RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Métodos de compensación de variabilidad en el reconocimiento de locutores

Flavio J. Reyes Díaz, Gabriel Hernández Sierra y José R. Calvo de Lara

RT_072 **marzo 2015**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada
MINISTERIO DE LA INDUSTRIA BÁSICA

REPORTE TÉCNICO
Reconocimiento de Patrones

Métodos de compensación de variabilidad en el reconocimiento de locutores

Flavio J. Reyes Díaz, Gabriel Hernández Sierra y José R. Calvo de Lara

RT_072 **marzo 2015**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Tabla de contenido

1. Introducción 1
2. Principales algoritmos para la verificación de locutores 3
   - 2.1. Compensación de la variabilidad de sesión en el espacio i-vector 3
     - 2.1.1. Normalización de la Covarianza Intra-clase (WCCN) 3
     - 2.1.2. Análisis discriminante lineal (LDA) 4
     - 2.1.3. Verificación de locutores en el espacio i-vector con compensación de variabilidad de sesión 4
   - 2.2. Análisis discriminante lineal probabilístico 5
     - 2.2.1. Verificación de locutores basado en el modelo PLDA sobre el espacio i-vector 6
3. Variabilidad de sesión en el reconocimiento de locutores 7
   - 3.1. Variabilidad en la duración de las señales 8
   - 3.2. Variabilidad por el tipo e intensidad del ruido en las señales 11
   - 3.3. Variabilidad debida a la reverberación en las expresiones de voz 13
4. Resultados experimentales 15
   - 4.1. Análisis del efecto en la variabilidad producido por los segmentos de corta duración en el espacio i-vector 16
5. Conclusiones Generales 20

Métodos de compensación de variabilidad en el reconocimiento de locutores

Flavio J. Reyes Díaz, Gabriel Hernández Sierra y José R. Calvo de Lara

Equipo de Procesamiento de Imágenes y Señales, Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV), La Habana, Cuba
{freyes, gsierra, jcalvo}@cenatav.co.cu

RT_072, Serie Azul, CENATAV
Aceptado: 19 de febrero de 2015

Resumen. La variabilidad existente en las señales de voz en ambiente reales representan un obstáculo para alcanzar robustez en el reconocimiento de locutores. En este trabajo se describen algunos de los tipos de variabilidad más comunes y se realiza un análisis de los métodos desarrollados para enfrentar esta problemática. Se realiza un grupo de experimentaciones para confirmar los resultados obtenidos por otros autores ante estos tipos de variabilidad y se propone un algoritmo de compensación de variabilidad para enfrentar la variabilidad que se origina por las diferentes duraciones de los segmentos de habla de entrenamiento y prueba, obteniendo una mejora de 3,4 % y 5,6 %, utilizando la distancia del coseno como medida de similitud y como clasificador el Análisis Discriminante Lineal probabilístico (PLDA), respectivamente.

Palabras clave: reconocimiento de locutores, variabilidad de la voz, duración de la señal de voz, ruido, reverberación, métodos de compensación de variabilidad.

Abstract. The variability in speech signals in real environment represents an obstacle to achieving robustness in speaker recognition. This paper describes some of the most common types of variability and presents an analysis of the developed methods to address this problem. A group of experiments to confirm the obtained results by other authors to these types of variability were done. A variability compensation algorithm is proposed to deal with different durations of the speech segments of training and testing. An improvement of 3.4 % and 5.6 % was obtained using cosine distance and probabilistic linear discriminant analysis (PLDA) as classifiers, respectively.

Keywords: speaker recognition, voice variabilities, short utterance, noisy, reverberation, variabilities compensation methods.

1. Introducción

La voz es utilizada como la base de aplicaciones automáticas para facilitar la comunicación entre las personas, soportada sobre las nuevas tecnologías de infocomunicaciones:

- la identificación del idioma hablado
- el reconocimiento y síntesis del habla
- el reconocimiento biométrico de las personas por la voz

El reconocimiento biométrico de personas por su voz es el proceso mediante el cual se verifica[1] o identifica[2] a una persona por su voz. En esta área del reconocimiento biométrico se han desarrollado disímiles algoritmos con el objetivo de resolver las distintas problemáticas que han ido surgiendo de conjunto con el desarrollo de las tecnologías, principalmente en el procesamiento y reconocimiento sobre las señales de voz.

Con el objetivo de hacer más robustos los sistemas de reconocimiento de locutores y mejorar los resultados se han venido utilizando diversos algoritmos de clasificación como: los modelos de mezclas gaussianas (GMM) [1,2], la adaptación (MAP) de dichos modelos [2,3] y las máquinas de vectores de soporte (SVM) [4,5].

Posteriormente surgió un enfoque para los sistemas de verificación de locutor, basado en los supervectores [5] que posibilitó la compensación de la variabilidad de sesión[3]. En este nuevo enfoque la variabilidad es combatida a nivel de rasgos [6], mediante la utilización de ambos modelos, generativos y discriminatorios, las GMM y las SVM como se expone en [7] y [8] respectivamente; a través de normalizaciones de la puntuación, como: Hnorm, Tnorm, Znorm y ZTnorm [9,10,11]. Tambien han surgido técnicas para la compensación de la variabilidad de sesión como el Análisis de Factores (FA) [12], la cual modela la variabilidad de sesión basándose en enfoques estadísticos, mediante el algoritmo de máxima expectancia (EM); o la proyección de atributos no deseados (NAP) [8], que estima una matriz de proyección para intentar reducir la variabilidad de sesión.

Recientemente, en [13] se propone para la compensación de la variabilidad, la utilización de un espacio de variabilidad total (T), que incluye la variabilidad propia del locutor y de la sesión. Este espacio es definido por una matriz de vectores propios correspondientes a los mayores valores propios, obtenidos de la matriz de covarianza a partir de las estadísticas de Baum-Welch de 1^{er} y 2^{do} orden de una población de locutores. A partir de aquí se obtiene el vector intermedio o i-vector, que es una variable oculta que se corresponde con la media de una distribución gaussiana obtenida a partir de las estadísticas de Baum-Welch de un segmento de voz perteneciente a un locutor a reconocer.

Recientemente se ha logrado un incremento de la eficacia en el funcionamiento de los sistemas de reconocimiento de locutores a partir de un nuevo algoritmo como el Análisis discriminante lineal probabilístico (PLDA) [14]. El mismo se apoya en la ventaja de que los i-vectores, son definidos como una transformación de un segmento de habla a un vector de bajas dimensiones y los utiliza como datos de entrada.

Desde hace unos años los sistemas de reconocimiento de locutores han venido sufriendo un aumento del número de obstáculos, que empeoran los resultados en el reconocimiento de personas por la voz, principalmente cuando se insertan en aplicaciones que funcionan en condiciones reales. Esto viene dado a partir del aumento del desarrollo y la utilización cada vez mayor de nuevas tecnologías de infocomunicaciones que procesan la voz, desde casi cualquier ambiente y en cualquier circunstancia: telefonía por internet, inalámbrica, satelital y celular, controles de acceso por voz a instalaciones y redes, etc. Este incremento en la utilización de la voz bajo muy diversas condiciones conllevan nuevos retos al enfrentar el análisis forense de la misma debido a todos los factores que le incorporan variabilidad: ruidos, reverberaciones, distorsiones del canal, duraciones, estado emocional y cooperatividad del hablante, entornos comunicativos, etc.

[1] *Verificación de una identidad reclamada por el locutor, a partir de una muestra de voz con identidad desconocida.*
[2] *Dada una muestra de voz, señalar, dentro de un grupo de personas, los propietarios más probables de la muestra.*
[3] *La variabilidad de sesión en el reconocimiento de locutores se asocia con aquellos factores que provocan diferencias entre dos muestras de voz a comparar, captadas en diferentes momentos o sesiones.*

Este reporte propone describir, dentro de los métodos del estado del arte del reconocimiento de locutores, cuáles han sido las principales vías que se han utilizado para enfrentar algunas de las formas de variabilidad de sesión, presentes sobre todo en el análisis forense de la voz.

2. Principales algoritmos para la verificación de locutores

En la actualidad los sistemas para el reconocimiento de locutores sobre el espacio de los i-vectores enfrentan la variabilidad de sesión a partir de métodos de compensación integrados a los mismos. En esta sección se describen los principales sistemas y las variantes de compensación más comunes.

2.1. Compensación de la variabilidad de sesión en el espacio i-vector

Dado que T es una matriz de variabilidad total, que contiene la información de la variabilidad de los locutores y de las sesiones, es usual que en los i-vectores existan ambas informaciones y que persista el problema de la variabilidad. Se hace necesario aplicar técnicas capaces de compensar la variabilidad que es ajena a la información discriminatoria de los locutores.

2.1.1. Normalización de la Covarianza Intra-clase (WCCN)

La Normalización de la covarianza intra-clases[4] (WCCN) fue propuesta por Dehak et al. en [13], para la compensación de la variabilidad de sesión en el espacio i-vector.

El objetivo de esta variante de compensación de sesión es minimizar la variabilidad intralocutor (en un mismo locutor), a partir de la covarianza promedio de un conjunto de locutores, la cual se formula:

$$\varphi(x) = B'x, \tag{1}$$

donde x es un i-vector y B es una matriz de proyección de orden superior obtenida de la descomposición de Cholesky de $W_{-1} = BB'$, donde W es la covarianza promedio de la variabilidad intra clase de un conjunto locutores y se define como:

$$W = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{n_l}\sum_{i=1}^{n_l}(x_i^l - x_l)(x_i^l - x_l)', \tag{2}$$

donde L es el número de locutores, n_l es la cantidad de segmentos de habla por cada locutor s, x_l es el vector resultante del cálculo de la media entre todos los i-vectores correspondientes a un locutor y x_i^l es i-ésimo *ivector* del locutor.

