/iet-bmt.2016.0119

Kanagasundaram, A. (2014). *Speaker verification using i-vector features* (tesis doctoral). Speech and Audio Research Laboratory, Queensland University of Technology. https://eprints.qut.edu.au/77834/1/Ahilan_Kanagasundaram_Thesis.pdf.

Kenny, P., Mihoubi, M., & Dumouchel, P. (2003). New map estimators for speaker recognition. En *Interspeech.* https://www.crim.ca/perso/patrick.kenny/eurospeech2003.pdf.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., & Dumouchel, P. (2008). A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(5), 980-988. http://dx.doi.org/10.1109/TASL.2008.925147

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to super-vectors. *Speech Communication*, *52*(1), 12-40. http://dx.doi.org/10.1016/j.specom.2009.08.009

Leung, K., Mak, M., Siu, M., & Kung, S. (2006). Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification. *Speech Communication*, *48*(1), 71-84. http://dx.doi.org/10.1016/j.specom.2005.05.013

Li, L., Chen, Y., Shi, Y., Tang, Z., & Wang, D. (2017). Deep speaker feature learning for text-independent speaker verification. En *Interspeech* (pp. 1542-1546). http://dx.doi.org/10.21437/Interspeech.2017-452

Li, N., & Mak, M.-W. (2015). Snr-invariant PLDA modeling for robust speaker verification. En *Interspeech.* https://www.isca-speech.org/archive/interspeech_2015/papers/i15_2317.pdf.

Magrin-Chagnolleau, I., Durou, G., & Bimbot, F. (2002). Application of time-frequency principal component analysis to text-independent speaker identification. *IEEE Transactions on Speech and Audio Processing*, *10*(6), 371-378. http://dx.doi.org/10.1109/TSA.2002.800557

Mandasari, M. I., McLaren, M., & van Leeuwen, D. A. (2012). The effect of noise on modern auto-matic speaker recognition systems. En *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4249-4252). http://dx.doi.org/10.1109/ICASSP.2012.6288857

Manikandan, M. S., Yadav, A. K., & Ghosh, D. (2017). Elimination of impulsive disturbances from archive audio signals using sparse representation in mixed dictionaries. En *Tencon ieee region 10 conference* (pp. 2531-2535). http://dx.doi.org/10.1109/TENCON.2017.8228288

Moreno-Daniel, A. (2004). Speaker verification using coded speech. En A. Martínez Trinidad (Ed.), *Progress in pattern recognition, image analysis and applications. CIARP* (vol. 3287, pp. 366-373). http://dx.doi.org/10.1007/978-3-540-30463-0_45

Morrison, G. S. (2009a). *Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /ai/*. http://dx.doi.org/10.1558/ijsll.v15i2.249

Morrison, G. S. (2009b). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. *The Journal of the Acoustical Society of America*, *125*(4), 2387-2397. http://dx.doi.org/10.1121/1.3081384

Morrison, G. S. (2010). Expert evidence. En (cap. Forensic voice comparison). Thomson Reuters. http://expert-evidence.forensic-voice-comparison.net/.

Mustafá, K., & Bruce, I. C. (2006). Robust formant tracking for continuous speech with speaker variability. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 435-444. http://dx.doi.org/10.1109/TSA.2005.855840

Nakatani, T., & Irino, T. (2004). Robust and accurate fundamental frequency estimation based on dominant harmonic components. *Journal of the Acoustical Society of America*, *116*(6), 3690-3700. http://dx.doi.org/10.1121/1.1787522

Nielsen, A. S., & Stern, K. R. (1985). Identification of known voices as a function of familiarity and narrow band coding. *Journal of the Acoustical Society of America*, *77*, 658. http://dx.doi.org/10.1121/1.391884

Nongpiur, R. C. (2008). Impulse noise removal in speech using wavelets. En *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1593-1596). http://dx.doi.org/10.1109/ICASSP.2008.4517929

Poddar, A., Sahidullah, M., & Saha, G. (2018). Speaker verification with short utterances: A review of challenges, trends and opportunities. *IET Biometrics*, *7*(2), 91-101. http://dx.doi.org/10.1049/iet-bmt.2017.0065

Poddar, A., Sahidullah, M., & Saha, G. (2015). Performance comparison of speaker recognition systems in presence of duration variability. *Annual IEEE India Conference (INDICON) (*pp. 1-6*).* New Delhi. http://dx.doi.org/10.1109/INDICON.2015.7443464

Polacky, J., Jarina, R., & Chmulik, M. (2016). Assessment of automatic speaker verification on lossy transcoded speech. En *4th International Conference on Biometrics and Forensics (IWBF)* (pp. 1-6). http://dx.doi.org/10.1109/IWBF.2016.7449679

Reynolds, D.A. (1997). Comparison of background normalization methods for text-independent speaker verification. En *Proc. of 5th European Conf. on Speech Communication and Technology (EuroSpeech)* (vol. 2, pp. 963-966). https://pdfs.semanticscholar.org/f5ad/e2e149b2d4bc0a4c679207b2bf858692af7a.pdf.

Reynolds, D.A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, *10*(1), 19-41.

Romito, L., & Galatà, V. (2004). Towards a protocol in speaker recognition analysis. *Forensic Science International*, *146*, S107-11. http://dx.doi.org/10.1006/dspr.1999.0361

Rosas, C., & Sommerhoff, J. (2009). Efectos acústicos de las variaciones fonopragmáticas y ambientales. *Estudios Filológicos*, *44*, 195-210. http://dx.doi.org/10.4067/S0071-17132009000100012

Rose, P. (2002). *Forensic speaker identification* (F. S. Series, Ed.). Taylor & Francis.

Saedi, R. *et al.* (2013). I4U submission to NIST-SRE 2012: A large-scale collaborative effort for noise-robust speaker verification. En *Interspeech.* https://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_1986.pdf.

Sarkar, A., Driss, M., Bousquet, P.-M., & Bonastre, J.-F. (2012). Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. *Proceedings of the Annual Conference of the International Speech Communication Association. Interspeech*.

Schmidt-Nielsen, A., & Crystal, T. H. (2000). Speaker verification by human listeners: Experiments comparing human and machine performance using the nist 1998 speaker evaluation data. *Digital Signal Processing*, *10*(1), 249-266. http://dx.doi.org/10.1006/dspr.1999.0356

Snyder, D., García-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. En *Interspeech* (pp. 999-1003). http://dx.doi.org/10.21437/Interspeech.2017-620

Tosi, O. (1979). *Voice identification: Theory and legal applications*. University Park Press: Baltimore, Maryland.

Univaso, P., Ale, J. M., & Gurlekian, J. A. (2015). Data mining applied to forensic speaker identification. En *IEEE Latin America Transactions*, *13*(4), 1098-1111. http://dx.doi.org/10.21437/Interspeech.2017-620

Univaso, P. (2017). Forensic speaker identification: A tutorial. En *IEEE Latin America Transactions*, *15*(9), pp. 1754-1770. http://dx.doi.org/10.21437/Interspeech.2017-620

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I. Recognition of backward voices. *Journal of Phonetics*, *13*, 19-38.

Wan, H., Ma, X., & Li, X. (2018). Variational bayesian learning for removal of sparse impulsive noise from speech signals. *Digital Signal Processing*, *73*, 106-116. http://dx.doi.org/10.1016/j.dsp.2017.11.007

Zheng, N., Lee, T., & Ching, P. C. (2007). Integration of complementary acoustic features for speaker recognition. *IEEE Signal Processing Letters*, *14*(3), 181-184. http://dx.doi.org/10.1016/j.dsp.2017.11.007