Título: **Métodos de representación y verificación del locutor texto-independiente.**

Unidad Ejecutora Principal del Resultado:
División de Investigaciones, CENATAV, DATYS

Autores:

(i) Autor Principal:
Dr.C. Gabriel Hernández Sierra (CENATAV, DATYS, Cuba)
(ii) Otros Autores:
Dr.C. José R. Calvo de Lara (CENATAV, DATYS, Cuba)
Dr. Jean-François Bonastre (Universidad de Avignon, Francia)

Colaboradores:
Dr.C. Pierre-Michel Bousquet (Universidad de Avignon, Francia)
Dr.C. Driss Matrouf (Universidad de Avignon, Francia)
Dr.C. Xavier Anguera (Telefonica Research, Barcelona, España)
Ing. Flavio J. Reyes (CENATAV, DATYS, Cuba)
Lic. Manuel Aguado Martínez) (CENATAV, DATYS, Cuba)
Lic. Rafael Fernández (CENATAV, DATYS, Cuba)
MSc. Argel González Padilla (ISPJAE, Cuba)
Ing. Dayana Ribas González (CENATAV, DATYS, Cuba)
Lic. Yannia Castellanos (Seguridad Tecnológica, DATYS, Cuba)
Ing. Orlando Jiménez (Seguridad Tecnológica, DATYS, Cuba)
Ing. Adamis Rodriguez (Seguridad Tecnológica, DATYS, Cuba)
Lic. Ana Montalvo (CENATAV, DATYS, Cuba)
Ing. Humberto Sierra (Seguridad Tecnológica, DATYS, Cuba)
Ing. Claudia Bello (CENATAV, DATYS, Cuba)

Resumen:
En concordancia con la estrategia trazada por la dirección del Ministerio del Interior y la empresa DATYS, con el objetivo de contar con herramientas computacionales propias para aumentar la eficacia del enfrentamiento de actividades delictivas y a su vez la obtención de productos competitivos a nivel mundial que contengan conocimiento propio, se desarrolló esta investigación. El reconocimiento del locutor independiente del texto, es un método de reciente incorporación en los sistemas biométricos y su auge se refleja en las competencias internacionales. A su vez es una necesidad identificada por los diferentes órganos operativos del MININT, el contar con herramientas que permitan la identificación de la persona que esta hablando en disimiles circunstancias e independiente del contenido del habla. En el área existen varios sistemas de referencias, donde la gran mayoría están basados en enfoques estadísticos y aunque presentan buenos resultados, la eficacia de los algoritmos de reconocimiento se encuentra afectada por la insuficiente información discriminatoria del locutor. En esta investigación, conducente a la Tesis de Doctorado del autor, se realizó un estudio donde se identificaron dos principales debilidades en

las representaciones actuales de la voz del locutor. Primero, no se tiene en cuenta el comportamiento temporal de la voz, siendo este un rasgo discriminatorio del locutor. Segundo, los eventos pocos frecuentes dentro de una población de locutores pero frecuentes en un locutor dado, apenas son tenidos en cuenta en los enfoques actuales, lo cual es contradictorio cuando el objetivo es discriminar los locutores. Motivado por la solución de estos problemas se obtuvieron como resultados principales:

- Una nueva representación binaria de la expresión de voz, capaz de contener los eventos discriminatorios desde los frecuentes hasta los pocos frecuentes.
- Se introdujo la información común de los locutores en los algoritmos de compensación de la variabilidad de sesión, implicando una mejora en su desempeño.
- Se desarrolló un algoritmo para obtener la información temporal de la expresión de voz y se incorporó en la nueva representación.

Los algoritmos propuestos se compararon contra los mejores reportados en la literatura más actual, sobre conjuntos de datos del repositorio internacional NIST SRE. Adicionalmente, parte de los algoritmos propuestos ya se encuentran aplicados en el proyecto “Sistema de verificación remota de personas por la voz (Verivoz)” desarrollado de conjunto con la Dirección General de la Policía Nacional Revolucionaria (DGPNR), y en el proyecto “Conjunto de funciones para la identificación biométrica por la voz (BioVoz)” que se encuentra en desarrollo por parte de nuestra empresa (Datys). Adicionalmente todos los resultados obtenidos en la investigación tributarán a proyectos futuros, como es la segmentación de locutores en conversaciones telefónicas, permitiendo contar con sistemas competitivos internacionalmente donde intervenga el Reconocimiento de Locutores.