Existen dos variantes para obtener la similitud entre dos i-vectores x_1 y x_2 compensados en función de la variabilidad intra-clase:

$$S(x_1, x_2) = (B'x_1)'(B'x_2), \tag{3}$$

y la propuesta por Dehak [13]:

$$S_{cos}(x_1, x_2) = \frac{(B'x_1)'(B'x_2)}{||B'x_1||||B'x_2||}. \tag{4}$$

[4] *Variabilidad contenida en una clase (un locutor)*

2.1.2. Análisis discriminante lineal (LDA)

El análisis discriminante lineal [15] es un método de reducción de dimensionalidad muy utilizando actualmente en el espacio i-vector, propuesto inicialmente por Dehak et al. en [13], como otra vía para la compensación de la variabilidad de sesión. El objetivo principal del `LDA` es aumentar la diferencia inter-clases[5] mediante la maximización de la covarianza entre las clases (de una población de locutores) y la minimización de la covarianza intra-clase (en un locutor). Estas varianzas son obtenidas a partir de las siguientes ecuaciones:

$$S_b = \sum_{l=1}^{L}(x_l - \bar{x})(x_l - \bar{x})', \tag{5}$$

$$S_w = \sum_{l=1}^{L}\frac{1}{n_l}\sum_{i=1}^{n_l}(x_i^l - x_l)(x_i^l - x_l)', \tag{6}$$

donde S_b y S_w son las matrices de covarianza inter-clases e intra-clase respectivamente, $\bar{x}$ es la media de todos los i-vectores de una población de locutores dada. L es el número de locutores y n_l es el número de segmentos de habla pertenecientes a un mismo locutor. A partir de estas dos matrices de variabilidad se define una matriz de proyección A, como se describe en la ecuación 7, que se compone por los vectores propios correspondientes a los mayores valores propios.

$$A = \frac{v' S_b v}{v' S_w v}. \tag{7}$$

La variante utilizada para obtener el valor de similitud (referencia) entre dos i-vectores es:

$$S_{cos}(x_1, x_2) = \frac{(A'x_1)'(A'x_2)}{||A'x_1||||A'x_2||}. \tag{8}$$

2.1.3. Verificación de locutores en el espacio i-vector con compensación de variabilidad de sesión

El sistema de verificación de locutores mediante la variante de compensación `WCCN` y/o `LDA` sobre el espacio i-vector se observa en la figura 1.

En la fase de entrenamiento se obtienen las matrices y modelos necesarios para la extracción de los i-vectores y para la compensación de sesión, como son el modelo universal de background (UBM) y la matriz de variabilidad total (T), representa el espacio de variabilidad total; en ambos casos se realiza un entrenamiento a partir de un conjunto de segmentos que caracterizan a una población, dado que es necesario obtener la mejor representación posible de las clases acústicas de la población en el UBM y la mayor variabilidad posible en la matriz T. Para la obtención de las matrices de proyección A y B, donde A se obtiene mediante el algoritmo LDA y B mediante el algoritmo WCCN se parte de un conjunto de segmentos de voz etiquetados por locutor, adquiridos en sesiones diferentes.

A la llegada de dos segmentos de voz a verificar se les calculan las estadísticas de Baum Welch de 1^{er} y 2^{do} orden de los segmentos de voz, respecto a las componentes gaussianas del modelo UBM. Posteriormente con las estadísticas obtenidas y la matriz T se extraen los i-vectores correspondientes.

Para realizar la compensación de la variabilidad de sesión de ambos i-vectores y el cálculo de la puntuación entre ellos se utiliza la ecuación 8 en el caso de que se desee realizar la compensación

[5] *Variabilidad entre las clases (entre locutores)*

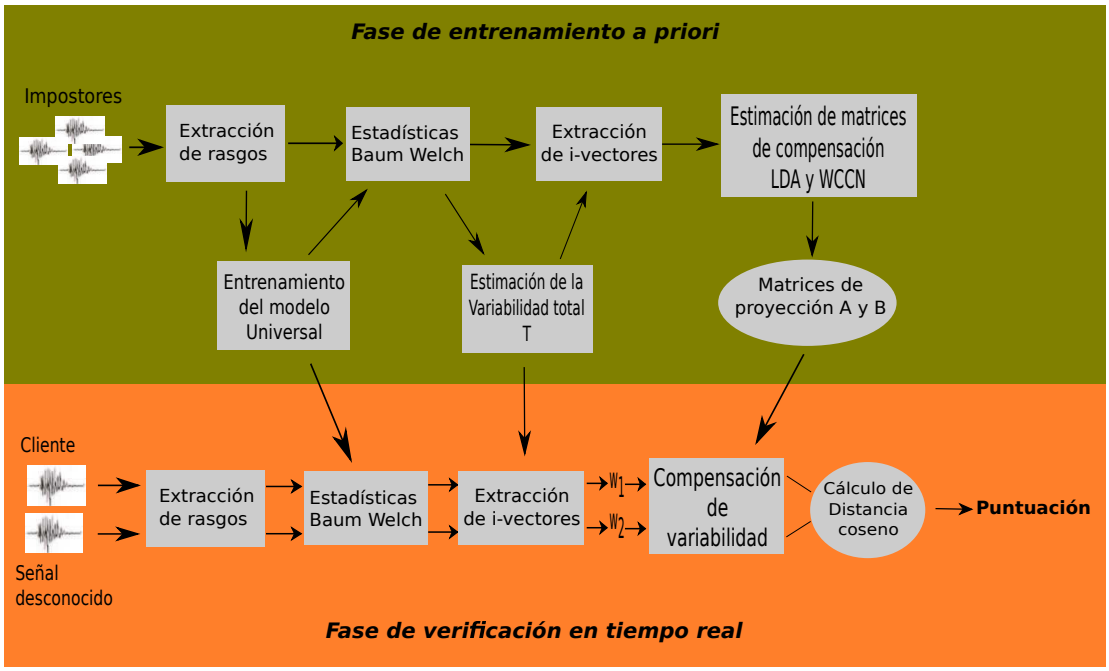

Fig. 1. Sistema de verificación de locutores en el espacio i-vector, realizando compensación de la variabilidad de sesión.

de la variabilidad inter-clases e intra-clase con el algoritmo LDA, o la ecuación 4 en el caso que solo sea de la compensación de la variabilidad intra-clase con el algoritmo WCCN.

Para compensar la variabilidad de sesión y obtener la puntuación entre dos i-vectores se puede utilizar la combinación de los algoritmos LDA y WCCN mediante la ecuación 9, siempre y cuando los i-vectores utilizados para estimar de la matriz de proyección B, estén compensados en función de la variabilidad de sesión con la matriz A.

$$S_{cos}(x_1, x_2) = \frac{(A'x_1)'B^{-1}(A'x_2)}{\sqrt{(A'x_1)'B^{-1}(A'x_1)}.\sqrt{(A'x_2)'B^{-1}(A'x_2)}}. \tag{9}$$

2.2. Análisis discriminante lineal probabilístico

El Análisis discriminante lineal probabilístico (PLDA) es un modelo generativo desarrollado por Price et al. en [16]. Este modelo mantiene una estrecha relación con la técnica JFA.

En el reconocimiento de locutores la variante más utilizada es el Análisis discriminante lineal probabilístico de tipo gaussiano(G-PLDA), propuesto por kenny et. al en [14], donde considera que cada i-vector x correspondiente al locutor L, está constituido por:

$$x = \mu + \phi y_s + \epsilon, \tag{10}$$

donde μ es el vector de medias de las clases de una población (UBM), ϕ es una matriz rectangular que representa el espacio de la información del locutor (voces propias), y_s es una matriz de identidad latente que mantiene una distribución normal y ϵ representa los términos ruidosos, los cuales mantienen una distribución gaussiana, con media cero y matriz de covarianza completa

$\sum$. Estos parámetros correspondientes al modelo son estimados a partir del algoritmo de máxima expectancia (EM) como se describe en [16].

El cálculo de puntuación o similitud entre dos i-vectores x_1 y x_2 mediante el G-PLDA se obtiene a partir de la razón de probabilidades condicionales entre dos i-vectores, como se observa en la siguiente ecuación:

$$similitud(x_1, x_2) = log \frac{prob(x_1, x_2|H_1)}{prob(x_1|H_0)prob(x_2|H_0)}, \tag{11}$$

donde los i-vectores x_1 y x_2, son los correspondiente al cliente y a la identidad desconocida respectivamente y H_1 y H_0, constituyen las hipótesis de correspondencia de los dos i-vectores a un mismo locutor o la correspondencia a locutores diferentes, respectivamente.

2.2.1. Verificación de locutores basado en el modelo PLDA sobre el espacio i-vector

Un sistema de verificación de locutores basado en PLDA se observa en la figura 2, para entrenar un modelo PLDA se hace necesario realizar los entrenamientos previos del modelo UBM y la matriz de variabilidad total T, como se describe en la fase de entrenamiento de la figura 1; con el objetivo de obtener los i-vectores correspondientes al conjunto de segmentos de habla de una población de impostores que será la utilizada para el entrenamiento del modelo G-PLDA. Esta población de impostores puede ser la misma utilizada en las estimaciones del modelo UBM o la matriz T. Estos i-vectores se compensan en función de la variabilidad de sesión mediante la matriz de proyección obtenida con el algoritmo LDA, descrito en la sección 2.1.2.

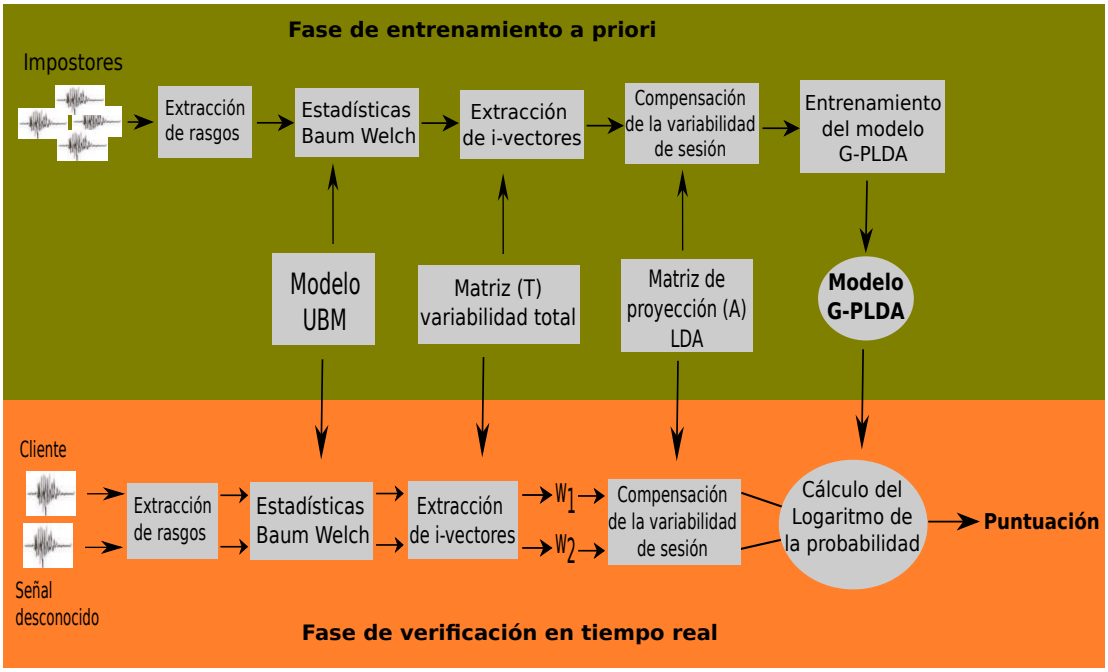

Fig. 2. Diagrama de un sistema IV-PLDA para verificación de locures.

En el momento de la verificación son pre-procesadas los dos segmentos de habla a comprobar, donde se les extraen y compensan sus i-vectores. Para calcular la puntuación o similitud entre los dos i-vectores se utiliza la ecuación 11.

3. Variabilidad de sesión en el reconocimiento de locutores

Teniendo en cuenta que los sistemas automáticos de reconocimiento de locutores han demostrado alcanzar una gran precisión en condiciones controladas, representan una variante atractiva para emplear en tareas de reconocimiento de locutores sobre escenarios forenses, basándose en que estos escenarios se vinculan a grandes cantidades de muestras de voces a procesar en un limitado tiempo que presentan las investigaciones.

En escenarios forenses cuando el especialista necesita precisar, con la mayor exactitud posible, cuál o cuáles personas sospechosas son las más probables "dueñas"de la expresión de habla identificar, existe gran variabilidad en las condiciones de las muestras de habla, teniendo en cuenta que las grabaciones son realizadas en ambientes reales. El reconocimiento automático de locutores en condiciones forenses, dista mucho aún de tener una eficacia alta, relativa a la eficacia obtenida cuando no existe tal variabilidad; debido a la variabilidad de sesión existente entre la(s) muestras de entrenamiento y la muestra a identificar.

En estos escenarios existen diferentes tipos de variabilidades, algunas de estas, están más presentes que otras en el ámbito forense: el tipo de canal y sus implicaciones en cuanto a compresión y codificación de la señal de habla, el tipo y la intensidad del ruido, la reverberación, el estado emocional y el esfuerzo vocal del hablante, la duración de la expresión de habla, constituye un reto enfrentarse al problema de la variabilidad en el reconocimiento de locutores, sobre todo cuando sus tipos pueden estar combinados en las muestras a comparar, lo cual es muy común.

Un grupo del proyecto SRI[6] en [17] llevan a cabo una agrupación las variaciones en las muestras de habla que afectan la calidad del reconocimiento del locutor.

Las variaciones intrínsecas son aquellas que dependen de la persona y del contexto como:

- Gritar
- Estilo del habla (¿qué se dice?)
- Condición física del locutor (emoción, intoxicación, enfermedad)
- Esfuerzo vocal
- idioma
- envejecimiento
- Duración de la muestra de voz

Las variaciones extrínsecas son los aspectos independientes a la persona, que afectan la calidad de las señales acústicas:

- ambientes acústicos, ruido, reverberación
- Variabilidad de canal (micrófonos, diferentes tipos de handset, diferentes tecnologías de grabación)
- Duración de la muestra de voz
- alta relación señal a ruido (buena calidad)
- degradación del audio debido a la compresión

No existen bases de datos para reconocimiento de locutores de acceso público y no restringido, con características propiamente forenses. Algunos de los tipos de variabilidad, existentes en los escenarios forenses, pueden encontrarse en diferentes bases de datos, utilizadas para la evaluación de los sistemas de reconocimiento de locutores. En la tabla 1 se muestran los tipos de variabilidad

[6] Laboratorio de investigación y tecnologías del habla (http://www.sri.com/)

que están presentes en las bases de voces de las evaluaciones SRE-NIST[7] [18] y otras [19,20] utilizadas para evaluar la robustez de los sistemas de reconocimiento de locutores.

Tabla 1. Relación de variabilidades existentes en las señales de las bases de datos utilizadas en la evaluación de los sistemas de reconocimiento de locutores.

Variabilidad/Bases	2004	2005	2006	2008	2010	2012	RSR2015 [19]	PRISM [20]
Canal	X	X	X	X	X	X	X	X
Idioma		X	X	X		X		X
Duración					X	X	X	X
Esfuerzo vocal					X	X		X
Ruido						X		X
Reverberación						X		X

La variabilidad en la duración, en el tipo y la intensidad del ruido, la reverberación y la compresión pueden ser generadas de forma artificial, como por ejemplo: la herramienta pública FaNT(http://dnt.kr.hsnr.de/download.html) permite la adición de ruido a las señales; en el caso de la duración, las señales pueden truncarse manualmente, seleccionando el número de tramas con las que se desee trabajar o basándose en la energía de la señal y en el caso de la compresión, se pueden utilizar diferentes herramientas para la codificación y compresión de señales.

A continuación se abordan los tipos de variabilidad más comunes en las señales adquiridas en escenarios forenses, como la variabilidad en la duración de los segmentos de habla, la variabilidad en el tipo e intensidad del ruido y la variabilidad en la reverberación de los segmentos de habla. Estos tipos de variabilidad pueden ser simulados, lo que ha posibilitado su estudio y la propuesta de diversas soluciones para compensarlos.

3.1. Variabilidad en la duración de las señales

El análisis de la variación de la duración de las señales en la verificación de locutores es un aspecto que ha venido tomando auge, debido al incremento de aplicaciones de autenticación biométrica e identificación forense por la voz, que están expuestas al trabajo con segmentos de acústicos adquiridos en ambientes reales, con grandes cambios en el tiempo de duración de los mismos. Se han venido desarrollando experimentaciones donde el factor principal es la diferencia de duración que existe entre los segmentos de habla que intervienen en los entrenamientos de los modelos y matrices (modelos UBM y PLDA, matrices de compensación de sesión) y los segmentos de habla utilizados en la prueba.

Un análisis del efecto causado por la duración de las señales en la calibración de las puntuaciones sobre el espacio de los i-vectores es propuesto por Mandasari y otros en [21] para aplicaciones forenses. La clasificación y calibración en los sistemas i-vectores son muy sensibles a las variaciones en la duración de los segmentos de habla para el entrenamiento y la prueba, y a la no homogeneidad respecto a la duración de los segmentos de habla en los datos de entrenamiento. Los autores plantean que esta problemática puede ser combatida mediante la utilización de similares duraciones de los segmentos de habla empleados en el entrenamiento y en las pruebas.

[7] NIST-SRE: evaluaciones de reconocimiento de locutores llevadas a cabo por el instituto Nacional de Estandarización de los EEUU

También para una eficaz recalibración de los datos, para la normalización de las puntuaciones, en presencia de segmentos de habla con variación en sus duraciones, es preciso utilizar segmentos de habla con similares duraciones, ya sean creados en laboratorio o no. Otras de las conclusiones alcanzadas por los autores es almacenar los valores de calidad de cada segmento de habla que intervienen en los entrenamientos, la cual es representada por la duración de estos segmentos; para poder realizar una calibración del modelo con segmentos de habla con duraciones similares.

En [22] Kanagasundaram y otros realizan una evaluación del uso de diferentes técnicas de compensación, WCCN, NAP [8] y LDA y de los algoritmos JFA y G-PLDA; en verificación de locutores sobre el espacio de los i-vectores ante la presencia de segmentos cortos y analizan el efecto que produce en la eficacia del reconocimiento al variar las duraciones de las muestras de entrenamiento y evaluación. Como resultado del estudio se demostró empíricamente que los sistemas G-PLDA y JFA no presentan grandes diferencias de comportamiento entre ellos pero son superiores en rendimiento a los sistemas que utilizan la distancia del coseno como medida de similitud y las técnicas de compensación sobre el espacio de los i-vectores, ante la variabilidad en la duración de los segmentos de habla.

Otro de los trabajos que analizan el efecto de la señales de corta duración en la verificación de locutores, son los propuestos por Sarkar y otros en [23], donde se analiza el efecto de las señales de corta duración en el entrenamiento de los parámetros[8] ante señales de corta duración que intervienen en la evaluación, utilizando PLDA sobre el espacio de los i-vectores y técnicas de normalización estándar y esférica. Para simular las señales cortas se basaron en la energía, seleccionando el número de tramas de mayor energía para obtener la duración requerida de los segmentos. Se concluye que cuando se realiza una evaluación con segmentos de corta duración o con segmentos de diferentes duraciones es necesario ajustar los datos que intervienen en el entrenamiento de los parámetros a las duraciones reales de los segmentos que intervienen en la evaluación.

Otro de los trabajos en esta área es la investigación realizada por Larcher y otros en [24], el cual analiza cómo afecta el contenido fonético en señales de corta duración en la eficacia del reconocimiento de locutor basados en la plataforma i-vector, usando los métodos de compensación WCCN y EFR [25]. Los autores se enfocaron en ajustar el contenido fonético de los segmentos de habla del entrenamiento (cliente) con relación a los segmentos que intervienen en la evaluación (prueba). Para el ajuste de los datos se utilizaron 3 conjuntos: contraseñas, comandos y la base Switchboard, los cuales presentan un promedio de duración de 0.75, 0.43 y 79.66 segundos por segmento, respectivamente. Los conjuntos de contraseñas y comandos fueron extraídos de la base de datos RSR2015, base de voces creada específicamente para la verificación de locutores dependientes del contexto se utilizaron además otras frases de la base Switchboard [26], para el entrenamiento del modelo UBM y de las matrices de variabilidad total y de compensación de sesión.

Con el objetivo de evaluar el efecto causado por la poca información fonética existente en las señales de corta duración en la verificación de locutor, se experimentó utilizando diferentes informaciones fonéticas en los segmentos de habla, para los clientes y las pruebas utilizadas en la evaluación, en comparación con la información fonética utilizada en los entrenamientos de los modelos y matrices.