1. **Problemática**

La voz es una de las características Biométricas sobre la que se ha investigado y desarrollado un grupo significativo de sistemas de reconocimiento automático de la persona, llamados también sistemas de reconocimiento automático del locutor (ASR por sus siglas en inglés). El objetivo de estos sistemas consiste en la verificación[1] o identificación[2] automática de una persona a través de su voz y en nuestro caso, texto-independiente[3]. El hecho de poder distinguir un locutor de otro está relacionado mayoritariamente con las características fisiológicas y los hábitos lingüísticos de cada uno de ellos. El reconocimiento conlleva un procesado de la voz, permitiendo extraer rasgos acústicos inherentes al locutor y la posterior búsqueda de posibles coincidencias mediante un proceso de reconocimiento de patrones.

Hoy en día, el desarrollo de nuevos algoritmos en el área del reconocimiento del locutor texto-independiente continúa siendo objeto de interés debido a su amplia variedad de aplicaciones y los complejos factores que afectan las señales de voz, estos son: la variabilidad del canal de transmisión, el ruido en la señal de voz, el estado emocional, etc. En general, cualquier variación provocada por estos factores se conoce como variabilidad de sesión, esta variabilidad se describe como la diferencia de condiciones entre el objeto de entrenamiento y el objeto de prueba, y constituye el principal reto a que se enfrentan los ASR.

La mayoría de los algoritmos reportados en la literatura para los sistemas de ASR se basan fundamentalmente en la representación del locutor en el contexto de los Modelos de Mezclas Gaussianas (GMM, por sus siglas en inglés) y su representación universal, conocida como Modelo Universal de Fondo (UBM, por sus siglas en inglés). El UBM ha de contener información acústica de los más diversos locutores posibles, para luego obtener un modelo por cada locutor derivado del UBM, utilizando la adaptación Máximo a Posterior (MAP, por sus siglas en inglés). Este modelo adaptado contiene entre otra información, una gran cantidad de información discriminatoria correspondiente al locutor.

Hace algunos años el enfoque de los super-vectores (SV) se impuso como representación del locutor en los sistemas ASR, dicho enfoque tiene como base la adaptación MAP del UBM a los modelos GMM de los locutores, posibilitando en este marco que cada expresión de voz se represente por un super-vector que se obtiene de la concatenación de todos los vectores de medias de los componentes Gaussianos del GMM. Estos super-vectores forman un espacio de representación de altas dimensiones, permitiendo el modelado directo de la variabilidad de sesión[4] para su compensación. Más recientemente, dos nuevas soluciones han sido propuestas en el marco del enfoque de los super-vectores: el Análisis de Factores Conjunto (JFA, por sus siglas en inglés) y el vector identidad (i-vector) en el sentido del reconocimiento del locutor.

Los algoritmos desarrollados sobre los super-vectores y los i-vectores han mostrado los mejores niveles de rendimiento en las competencias del Instituto Nacional de Estándares y Tecnología (NIST, por sus siglas en inglés), pero presentan *dos inconvenientes principales*:

[1] En la verificación del locutor el objetivo consiste en determinar si una persona es quien afirma ser o no, a partir de una muestra de su voz.

[2] En la identificación del locutor el objetivo es comparar una la muestra de voz de un locutor desconocido entre muestras de hablantes conocidos, determinando si alguna de las muestras de hablantes conocidos proviene del locutor desconocido.

[3] Se refiere a la independencia del contenido fonético del habla.

[4] Dicha variabilidad se observa al contar con varias expresiones de locutores en diferentes momentos y por diferentes canales telefónicos.

1. Resulta imposible trabajar con la información temporal del habla, ya que cada conjunto de vectores acústicos está representado por solo un punto en el espacio de los i-vectores.
2. Estos enfoques se basan en evaluaciones de modelos estadísticos, donde la influencia de una información específica se determina principalmente por la frecuencia de esta información. Es decir: si se produce un evento a menudo para un locutor dado, pero muy rara vez para los otros, apenas será tenido en cuenta por estos enfoques, lo cual podría parecer como una paradoja cuando el objetivo es discriminar los locutores.