Los segmentos que intervienen en las evaluaciones (cliente-prueba) se distribuyeron de forma tal que se pudiera evaluar la utilización de diferentes y similares contenidos fonéticos. Los autores

[8] Matriz de variabilidad total, Matriz que representa el espacio de la información del locutor y Matriz de covarianza intra-clase

concluyen que la información característica del locutor, que se encuentra intrínseca en el i-vector; que brindan los fonemas, puede utilizarse para mejorar los resultados en la evaluación con señales de corta duración.

Otro de los trabajos donde se analiza el efecto de la información que brinda el contenido fonético para el reconocimiento de locutores independiente del contenido con señales cortas es el presentado por Hasan et al. en [27]. En este trabajo los autores demuestran que con el truncado de las señales, disminuye exponencialmente el número de fonemas únicos, que son los que contienen mayor información característica del locutor, y reconocen que la cantidad de fonemas únicos a utilizar en entrenamiento y prueba requiere aún de una investigación más profunda. Realizan una comparación entre la media y la varianza de los i-vectores de un locutor extraídos de señales truncadas, respecto al i-vector extraído de la señal de mayor duración.

Partiendo de los análisis de la cantidad de fonemas únicos presentes en la señales de corta duración se proponen 3 variantes para compensar la variabilidad en la duración:

1 Realizar un entrenamiento del modelo PLDA con gran variabilidad en la duración de las señales utilizadas.
2 Utilizar el logaritmo de las duración de las señales como una función de la medida de calidad para la calibración de la puntuación.
3 Realizar un entrenamiento del modelo PLDA usando i-vectores extraídos de señales cortas con información fonética controlada, la cual sería la utilización de fonemas similares a los utilizados en las evaluaciones; obtenidas en el laboratorio.

En [28], Kanagasundaram y otros realizan experimentaciones sobre el funcionamiento de los algoritmos HT-PLDA y G-PLDA [14], ante la variabilidad de la duración de las señales de entrenamiento, en la evaluación y en la normalización de puntuaciones. Este mismo autor y otros proponen en [29] una nueva técnica para compensar la variabilidad de sesión con señales de corta duración sobre el espacio de los i-vectores. Se propone una modificación en el cálculo de la covarianza entre clases para el método LDA; mediante la siguiente expresión:

$$S_b = \alpha_l S_b^l + \alpha_c S_b^c, \tag{12}$$

donde α_l y α_c son pesos para señales de larga y corta duración respectivamente, y S_b^l y S_b^c son las estimaciones de la dispersión entre las clases para señales de larga y corta duración respectivamente.

Los autores plantean que el uso de señales de corta duración no afecta la variabilidad entre clases y si afecta la variabilidad intra-clase debido a las grandes variaciones en el contenido fonético. También concluyen que la utilización de la información que brinda variabilidad intra-clase no representa un beneficio para la eficacia de los algoritmos de reconocimiento de locutor, solo afectaría su rendimiento.

Otro de los trabajos donde se introduce una nueva estrategia para compensar la variabilidad de la duración es el propuesto por Hautamäki y otros en [30]. Los autores sustituyen la estimación de estadísticas de Baum-Welch por la estimación *minimax* [31] para estimar las estadísticas suficientes de primer orden y robustecer la extracción de los i-vectores, obteniendo un mejora de un 2 por ciento respecto a la línea base.

Como conclusión se tiene que la variabilidad de la duración entre las señales de entrenamiento y prueba representan un problema no resuelto en la verificación de locutores. La mayoría de los trabajos expuestos atacan esta problemática mediante el ajuste de los datos de entrenamiento en dependencia de los datos que intervienen en la prueba, lo que no es una solución factible

en aplicaciones reales, pues se desconoce que condiciones presentan las señales de prueba. En algunos trabajos como [21], [22] y [29], no evalúan los algoritmos propuestos con segmentos de habla de más de 1 minuto de duración, por lo que no pueden comparar la eficacia obtenida por los algoritmos propuesto en presencia de muestras en óptimas condiciones. Y otros que sí realizan la comparación como [23] no alcanzan la eficacia que presentan los algoritmos de reconocimiento de locutor actuales. Por último se observa que cuando se ajusta un algoritmo para enfrentar la problemática de las diferencias en la duración de las señales, se afecta la eficacia cuando se evalúa con segmentos de habla que no presentan esa variabilidad.

3.2. Variabilidad por el tipo e intensidad del ruido en las señales

Todas las señales de audio, incluida la voz, de alguna forma están expuestas al deterioro de su calidad, dado que en cualquiera de las etapas por las que puede pasar, emisión, propagación, captura, transmisión, almacenamiento y reproducción se puede introducir ruido. Se denomina ruido a toda señal no deseada que se mezcla con la señal deseada, en este caso, la voz. Las señales de voz son afectadas por diferentes tipos de ruido, como se describen en [32]: el ruido ambiental, la distorsión propia del teléfono, los ruidos propios del canal por donde se transmite la voz, los ruidos de cuantificación y codificación, entre otros.

Donde el ruido ambiental esta constituido por todos los sonidos que estén en el entorno donde se realiza la propagación y captura de la señal de voz, así como, el eco y la reverberación provocados por la acústica que presenta el área donde se propaga y captura el sonido. Ambos ruidos se observan como una perturbación aditiva y no correlacionada con la señal de voz, a diferencia de los ruidos por efecto del canal, los cuales se combinan con la señal de forma convolucional.

El ruido introducido por la distorsión del teléfono provoca efectuaciones de la señal de voz al saturarse el micrófono o el amplificador debido, por ejemplo, a que la señal de voz sobrepasa el rango de amplitud soportado por dichos dispositivos, esto puede ocurrir al hablar demasiado alto o al gritar. Los ruidos de cuantificación son las distorsiones introducidas en el proceso de conversión analógico-digital.

El reconocimiento de locutores no está ajeno a esta problemática dado que la eficacia de reconocimiento puede ser disminuida por los efectos de la variabilidad se sesión debido al ruido existente en las señales de voz [33], teniendo en cuenta que se provoca una degradación en la calidad de la señal de voz.

En el año 2012 se realizó la ultima evaluación NIST de los sistemas de reconocimiento de locutores, la cual presentó por primera vez en su data segmentos, que intervienen en la evaluación; con variabilidad por ruido y por duración. Varias empresas y equipos de investigación se unieron para presentar propuestas de reconocimiento de locutores aplicando variantes robustas ante la variabilidad de sesión por el ruido enfrentando esta problemática de forma simultánea, desde diferentes espacios [17,18]:

a. Preprocesamiento de las señales

   Con el objetivo de reducir el efecto ocasionado por el ruido, se utilizaron 3 métodos de filtrado: el filtrado de wiener de dos etapas [34], la sustracción espectral [35] y el método $RASTA_{LP}$ [36]. El filtrado de Wiener minimiza el error medio cuadrático entre la señal de voz limpia y la ruidosa, la sustracción espectral se enfoca en cancelar el ruido aditivo a partir de la señal de voz ruidosa, para esto se estima el ruido que presenta la señal y luego se le substrae para

obtener un estimado de la señal original; el método $RASTA_{LP}$ es un filtro paso bajo que se aplica a la señal ruidosa para reducir el efecto del ruido y la reverberación.

b. Extracción de rasgos
En estos sistemas se utilizaron varios tipos de rasgos de bajo nivel como los de Predicción Lineal Perceptual (PLP) [37], los Coeficientes Cepstrales en Frecuencia Mel (MFCC) [38], los Coeficientes Cepstrales de Predicción Lineal (LPCC) [39], así como los rasgos prosódicos [40] También se utilizaron nuevos rasgos diseñados específicamente para robustecer los sistemas de reconocimiento de locutores ante la variabilidad de sesión debido al ruido [17], como el cepstrum de la Modulación de la duración media, ($MDMC$) [41], los Coeficientes Cepstrales con Normalización de la Potencia, ($PNCC$) [42] y los Coeficientes de la Envolvente de Hilbert, ($MHEC$) [43].
Otros rasgos utilizados han sido los Coeficientes Cepstrales con bancos de Filtros Rectangulares, ($RFCC$) [44], propuestos para enfrentar el efecto *Lombard* y los rasgos $MFCC - QCN - RASTA_{LP}$, que son rasgos $MFCC$, que se procesan mediante la Normalización Cepstral Cuantil (QCN) [44].

c. Compensación de variabilidad de sesión
Para la compensación de la variabilidad de sesión fueron utilizados los algoritmos WCCN, LDA y la combinación LDA-WCCN, con el objetivo de reducir la variabilidad entre clases e intra-clases. Y para reducir el efecto producido por el canal se utilizó la compensación con el algoritmo NAP. Todas las matrices de compensación obtenidas mediante estos algoritmos se entrenaron a partir de segmentos de habla con variabilidad provocada por el tipo y su intensidad.

d. Algoritmos de clasificación
Como métodos de clasificación se apoyaron en el Análisis de factores conjunto (JFA) [45] y el modelo $PLDA$ para el espacio i-vector, además utilizaron la la distancia del coseno como medida de similitud. Otro de los modelos utilizados que mejores resultados obtuvo fue $Anti - modelKL - SVM - NAP$ [46], el cual se basa en un modelo SVM con un kernel de divergencia $Kullback - Leibler$, con compensación del canal mediante la proyección de atributos no deseados (NAP), $KL - SVM - NAP$ [47]; y con la adición de un anti-modelo.

e. Entrenamiento multicondición
Para los entrenamientos multicondición de los modelos, las matrices de variabilidad total y de proyección utilizadas en la compensación de la variabilidad de sesión, se utilizó la combinación de segmentos limpios con segmentos contaminados con ruido. Para esto se contaminaron los segmentos de habla de la data de entrenamiento de forma artificial adicionándole ruido HVAC(calefacción, ventilación y aire acondicionado), con 2 niveles de SNR, 6db y 15db, obtenidos de la base de datos de sonidos en: www.freesound.org. Para la contaminación de los segmentos se utilizó la herramienta FaNT (http://dnt.kr.hsnr.de/download.html).

f. Fusión de clasificadores
La mayoría de las propuestas utilizaron la fusión como una herramienta indispensable para elevar la robustez de la verificación de locutores, mediante la combinación de sistemas que trabajaban con diferentes tipos de rasgos, diferentes algoritmos de clasificación o diferentes modelos. La fusión se llevó a cabo sobre el espacio de las puntuaciones mediante fusiones lineales o fusiones lineales pesadas utilizando el tiempo de duración y el valor de SNR como medidas de calidad de los segmentos de habla. En la propuesta de [48] realizaron la fusión de i-vectores obtenidos con diferentes rasgos calculando un vector promedio de todos estos i-vectores.

Tabla 2. Principales resultados en la evaluación NIST-SRE 2012 en base al minDCF, bajo las condiciones segmentos de prueba telefónicos sin ruido (CC2), segmentos de prueba telefónicos con adición de ruido (CC4) y segmentos de prueba telefónicos con ruido real (CC5).

Condiciones/Lugares	1	2	3
CC2	ABC	SRI	SHDragon
CC4	SRI	ABC	IIR
CC5	ABC	SRI	SHDragon

En la tabla 2, se exponen las tres mejores propuestas en tres de los experimentos de la evaluación [49], medidos con la función del costo de detección [18]. Donde de 9 instituciones, tres compañías, tres universidades y tres laboratorios; solo 4 propuestas obtuvieron los mejores resultados:

- SHDragon: Shanghai Dragon Voice, China
- ABC: Laboratorios Agnitio, Universidad de Berno en la República Checa y Laboratorio CRIM en Canadá
- SRI: SRI International, USA
- IIR: Instituto para investigaciones de las infocomunicaciones, Singapore.

Del análisis de los resultados de las evaluaciones NIST 2012, se concluye, que para enfrentar el efecto del ruido en el reconocimiento de locutores es de gran ayuda el uso de entrenamiento multicondición, que no es más que ajustar los datos de entrenamiento en función del dato que se vaya a evaluar. No obstante, se aprecia en los resultados, que son mas eficaces los algoritmos de reconocimiento de locutor, donde intervienen los nuevos rasgos y métodos de compensación ajustados al tipo de variabilidad y el entrenamiento multicondición con los tipos de ruido a enfrentar, cuando los segmentos de prueba están contaminados con ruido que cuando están limpios, lo que demuestra que existe un bache cuando se tiene un sistema ajustado para enfrentar los espacios ruidosos y se analizan segmentos de habla limpios, dado que disminuye la eficacia del mismo.

3.3. Variabilidad debida a la reverberación en las expresiones de voz

Los sonidos pueden recibirse por 2 vías fundamentales: el sonido directo o el sonido reflejado en algún obstáculo. En el caso de que las ondas de sonido reflejadas sufran un retardo en el tiempo mayor a los 100 milisegundos y sean reconocidas por el ser humano como un segundo sonido se les denomina eco, pero si el retardo es menor y se perciben como una adición que modifica el sonido original se denomina reverberación. Es muy común observar la reverberación, debido al efecto causado por la reflexión de los sonidos en locales cerrados, como oficinas, salas de reuniones, cabinas telefónicas, etc. y se representa como una ligera permanencia del sonido una vez que la fuente original ha dejado de emitirlo.

La reverberación es un parámetro que cuantifica la acústica de un local, la cual se analiza a partir del tiempo de reverberación y se denomina **intensidad de la reverberación**.

La variabilidad por la intensidad de la reverberación en el habla es un aspecto que afecta la robustez de los sistemas de reconocimiento de locutores, siendo considerado como un ruido aditivo a la señal. Esta variabilidad es enfrentada en la actualidad mediante la utilización de

nuevos rasgos, las normalizaciones de los mismos y el entrenamiento de los modelos, matrices y clasificadores a partir de datos ajustados para enfrentar esta variabilidad.

Para la compensación de la reverberación en el espacio de los rasgos se proponen rasgos de Predicción Lineal en el dominio de la Frecuencia (FDLP) [50] para compensar esta variabilidad. Los autores en [51] comparan el funcionamiento de los sistemas JFA, PLDA y KPLS [52] usando 5 tipos de rasgos distintos en presencia de variabilidad en los segmentos de habla por la reverberación. Comprobando que los sistemas son más robustos ante la variabilidad por reverberación cuando utilizan los rasgos FDLP-lineal y FDLP-mel. En comparación con el tipo de clasificador utilizado obtuvieron los mejores resultados con el PLDA.

Una representación cortical propuesta en [53] es aplicada al reconocimiento del locutor en [54].

Zhou y otros en [55] realizan una comparación entre los rasgos MFCC y LPCC en el reconocimiento de locutores con segmentos de habla con varias intensidades de reverberación, comprobando que los LPCC son más robustos que los MFCC.

Hasan y otros en [56], utilizan rasgos desarrollados específicamente para enfrentar el ruido y la reverberación en los segmentos de habla, como, los MHEC. Otro aspecto a destacar en este trabajo es la utilización de técnicas de normalización de rasgos como QCN y $RASTA_{LP}$ para reducir el efecto de la reverberación en los segmentos de habla.

García-Romero y otros en [57] utilizan el entrenamiento multicondición en un conjunto de subsistemas independientes, todos basados en el clasificador PLDA gaussiano; ajustados a condiciones específicas para enfrentar la reverberación, analizada como un ruido aditivo. Se proponen 3 alternativas de entrenamiento multicondición para obtener los hiperparámetros correspondientes al modelo PLDA para cada condición, $\{\mu, \phi, \epsilon\}_1^K$, descrito en la sección 2.2, las cuales consisten en modelar los clasificadores independientes, atados o agrupados. El modelo independiente consiste en modelar cada condición de forma independiente al resto, obteniendo los parámetros del modelo, expuestos en 2.2; como se describe en [14]. El modelo atado, consiste en la generación de i-vectores utilizando la misma variable de identidad latente descrita en 2.2 para las diferentes condiciones, obteniéndose un modelo con la variable de identidad común mientras que el resto de los parámetros del mismo presentan un valor independiente por cada condición utilizada en el entrenamiento. El modelo agrupado consiste en agrupar los datos de entrenamiento de todas las condiciones y estimar los parámetros del modelo. Finalmente, teniendo en cuenta que en los entrenamientos multicondición los modelos de los clientes están representados por varios i-vectores, se obtiene la puntuación final de la comparación mediante la combinación pesada de la puntuación de cada subsistema, en este caso los autores consideraron el peso equiprobable. Se pudo comprobar que los sistemas que utilizan la modelación mediante agrupación y atado de los datos obtienen una mejor eficacia ante la variabilidad por la reverberación ya que se logra equilibrar los datos de entrenamiento.

Como resumen se observa que los trabajos que enfrentan la variabilidad del habla por la reverberación, han explorado dos áreas fundamentales de los sistemas de reconocimiento de locutores. Una de las áreas ha sido la extracción de rasgos y su normalización, donde se mejoran los resultados de los sistemas con los rasgos tradicionales, MFCC y LFCC, mediante la utilización de rasgos desarrollados específicamente para enfrentar esta variabilidad, como los FDLP-lineal, FDLP-mel y MHEC; y la técnica de normalización de rasgos QCN, la cual es más robusta que otras técnicas de normalización en el reconocimiento del habla como se muestra en [58]. La otra área explorada ha sido el entrenamiento multicondición, ya sea en el modelo o en la representación del cliente mediante el uso de diversos i-vectores bajo diferentes condiciones de reverberación; observándose el aumento de la robustez de los sistemas ante la variabilidad por la reverberación.

4. Resultados experimentales

Del estudio realizado en la sección 3.1, se observa que la variabilidad en la duración los segmentos de habla en el entrenamiento y la prueba, constituye una problemática a resolver, en reconocimiento de locutores los resultados obtenidos mediante el ajuste de los datos, no logran alcanzar la eficacia obtenida en presencia de señales de más de 1 minuto de duración. Debido a que los modelos se ajustan para enfrentar una variabilidad específica y, en casos reales, se desconoce la variabilidad existente en las muestras de voz o es muy distinta a la ajustada, conlleva a que los métodos no brinden la mejor eficacia.

Para poder conocer el comportamiento de los diversos métodos propuestos, ante la variabilidad de la duración, nos propusimos implementar un experimento utilizando las bases con que contamos, Nist 2004 y 2005 para el entrenamiento y Nist 2008 para la evaluación.

Inicialmente se comprobó la eficacia, medida utilizando el EER [18] y mínimo DCF [18], de los métodos del estado del arte descritos en la sección 2, ante la variabilidad de la duración. Utilizando los métodos IV y $PLDA$ gaussiano, se aplica la compensación de la variabilidad inter-clases e intra-clases, algoritmo LDA descrito en la sección 2.1.2. Para la clasificación se utiliza la distancia del coseno y el algoritmo $PLDA$, en el experimento se utilizaron diferentes configuraciones respecto a la duración de los segmentos de habla en el cliente y la prueba:

Tabla 3. Duración de los segmentos de habla en la evaluación (cliente-prueba).

Experimentos	**Cliente**	**Prueba**
1	5 segundos	5 segundos
2	10 segundos	10 segundos
3	15 segundos	15 segundos
4	20 segundos	20 segundos
5	más de 60 segundos	más de 60 segundos

Tabla 4. Resultados de la verificación de locutores, EER(minDCF); ante la variabilidad de la duración (segundos) de los segmentos que intervienen en la evaluación.

	Full-Full	20-20	15-15	10-10	5-5
IV	3.70(0,022)	9,56(0,039)	11,1(0,050)	**13,9(0,064)**	22,1(0,084)
PLDA	**2.96(0,022)**	**7,97(0,039)**	**10,0(0,048)**	14,1(0,061)	**21,4(0,086)**

En la tabla 4 se aprecia la mayor robustez del método PLDA ante señales cortas en comparación con el método IV y se confirma que la eficacia decrece mientras el tiempo de duración de los segmentos disminuye, por lo que podemos decir que la variabilidad de sesión está dada por la corta duración de ambas muestras en comparación con las muestras de voz utilizadas en el entrenamiento del modelo UBM, matriz T y modelo PLDA y no en la diferencia de duración entre ambas muestras de la evaluación.

Ante este comportamiento ineficaz de los métodos, debido a la corta duración de las muestras de entrenamiento y prueba, se propuso evaluar el efecto de la variabilidad por la duración en los i-vectores y proponer métodos para reducirla a partir de la compensación de la variabilidad intra locutor e inter locutor.