2. **Novedad científica**

El principal impacto científico de este trabajo lo constituyen los aportes que se hacen al Reconocimiento del Locutor texto-independiente, específicamente este nuevo enfoque parte de una representación del habla, que se desplaza del área de trabajo probabilística continua a un espacio discreto y binario. Esta representación se basa en decisiones binarias locales, tomadas por cada una de las observaciones (vectores acústicos). A diferencia de los enfoques anteriores, esta área de trabajo binaria es capaz de modelar los eventos discriminatorios frecuentes y poco frecuentes, dado que representa un extracto de la voz mediante una matriz binaria, donde cada vector acústico está representado por un vector binario. Además la matriz binaria está ordenada temporalmente permitiendo la extracción de la información temporal discriminatoria del locutor. Los principales aportes son:

1. Se desarrolló un método para evaluar la presencia de información redundante en el espacio de los super-vectores, utilizando algoritmos basados en técnicas no lineales de reducción: Isomap y Laplaciano. Los resultados de este método mostraron claramente la información redundante presente en la representación por super-vectores, utilizando la técnica Isomap se logró una reducción de dimensión en un factor de cuatro y no hubo prácticamente pérdidas en términos de eficacia. Además, utilizando la técnica del Laplaciano para reducir la dimensión en un factor de dos, se superó el resultado obtenido por una técnica lineal de reducción (Análisis de Componentes Principales), lo que muestra la importancia de tener presente la naturaleza interna de los datos. Este resultado constituyó uno de los primeros pasos en el campo de reconocimiento del locutor, donde se utilizó la información topológica contenida en el espacio acústico. Ver las publicaciones 1 y 2 de la lista de publicaciones.

2. Dentro de un nuevo enfoque basado en una representación binaria de la expresión de voz para reconocimiento del locutor, se propuso un método para la obtención del modelo generador y una nueva medida de similitud asociada con una representación global (vector acumulativo) de la información existente en la matriz binaria, dicha representación tiene en cuenta los eventos discriminatorios, desde los más frecuentes hasta los pocos frecuentes, en una señal de voz. Al comparar los resultados con los obtenidos por los métodos en el estado del arte, se comprobó que requieren considerablemente menos recursos de cómputo y memoria, mostrando un nivel de rendimiento similar al enfoque GMM-MAP. Este enfoque binario abrió nuevos caminos para continuar las investigaciones en el enfrentamiento a los problemas de variabilidad de sesión y a la explotación de la información temporal existente en la voz. Ver las publicaciones 3 y 4 de la lista de publicaciones.

3. Se propuso un nuevo método de compensación de variabilidad de sesión que incluye la información común entre los locutores, dicha propuesta modificó tres algoritmos de compensación de variabilidad del estado del arte. Esta modificación consiste en utilizar una nueva variante de la matriz de dispersión intra-clase para los vectores acumulativos e i-vector, que es capaz de contener no solo los atributos no deseados, sino también la información común a los locutores, que tampoco es deseada. Se logró, con la nueva variante propuesta para la compensación sobre la representación binaria, una mejora general de la eficacia en los algoritmos para la verificación del locutor. Este resultado permitió comprobar la importancia de incluir la información común en la matriz de dispersión intra-clase, lo que confirmó el análisis de la varianza espectral realizado. También se propuso un método simple (máscara) basado en la varianza de los vectores acumulativos de los locutores, para seleccionar los componentes Gaussianos más discriminatorios. Se mostró, además, que estos algoritmos de compensación pueden ser generalizados a otros enfoques (por ejemplo i-vector), donde se realizaron comparaciones con los métodos del estado-del-arte, mostrando que su empleo mejora la eficacia en el reconocimiento del locutor. Ver las publicaciones 5, 7 y 8 de la lista.
4. Se propusieron dos nuevos métodos para capturar e incluir, la información temporal existente en las expresiones de voz, en el reconocimiento del locutor. Se propuso un método para extraer la información temporal a nivel de segmentos de la voz, logrando que el vector acumulativo de un segmento reflejé la distribución de las especificidades relativas al contenido fonético, lo cual es útil y aplicable en diferentes áreas de procesamiento de la voz. Este resultado demostró un nivel de rendimiento comparable, respecto a la eficacia, con el enfoque GMM-MAP, pero con una mejor eficiencia. Se propuso un método para obtener una representación dinámica de la expresión de voz que refleje la información temporal a nivel de tramas, que demostró una mejora en la eficacia respecto a la obtenida con los vectores acumulativos. Se aplicó a dichas representaciones temporales una nueva variante del método de compensación de variabilidad ya propuesto, logrando una mejora apreciable del rendimiento en los algoritmos de verificación del locutor. Se observó además, al fusionarse las puntuaciones obtenidas, que ambos dominios (el enfoque binario y el i-vector) contienen información complementaria. Notar que la representación de la información temporal sobre el enfoque binario, constituye un nuevo paso en el reconocimiento del locutor el cual es imposible obtener en el enfoque i-vector. Ver las publicaciones 3, 6 y 8 de la lista.