4.1. Análisis del efecto en la variabilidad producido por los segmentos de corta duración en el espacio i-vector

Para mostrar el efecto producido por la variabilidad en la duración de los segmentos de habla, en el espacio i-vector, se transformó el espacio i-vector, utilizando la técnica PCA, para obtener los componentes principales, mostrando los dos primeros en un espacio bidimensional. Se aplicó dicha técnica a i-vectores extraídos de segmentos de habla con diferentes duraciones (3 segundos, 5 segundos, 10 segundos, 15 segundos, 20 segundos y toda la duración) de 10 clases distintas, donde cada clase corresponde a un locutor en específico.

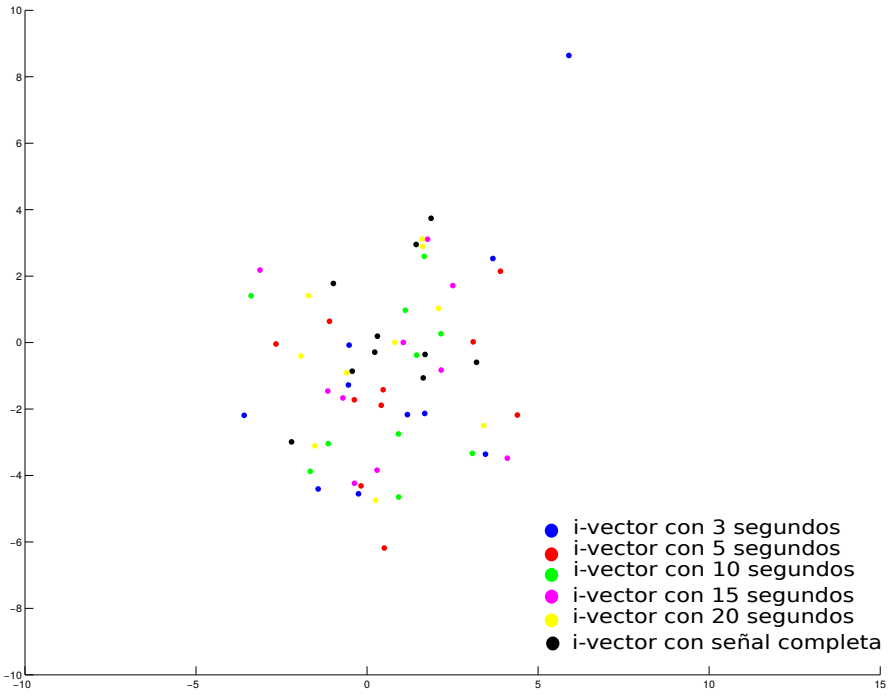

Fig. 3. Representación de las dos primeras componentes principales de los i-vectores sin compensar la variabilidad de sesión, para 10 locutores y diferentes duraciones en los segmentos de habla.

Como se observa en la figura 3, es imposible detectar y agrupar los i-vectores correspondientes a cada clase debido a la dispersión y solapamiento que presentan en el espacio de las dos primeras componentes principales.

Se procedió entonces a compensar la variabilidad de sesión de los i-vectores utilizando el algoritmo LDA descrito en la sección 2.1.2, utilizando la duración completa de los segmentos de habla que intervienen en el entrenamiento de la matriz de proyección `LDA`. Aún se observa la compensación LDA logra agrupar en clases los i-vectores de cada locutor, reduciendo apreciablemente el efecto de la variabilidad intra-clase debido a las diferentes duraciones de las muestras, aunque no de forma óptima, ya que no logra reducir a un mínimo la separación existente entre los i-vectores de una misma clase, para las diferentes duraciones.

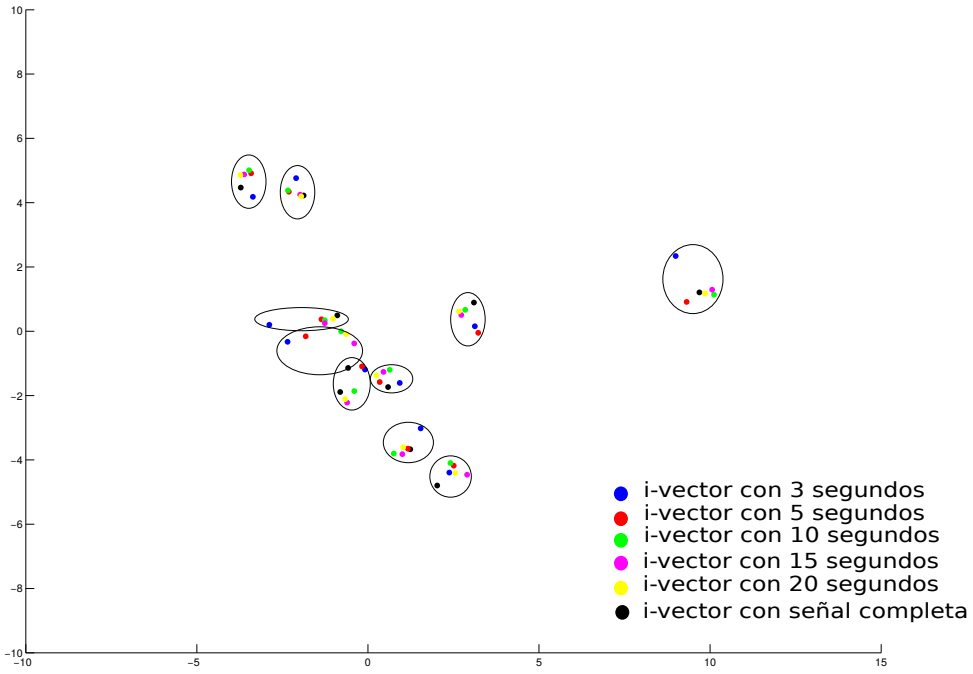

Fig. 4. Representación de las dos primeras componentes principales de los i-vectores compensados ante la variabilidad de sesión, para 10 locutores y diferentes duraciones del segmento de habla.

En el trabajo de kanagasundaram [29]; se propone una nueva variante de compensación de sesión (SUN-LDA) para enfrentar los problemas causados por las cortas duraciones en las expresiones de habla utilizadas en el entrenamiento del cliente y la prueba, mediante la incorporación, en la estimación de la covarianza del algoritmo LDA, de la información de la variabilidad entre clases de las señales. Los autores plantean que en la estimación de la covarianza intra-clases no era conveniente utilizar la información que proveen la variabilidad de las señales cortas. Teniendo en cuenta que uno de los aspectos principales de esta compensación de sesión, mediante el algoritmo LDA, es minimizar la variabilidad intra-clase, *se hace necesario obtener la mayor cantidad de representaciones de un mismo locutor con distintas duraciones, para así poder minimizar aún más la variabilidad intra-clase.*

Se comprobó el funcionamiento de la variante SUN-LDA propuesta en [29] y se obtuvo un EER de 14,57 % con segmentos de habla de 10 segundos para el cliente y la prueba. Se modificó la misma con una nueva propuesta consistente en incorporar la información de la variabilidad intra-clase de señales cortas para estimar la covarianza (SUN-LDA-2), obteniéndose un EER de 14,08 %, con los mismos segmentos de habla. Se pudo comprobar que la información de la variabilidad intra-clase contenida en las señales cortas, lejos de afectar, benefician los resultados del reconocimiento de locutores en presencia de las señales cortas, comparación con los resultados obtenidos en [29].

Se propuso adicionalmente un método, denominado IV-CVD; para minimizar la variabilidad entre los i-vectores de una misma clase que se ven afectados por diferentes duraciones de los segmentos del habla. Para esto se estimó las variabilidades entre clases e intra-clase utilizando

i-vectores extraídos de segmentos de habla con variabilidad debido a la duración y compensados para reducir la variabilidad de sesión, mediante el algoritmo LDA.

Se obtuvieron 3285 i-vectores de 262 impostores (de las bases de datos Nist-SRE 04-05), con diferentes duraciones de habla (3, 5, 10, 15, 20 segundos y señal completa), y se llevo a cabo la compensación ante la variabilidad de sesión con una matriz de proyección obtenida con el algoritmo LDA, .

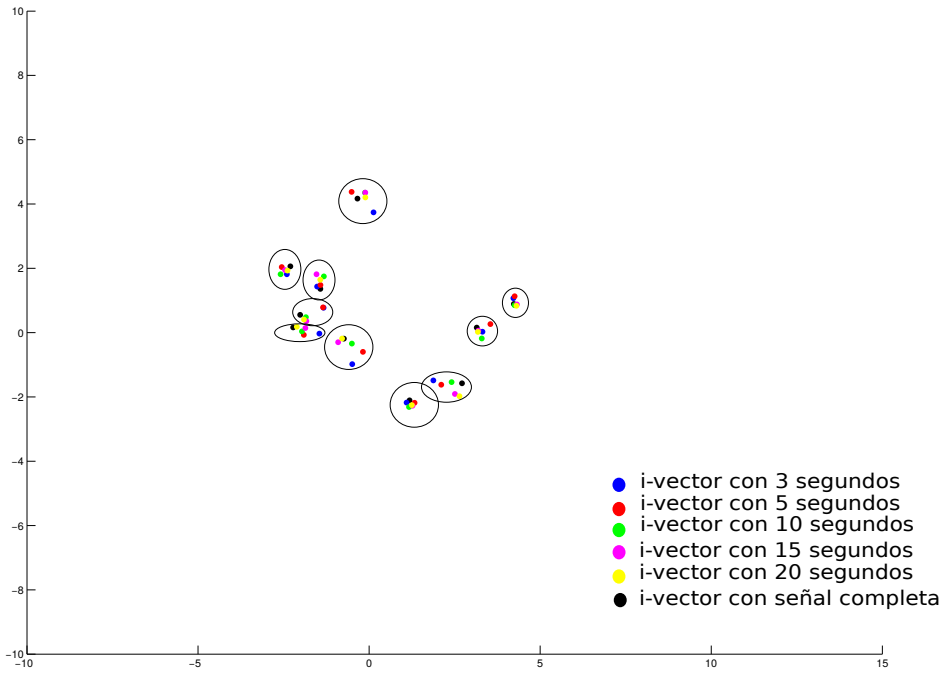

Fig. 5. Representación de las dos primeras componentes principales de los i-vectores compensados ante la variabilidad de sesión y variabilidad en la duración de los segmentos de habla para 10 locutores.

En la figura 5 se observa que se logró minimizar aún más la variabilidad intra-clase, en comparación con la figura 4.

Se comprobó el efecto que causa la compensación de la variabilidad de sesión debido a la duración de los segmentos de habla con el algoritmo propuesto, IV-CVD, mediante el cálculo de la varianza que existe intra-clase e inter-clases entre los 10 locutores.

En la tabla 5 se observa una disminución de las varianzas intra-clase de los 10 locutores cuando se utiliza el algoritmo `IV-CVD` en comparación con el `IV`, la cual representa una reducción del 50.20 % entre los 10 locutores analizados. La variabilidad inter-clases no se comportó de igual forma que la variabilidad intra-clase, siendo esto aún un problema a resolver.

Por último en la tabla 6 se muestran los resultados utilizando dos clasificadores, la distancia del coseno y PLDA, sin perder de vista que el PLDA presenta una gran pérdida de eficiencia a costa de un incremento sustancial de la eficacia.

En estas experimentaciones se realizó un entrenamiento multicondición del modelo PLDA y la matriz de compensación de variabilidad de sesión LDA, el cual consiste en utilizar muestras

Tabla 5. Relación de la varianza intra locutor e inter locutor en el espacio de los i-vectores entre 10 locutores.

LOC	IV		IV-CVD		% Reducción	% Reducción
	Inter-clases	Intra-clase	Inter-clases	**Intra-clase**	Inter-clases	Intra-clase
1	**88.19**	4.51	38.70	**2.59**	56.11	42.57
2	**79.41**	8.29	34.69	**4.08**	56.31	50.78
3	**87.83**	2.47	35.19	**1.23**	59.93	50.20
4	**115.0**	11.0	44.92	**4.98**	60.93	54.72
5	**84.37**	4.12	34.48	**1.75**	59.13	57.52
6	**83.78**	2.84	34.13	**1.53**	59.26	46.12
7	**84.19**	8.95	34.67	**4.70**	58.81	47.48
8	**82.59**	1.92	33.59	**1.09**	59.32	43.22
9	**85.34**	5.74	34.71	**2.96**	59.32	48.43
10	**86.47**	5.54	35.12	**2.16**	59.38	61.01

de voz con diferentes duraciones para estimar dichos parámetros. Las muestras de voz con corta duración se obtuvieron mediante el truncado de las muestras correspondientes a las bases de datos Nist-04-05, obteniendo conjuntos de muestras de voz de 3, 5, 10, 15 y 20 segundos, además de las muestras originales sin truncar (muestras completas).

Se evaluaron el método SUN-LDA, propuesto en [29]; nuestra variante utilizando la información intra-clase de señales cortas y la nueva variante propuesta IV-CVD, todas ellas enfrentan la variabilidad debido a la corta duración de los segmentos de habla en el cliente y prueba.

Tabla 6. Resultados con entrenamiento multi-condiciones respecto a la duración, en base al EER.

	full-full	20-20	15-15	10-10	5-5	Promedio
IV	5,50	8,53	10,9	**13,2**	20,7	11.7
SUN-LDA	**3,93**	9,56	10,78	14,2	22,5	12.2
SUN-LDA-2	4,78	9,08	10,7	14,1	22,3	12,1
IV-CVD	4,55	**8,42**	**10,4**	**13,2**	**20,2**	**11.3**
PLDA	**3,87**	8,22	9,33	12,5	18,2	10.4
SUN-LDA+PLDA	4.09	**7,86**	9,11	11,8	17.9	10.15
SUN-LDA-2+PLDA	4,27	7,97	**8,88**	11,8	18,81	10,34
IV-CVD+PLDA	4,32	7,97	8,90	**10,8**	**17,1**	**9,81**

Comparando los resultados de la tabla 6 con los mostrados en la tabla 4, se confirma que es necesario ajustar los clasificadores a las condiciones a que se enfrentarán. Los sistemas de reconocimiento de locutores basados en el clasificador PLDA y ajustados para enfrentar la variabilidad debido a la corta duración de los segmentos de habla en el cliente y la prueba, alcanzan una mayor eficacia que los sistemas IV.

El nuevo método propuesto, IV-CVD; obtiene menores errores para casi todas las combinaciones de duración de los segmentos del cliente y la prueba, con un más bajo error promedio, tanto en IV como PLDA.

Para analizar la robustez de los métodos evaluados, nos apoyamos, en el comportamiento de los EER de cada experimento desde los segmentos de habla de más corta duración hasta los experimentos donde se utilizan los segmentos completos. Como se observa en la figura 6,a menor

pendiente de la curva que une los puntos de EER para cada método, mayor robustez. También podemos decir:

- Los métodos con clasificador `PLDA` son, en general, más robustos que los métodos `IV`.
- Entre los métodos `IV`, el método más robusto es el método propuesto `IV-CVD`, seguido del propio `IV`.
- Para los métodos con clasificador `PLDA`, el método de compensación de variabilidad de sesión debido a la duración más robusto, es el método propuesto `IV-CVD`.

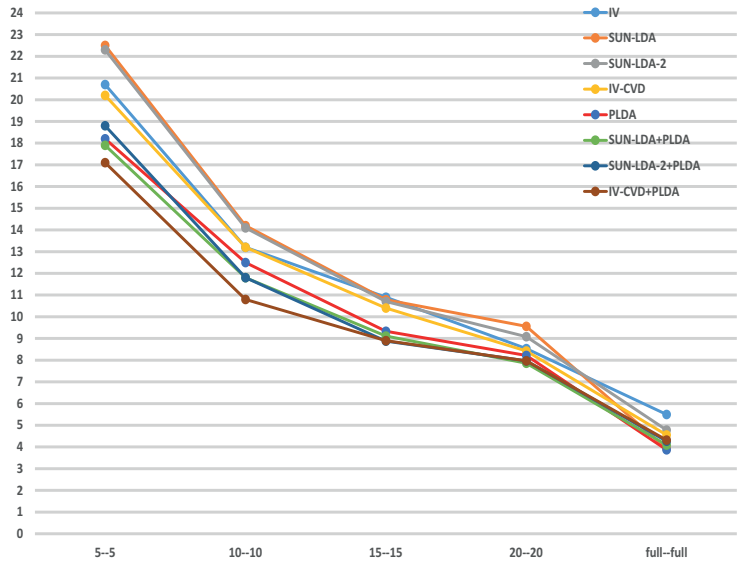

Fig. 6. Diferencia entre los errores (EER) de los experimentos con señales cortas en el cliente y en la prueba y el error en el experimento con la señal completa.

5. Conclusiones Generales

A partir del estudio realizado de los métodos utilizados en el reconocimiento de locutores para compensar la variabilidad en las señales, provocadas por cambios en su duración y en ruido; podemos asegurar que, cuando las señales son adquiridas en entornos no controlados (reales), se afecta la eficacia de los métodos utilizados en el reconocimiento. Teniendo en cuenta que se desconocen las condiciones y la calidad de las señales a procesar, es necesario desarrollar nuevos algoritmos capaces de aumentar la robustez y la eficacia frente a las condiciones reales en que se adquieren muchas de las expresiones de voz.

Una inmensa mayoría de los métodos que enfrentan esta problemática, se apoyan principalmente en la utilización de muchos datos de entrenamiento obtenidos en diferentes condiciones(entrenamiento multicondición) para ajustar, entre otros aspectos, los parámetros de los métodos de compensación. Esto provoca que se vea afectada la eficacia del reconocedor, como se observa por ejemplo, cuando debido a la variabilidad de la duración, los métodos se ajustan para enfrentar ese tipo de variabilidad y las condiciones reales en que se presenta dicha variabilidad son otras, como se observa en la configuración $full$-$full$ al comparar la tabla 4 con la tabla 6.

Al comparar los resultados experimentales en las primeras columnas de las tablas 4 y 6, con las muestras de cliente y prueba de mayor duración, full-full, se observa un aumento del error en los resultados cuando se aplica el entrenamiento multicondición, lo cual se debe al ajuste de los modelos de los locutores para enfrentar la variabilidad en la duración de las señales, pues en este experimento los segmentos de habla utilizados presentan duraciones óptimas (cuando la duración de los mismos exceden los 60 segundos). **Lo cual representa un problema en aplicaciones reales de verificación o identificación de locutores**. De ahí que podemos decir que aún existen obstáculos que afectan los sistemas de reconocimiento de locutores en ambientes reales.

Recomendaciones

1 Profundizar en la utilización del la normalización del espectro de la varianza [59] y de las gráficas espectrales de varianza [59], para compensar los efectos de corta duración y ruido.
2 Aplicar los métodos propuestos de compensación de variabilidad utilizando LDA ante ruido y reverberación.
3 Proponen un vía de solución al problema forense, a partir de una fusión de clasificadores ajustados a diferentes fuentes de variabilidad, cuyos resultados se compensen con la información de las fuentes de variabilidad que se hayan detectado en las muestras del cliente y la prueba, concretamente la duración, el nivel de ruido y la reverberación.