3. **Impacto del resultado**

En varios órganos operativos del MININT se llevan a cabo investigaciones de voz, que posibilitan el reconocimiento de locutores con fines periciales o investigativos. Todos los algoritmos desarrollados fueron implementados de forma sencilla y adecuadamente documentados para ser utilizados en la solución de dicha problemática.

Los algoritmos desarrollados constituyen aportes al conocimiento mundial en el área del Reconocimiento de Patrones. Los resultados alcanzados se encuentran publicados en revistas seriadas, eventos de prestigio internacional, eventos nacionales y una tesis doctoral. Además como fruto de la tesis se defendieron dos tesis de grado y una de maestría.
Los resultados teóricos obtenidos se han aplicado en la solución de problemas prácticos. Varios algoritmos de reconocimiento de locutores se encuentran introducidos en la práctica:

- Se implementaron algoritmos propios del estado del arte para el reconocimiento del locutor texto-independiente, sobre una plataforma que permitió sustituir la plataforma extranjera hasta entonces utilizada. Además, estos algoritmos se encuentran ya introducidos en los proyectos "Sistema de verificación remota de personas por la voz (Verivoz)" en explotación por la DGPNR en 6 municipios de la capital y "Conjunto de funciones para la identificación biométrica por la voz (BioVoz)" que se encuentra en desarrollo por parte de nuestra empresa (Datys).
- Se incorporó a ambos proyectos un nuevo algoritmo propio, desarrollado en la investigación, el cual compensa la variabilidad de sesión mejorando la eficacia del reconocimiento del locutor.

Por otra parte, los resultados de la investigación estarán presentes en el nuevo proyecto futuro, que permitirá segmentar y reconocer automáticamente la identidad de cada locutor durante la ejecución de llamadas telefónicas con múltiples hablantes. Además, contar con métodos desarrollados con *conocimiento propio*, constituye un aspecto básico para garantizar la sostenibilidad de los resultados alcanzados. No solo permite la actualización y modernización permanente de los mismos, sino que se llega a alcanzar la total independencia tecnológica en los aspectos desarrollados. Además permite seguir incursionando en la obtención de nuevos resultados, novedosos desde el punto de vista científico, para enfrentar nuevas problemáticas.

4. **Principales publicaciones**

1. Gabriel Hernández-Sierra and José R. Calvo and F. J. Reyes and R. Fernández: "Simple Noise Robust Feature Vector Selection Method for Speaker Recognition". CIARP'09, Lecture Notes in Computer Science, ISSN 0302-9743, vol. 5856, pp. 313–320, 2009.
2. Gabriel Hernández-Sierra and Jean-François Bonastre and Driss Matrouf and José R. Calvo: "Topological representation of speech for speaker recognition". Conference of the International Speech Communication, ISSN 1990-9772, pp. 2134-2137, 2010.
3. Jean-François Bonastre, Xavier Anguera, Gabriel Hernández-Sierra, Pierre M. Bousquet: "Speaker modeling using local binary decisions". Conference of the International Speech Communication, ISSN 2072-6287, pp. 13-16, 2011.
4. Gabriel Hernández-Sierra, Jean François Bonastre, José R. Calvo: "Speaker recognition using a binary representation and specificities models". CIARP 2012, Lecture Notes in Computer Science, ISSN 0302-9743, vol. 7441, pp. 732-739, 2012.
5. Manuel A. Martínez, Gabriel Hernández-Sierra, José R. Calvo: "Speaker Verification Using Accumulative Vectors with Support Vector Machines". CIARP 2013, Lecture Notes in Computer Science, ISSN 0302-9743, vol. 8259, pp. 350-357, 2013.
6. Gabriel Hernández-Sierra, José R. Calvo, Jean François Bonastre: "Temporal Information in a binary framework for Speaker Recognition". CIARP 2014, Lecture Notes in Computer Science, ISSN 0302-9743, vol. 8827, pp. 207–213, 2014.
7. Gabriel Hernández-Sierra, José R. Calvo, Jean-François Bonastre, Pierre M. Bousquet: "Session compensation using binary speech representation for speaker recognition". Journal of Pattern Recognition Letters, ISSN 0167-8655, vol. 49, pp. 17-23, 2014.
8. Gabriel Hernández-Sierra, José R. Calvo, Jean-François Bonastre: "Session variability compensation on the temporal information for speaker recognition". In: XVI Convention of Electrical Engineering, CIE-2015, ISBN: 978-959-312-025-8.