Referencias bibliográficas

1. Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., Reynolds, D.A.: A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process. **2004** (2004) 430–451
2. Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: From features to supervectors. Speech Communication **52**(1) (2010) 12 – 40
3. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing **10**(1–3) (2000) 19 – 41
4. Campbell, W., Campbell, J., Reynolds, D., Singer, E., Torres-Carrasquillo, P.: Support vector machines for speaker and language recognition. Computer Speech and Language **20**(2–3) (2006) 210 – 229
5. Campbell, W.M., Sturim, D.E., Reynolds, D.A.: Support vector machines using gmm supervectors for speaker verification. Signal Processing Letters, IEEE **13**(5) (2006) 308–311
6. Reynolds, D.A.: Channel robust speaker verification via feature mapping. In: Proceedings.(ICASSP'03). Volume 2. (2003)
7. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A study of interspeaker variability in speaker verification. Audio, Speech, and Language Processing, **16**(5) (2008) 980–988
8. Solomonoff, A., Campbell, W., Boardman, I.: Advances in channel compensation for svm speaker recognition. In: Proceedings.(ICASSP'05). Volume 1. (2005)
9. Reynolds, D.: The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus. In: ICASSP-96. Conference Proceedings. Volume 1. (1996)
10. Auckenthaler, R., Carey, M., Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digital Signal Processing **10**(1–3) (2000) 42 – 54
11. Aronowitz, H., Burshtein, D.: Efficient speaker identification and retrieval. In: Ninth European Conference on Speech Communication and Technology. (2005)
12. Kenny, P., Boulianne, G., Dumouchel, P.: Eigenvoice modeling with sparse training data. Speech and Audio Processing **13**(3) (2005) 345–354
13. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing, **19**(4) (2011) 788–798

14. Kenny, P.: Bayesian speaker verification with heavy tailed priors. In: Speaker and Language Recognition Workshop (Odyssey). (2010)
15. Lachenbruch, P.A.: Discriminant analysis. Wiley Online Library (1975)
16. Prince, S., Elder, J.: Probabilistic linear discriminant analysis for inferences about identity. In: Computer Vision, ICCV 2007. (2007)
17. Scheffer, N., Ferrer, L., Lawson, A., Lei, Y., McLaren, M.: Recent developments in voice biometrics: Robustness and high accuracy. In: Technologies for Homeland Security (HST),. (2013) 447–452
18. Gonzalez-Rodriguez, J.: Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014). Loquens **1**(1) (2014)
19. Lee, K.A., Larcher, A., Thai, H., Ma, B., Li, H.: Joint application of speech and speaker recognition for automation and security in smart home. In: In Proceedings of the 12th Annual Conference of the International Speech Communication Association. (2011) 3317–3318
20. Ferrer, L., Bratt, H., Burget, L., Cernocky, H., Glembek, O., Graciarena, M., Lawson, A., Lei, Y., Matejka, P., Plchot, O., et al.: Promoting robustness for speaker modeling in the community: the prism evaluation set. In: Proceedings of NIST 2011 Workshop. (2011)
21. Mandasari, M.I., McLaren, M., van Leeuwen, D.A.: Evaluation of i-vector speaker recognition systems for forensic application. In: In Proceedings of the 12th Annual Conference of the International Speech Communication Association. (2011) 21–24
22. Kanagasundaram, A., Vogt, R., Dean, D.B., Sridharan, S., Mason, M.W.: I-vector based speaker recognition on short utterances. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association. (2011) 2341–2344
23. Sarkar, A.K., Matrouf, D., Bousquet, P.M., Bonastre, J.F.: Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In: In Proceedings of the 13th Annual Conference of the International Speech Communication Association. (2012)
24. Larcher, A., Bousquet, P., Lee, K.A., Matrouf, D., Li, H., Bonastre, J.F.: I-vectors in the context of phonetically-constrained short utterances for speaker verification. In: Acoustics, Speech and Signal Processing (ICASSP). (2012)
25. Bousquet, P.M., Matrouf, D., Bonastre, J.F.: Intersession compensation and scoring methods in the i-vectors space for speaker recognition. In: In Proceedings of the 12th Annual Conference of the International Speech Communication Association. (2011) 485–488
26. Lamel, L., Gauvain, J.L.: Speaker recognition with the switchboard corpus. In: Acoustics, Speech, and Signal Processing, ICASSP. Volume 2. (1997) 1067–1070
27. Hasan, T., Saeidi, R., Hansen, J.H., van Leeuwen, D.A.: Duration mismatch compensation for i-vector based speaker recognition systems. In: Acoustics, Speech and Signal Processing (ICASSP). (2013) 7663–7667
28. Kanagasundaram, A., Vogt, R.J., Dean, D.B., Sridharan, S.: Plda based speaker recognition on short utterances. In: The Speaker and Language Recognition Workshop (Odyssey 2012), ISCA (2012)
29. Kanagasundaram, A., Dean, D., Gonzalez-Dominguez, J., Sridharan, S., Ramos, D., Gonzalez-Rodriguez, J.: Improving short utterance based i-vector speaker recognition using source and utterance-duration normalization techniques. In: In Proceedings of the 14th Annual Conference of the International Speech Communication Association, International Speech Communication Association (ISCA) (2013) 2465–2469
30. Hautamäki, V., Cheng, Y.C., Rajan, P., Lee, C.H.: Minimax i-vector extractor for short duration speaker verification. In Bimbot, F., Cerisara, C., Fougeron, C., Gravier, G., Lamel, L., Pellegrino, F., Perrier, P., eds.: In Proceedings of the 14th Annual Conference of the International Speech Communication Association, ISCA (2013) 3708–3712
31. Merhav, N., Lee, C.H.: A minimax classification approach with application to robust speech recognition. Speech and Audio Processing, **1**(1) (1993) 90–100
32. Dayana Ribas González, J.R.C.: Métodos de compensación de ruido en el reconocimiento de locutores. Technical report, Centro de Aplicaciones de Tecnologías de Avanzada (2009)
33. Ming, J., Hazen, T.J., Glass, J.R., Reynolds, D.A.: Robust speaker recognition in noisy conditions. Audio, Speech, and Language Processing **15**(5) (2007) 1711–1723
34. Agarwal, A., , Agarwal, A., Cheng, Y.M.: Two-stage mel-warped wiener filter for robust speech recognition. In: in Proc. ASRU. (1999) 12–15
35. Davis, G.M.: Noise reduction in speech applications. Volume 7. CRC Press (2002)
36. Boril, H., Hansen, J.H.: Ut-scope: Towards lvcsr under lombard effect induced by varying types and levels of noisy background. In: Acoustics, Speech and Signal Processing (ICASSP). (2011) 4472–4475
37. Hermansky, H.: Perceptual linear predictive (plp) analysis of speech. the Journal of the Acoustical Society of America **87**(4) (1990) 1738–1752

38. Reynolds, D.A.: Experimental evaluation of features for robust speaker identification. Speech and Audio Processing, IEEE Transactions **2**(4) (1994) 639–643
39. Atal, B.S.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. the Journal of the Acoustical Society of America **55**(6) (2005) 1304–1312
40. Dehak, N., Dumouchel, P., Kenny, P.: Modeling prosodic features with joint factor analysis for speaker verification. Audio, Speech, and Language Processing, **15**(7) (2007) 2095–2103
41. Mitra, V., Franco, H., Graciarena, M., Mandal, A.: Normalized amplitude modulation features for large vocabulary noise-robust speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP),. (2012) 4117–4120
42. Kim, C., Stern, R.M.: Power-normalized cepstral coefficients (pncc) for robust speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP). (2012) 4101–4104
43. Sadjadi, S.O., Hansen, J.H.: Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions. In: Acoustics, Speech and Signal Processing (ICASSP). (2011) 5448–5451
44. Boril, H., Hansen, J.H.: Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments. Audio, Speech, and Language Processing, **18**(6) (2010) 1379–1393
45. Kenny, P.: Joint factor analysis of speaker and session variability: Theory and algorithms. CRIM, Montreal,(Report) CRIM-06/08-13 (2005)
46. Sun, H., Lee, K.A., Ma, B.: Anti-model kl-svm-nap system for nist sre 2012 evaluation. In: Acoustics, Speech and Signal Processing (ICASSP). (2013) 7688–7692
47. Campbell, W.M., Sturim, D.E., Reynolds, D.A., Solomonoff, A.: Svm based speaker verification using a gmm supervector kernel and nap variability compensation. In: Acoustics, Speech and Signal Processing, (ICASSP). Volume 1. (2006)
48. Ferrer, L., McLaren, M., Scheffer, N., Lei, Y., Graciarena, M., Mitra, V.: A noise-robust system for nist 2012 speaker recognition evaluation. (2013)
49. Greenberg, C.S., Stanford, V.M., Martin, A.F., Yadagiri, M., Doddington, G.R., Godfrey, J.J., Hernandez-Cordero, J.: The 2012 nist speaker recognition evaluation. (2013)
50. Ganapathy, S., Thomas, S., Hermansky, H.: Front-end for far-field speech recognition based on frequency domain linear prediction. In: In Proceedings of the 9th Annual Conference of the International Speech Communication Association. (2008) IDIAP-RR 08-17.
51. Garcia-Romero, D., Zhou, X., Zotkin, D., Srinivasan, B., Luo, Y., Ganapathy, S., Thomas, S., Nemala, S., Sivaram, G.S., Mirbagheri, M., et al.: The umd-jhu 2011 speaker recognition system. In: Acoustics, Speech and Signal Processing (ICASSP). (2012) 4229–4232
52. Srinivasan, B.V., Garcia-Romero, D., Zotkin, D.N., Duraiswami, R.: Kernel partial least squares for speaker recognition. In: In Proceedings of the 12th Annual Conference of the International Speech Communication Association. (2011) 493–496
53. Zotkin, D.N., Chi, T., Shamma, S.A., Duraiswami, R.: Neuromimetic sound representation for percept detection and manipulation. EURASIP Journal on Applied Signal Processing **9** (2005) 1350
54. Nemala, S.K., Zotkin, D.N., Duraiswami, R., Elhilali, M.: Biomimetic multi-resolution analysis for robust speaker recognition. EURASIP Journal on Audio, Speech, and Music Processing **2012**(1) (2012) 1–10
55. Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., Shamma, S.: Linear versus mel frequency cepstral coefficients for speaker recognition. In: Automatic Speech Recognition and Understanding (ASRU). (2011) 559–564
56. Hasan, T., Liu, G., Sadjadi, S.O., Shokouhi, N., Bořil, H., Ziaei, A., Misra, A., Godin, K., Hansen, J.: Utd-crss systems for 2012 nist speaker recognition evaluation. In: Proc. NIST SRE Workshop. (2012)
57. Garcia-Romero, D., Zhou, X., Espy-Wilson, C.Y.: Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition. In: Acoustics, Speech and Signal Processing (ICASSP). (2012) 4257–4260
58. Boril, H., Grézl, F., Hansen, J.H.: Front-end compensation methods for lvcsr under lombard effect. In: In Proceedings of the 12th Annual Conference of the International Speech Communication Association. (2011) 1257–1260
59. Bousquet, P.M., Larcher, A., Matrouf, D., Bonastre, J.F., Plchot, O.: Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. In: Proc. Odyssey. (2012)

RT_072, marzo 2015
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba