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

Reconocimiento del locutor dependiente del texto con modelos acústicos del habla

Ing. Ivis Rodés Alfonso,
Dr. C. José Ramón Calvo de Lara

RT_009 **Noviembre 2009**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
Ciudad de 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

Reconocimiento del locutor dependiente del texto con modelos acústicos del habla

Ing. Ivis Rodés Alfonso,
Dr. C. José Ramón Calvo de Lara

RT_009 **Noviembre 2009**

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

Reconocimiento del locutor dependiente del texto con modelos acústicos del habla

Ing. Ivis Rodés Alfonso, Dr. C. José Ramón Calvo de Lara

Centro de Aplicaciones de Tecnología de Avanzada, 7a #21812 e/ 218 y 222, Siboney, Playa, Habana, Cuba
irodes@cenatav.co.cu

RT__009 CENATAV
Fecha del camera ready: 24 de marzo de 2009

Resumen: El reconocimiento o identificación de una persona por su voz o reconocimiento del locutor, es una modalidad de identificación biométrica que ha extendido su uso debido a su ubicuidad y adecuada relación costo-beneficio. Los métodos de reconocimiento del locutor que dependen del texto son muy utilizados en aplicaciones de identificación biométrica en comercio y banca electrónica y en aplicaciones personalizadas de correo de voz, entre otros. Este trabajo pretende brindar un estado del arte de la temática, identificar los problemas aún no resueltos y recoger una bibliografía actualizada sobre el tema.

Palabras claves: reconocimiento del locutor, reconocimiento del locutor dependiente del texto, modelos acústicos del habla

Abstract: Speaker recognition is a biometric technique that has extended its use due to its ubiquity and adequate cost-benefit ratio. Text-dependent speaker recognition methods are widely used in biometric identification systems for commerce and electronic banking and for voice mail personalized applications. This work attempts to propose a state of the art of the text-dependent speaker recognition with speech acoustic models, identify problems not yet solved and pick up an updated bibliography on the theme.

Keywords: Speaker Recognition, Text-dependent Speaker Recognition, Speech Acoustics Models

1 Introducción

Recientemente, la biometría ha emergido como una disciplina científica que tiene como objetivo capturar automáticamente las características identificativas de las personas. La voz se usa mucho como técnica biométrica para el reconocimiento de personas al igual que el ADN o las huellas digitales, a pesar de verse afectada por muchas fuentes de variabilidad, como el ruido de fondo, el del canal de transmisión si se usan líneas telefónicas y la variabilidad en el comportamiento de la persona. La variabilidad en el comportamiento puede ser voluntaria, constituyendo un reto para los sistemas de reconocimiento pues se enfrentan a personas que tratan de imitar la voz de otro, y la involuntaria, que está dada por el estrés, enfermedades, el estado anímico o por el hecho de que las personas no pueden decir la misma frase dos veces precisamente de igual forma, incluso aunque lo deseen.

El reconocimiento automático de locutor se divide en dos tareas fundamentales: identificación del locutor (fig. 1a) y verificación del locutor (fig. 1b). La identificación es el proceso de decidir en un conjunto de personas a quien pertenece una muestra de voz determinada, la respuesta es identificar a la persona cuyo modelo brinde la máxima similaridad. La verificación consiste en comprobar si una muestra de voz pertenece a una persona, por tanto a

diferencia de la identificación, la respuesta del sistema será binaria: aceptar o rechazar la identidad.

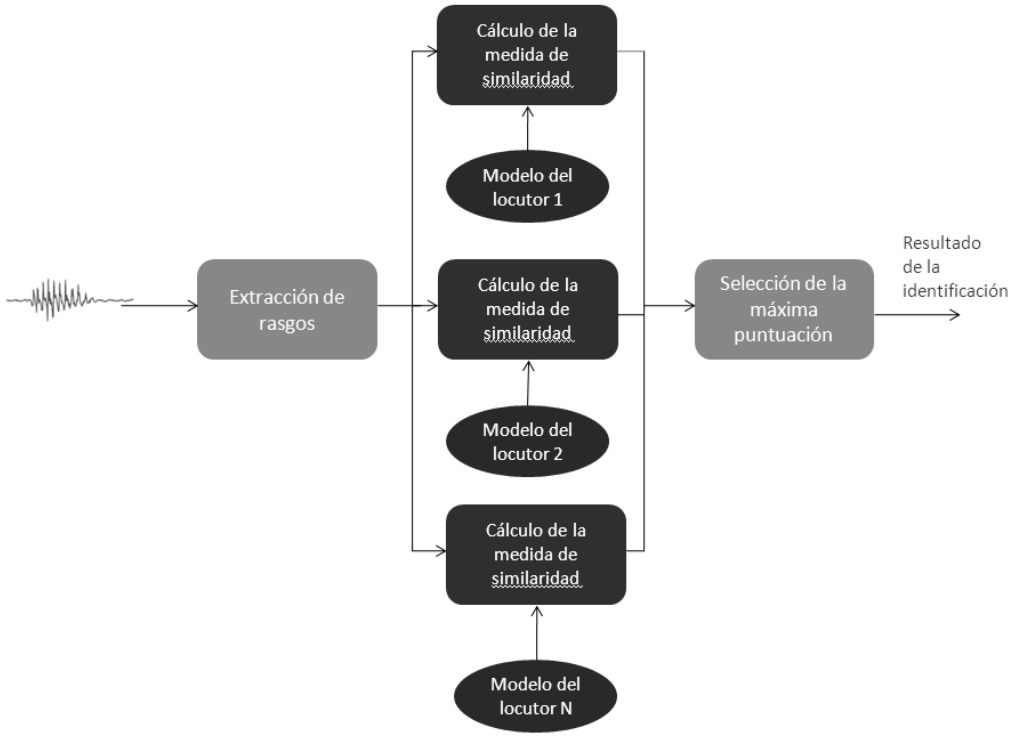

Fig. 1a. Identificación del locutor

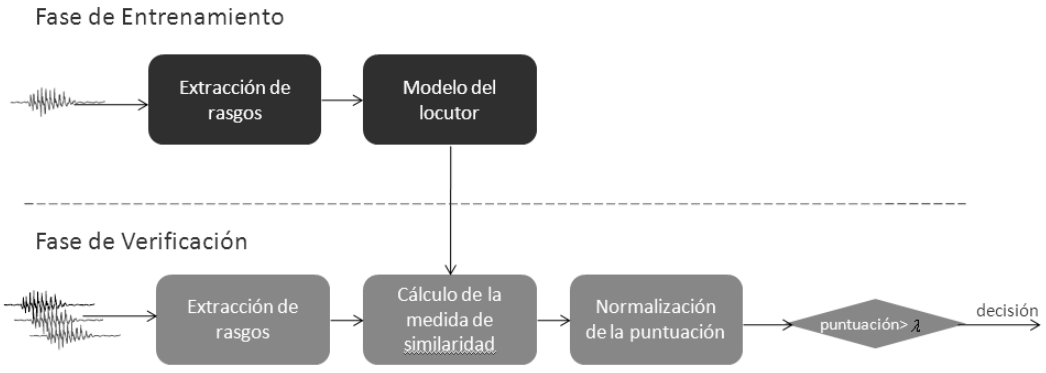

Fig. 1b. Verificación del locutor

El reconocimiento automático del locutor puede ser también dependiente del texto o independiente del texto. Existen diferencias y un solapamiento muy significativo entre ambos. Los avances en el reconocimiento del locutor independiente del texto obtenidos en las

evaluaciones NIST(National Institute of Standards and Technology) [1], pueden ser aplicados con éxito en el reconocimiento del locutor dependiente del texto haciendo sólo pequeñas modificaciones.

La diferencia fundamental está dada por el léxico que permite cada uno. El reconocimiento del locutor dependiente del texto puede utilizar la misma expresión para el entrenamiento y la prueba o simplemente no restringir el léxico para el entrenamiento, y asumir que el léxico activo durante la prueba es un subconjunto del léxico usado para el entrenamiento.

Esta limitación no existe para el reconocimiento del locutor independiente del texto, donde cualquier palabra puede ser pronunciada durante el entrenamiento y la prueba. En este caso, se necesita de mucha información en el entrenamiento (generalmente más de 30 segundos) para lograr buenos resultados. Por el contrario, en el reconocimiento del locutor dependiente del texto se obtienen resultados muy precisos con muy pocos datos (generalmente menos de 8 segundos de habla), porque se usan las mismas expresiones para el entrenamiento y la prueba.

Tradicionalmente, el reconocimiento del locutor independiente del texto estaba asociado con el reconocimiento del locutor en conversaciones completas. Recientemente, trabajos como los de [2] y [3] ayudaron a romper la brecha entre el reconocimiento del locutor independiente del texto y el dependiente del texto, usando para el entrenamiento las palabras más frecuentes pronunciadas en las conversaciones y aplicándoles a las mismas las técnicas del reconocimiento del locutor dependiente del texto. Estos trabajos dieron a conocer los beneficios de usar los algoritmos de reconocimiento del locutor dependiente del texto en tareas de reconocimiento del locutor independiente del texto. Este trabajo se enfocará particularmente a estudiar la verificación del locutor dependiente del texto debido a su amplia aplicación, brindando un estado del arte de la temática, identificando problemas aún no resueltos y recogiendo una bibliografía actualizada sobre el tema.

1.1 Rasgos del habla

Las primeras descripciones de sistemas de reconocimiento del locutor dependiente del texto se remontan a los principios de los 90.

Los rasgos acústicos que más se usan en sistemas de reconocimiento del locutor dependientes del texto son: los coeficientes cepstrales en escala Mel (MFCC, por sus siglas en inglés) [4] y los coeficientes de Predicción Lineal (LPC, por sus siglas en inglés) [5, 6]. La sustracción de la media cepstral y el alineamiento de los rasgos son reconocidos como técnicas muy efectivas para eliminar el ruido. Los rasgos dinámicos también han jugado un rol muy positivo en el reconocimiento del locutor dependiente del texto, [7]. También ha sido propuesto un método de mapeo de rasgos [8] similar al modelo de síntesis del locutor [9], que ha dado muy buenos resultados ante la robustez del canal.

1.2 Modelado Acústico

En el transcurso de los años han sido investigadas varias técnicas de modelado del habla. El esquema de modelado más común en los sistemas de reconocimiento del locutor es el modelado usando los modelos ocultos de Markov (HMM, por sus siglas en inglés) [10]. Las unidades modeladas por HMM dependen mucho del tipo de aplicación (fig. 2).

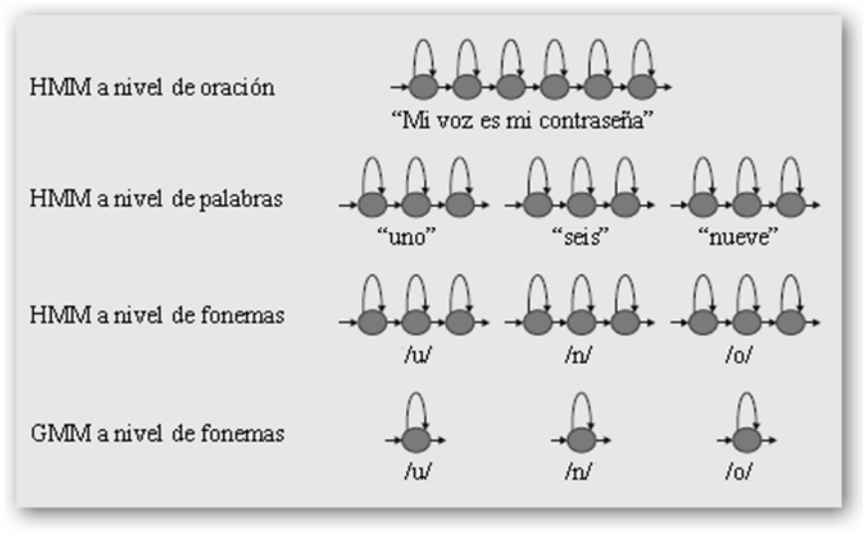

Fig. 2. Diferentes topologías de los HMM según el tipo de aplicación

En una aplicación donde el léxico del entrenamiento y la prueba son idénticos y están en el mismo orden, se puede usar un HMM a nivel de oración. Cuando el orden del léxico en las sesiones de entrenamiento y prueba no es el mismo, se usan HMM de palabras [6, 11]. Una aplicación de los HMM a nivel de palabras se presenta, por ejemplo, en un diálogo de reconocimiento del locutor basado en dígitos. En estos, todos los modelos de los dígitos son obtenidos en la fase de entrenamiento y durante la fase de prueba se solicita una secuencia de dígitos aleatoria. Los HMM a nivel de fonemas se usan para refinar la representación del espacio acústico [12, 13].

Se han usado regularmente, modelos HMM de derecha a izquierda con N-estados y modelos HMM de un solo estado llamados modelos de mezclas gausianas (GMM, por sus siglas en inglés) para modelar la acústica a nivel de fonemas, en el contexto del de reconocimiento del locutor dependiente del texto [14]. Son comunes dos tipos de modelado del habla usando GMM en reconocimiento del locutor dependiente del texto, el primero es el método GMM simple a nivel de oración o palabra, que es similar al modelo en tareas independientes del texto, y el segundo es una representación GMM a nivel de fonemas. Ambos métodos se caracterizan por la simplicidad, los resultados publicados y son un reflejo de la tendencia a la convergencia con los métodos de reconocimiento del locutor independiente del texto.

Para estos tipos de modelado, el entrenamiento se realiza mediante una adaptación bayesiana [15, 16] que altera los parámetros del modelo del locutor, usando los rasgos extraídos del habla obtenidos de la fase de entrenamiento. Esta forma de entrenamiento facilita la adaptación de los coeficientes, que es diferente para las medias, varianzas y pesos de las mezclas y tiene un gran impacto en el reconocimiento del locutor dependiente del texto.

Además de las tendencias actuales que son HMM y GMM, existen otros métodos de modelado del habla, menos difundidos. En [17] se proponen las máquinas de soporte vectorial (SVM, por sus siglas en inglés), método muy utilizado en el campo de reconocimiento del locutor independiente del texto [18, 19]. Trabajos en los que se usan las SVM para el reconocimiento del locutor dependiente del texto son los de [20, 21]. En ellos se evalúa la robustez del sistema basado en SVM para un léxico restringido. Los métodos de modelado por

Redes Neuronales [22] y los algoritmos de alineamiento dinámico en el tiempo (DTW, por sus siglas en inglés) [23] también han sido investigados y utilizados en el reconocimiento del locutor dependiente del texto.

1.3 Resultados de la clasificación por razón de similaridad

Como se dijo anteriormente, el reconocimiento del locutor puede ser dividido en: identificación del locutor y verificación del locutor. Para la verificación del locutor, el esquema estándar de clasificación está basado en la comparación de dos hipótesis:

H_0: La expresión de prueba es del locutor C que clama su identidad y está modelada por λ.

H_1: La expresión de prueba es de otro locutor que no es C y está modelada por $\overline{\lambda}$.

Matemáticamente, los resultados de de la clasificación (puntuación) se obtienen mediante la razón logarítmica de similaridad "*Log Likehood Ratio*" (LLR):

$$LLR(X|\lambda) = \log p(X|\lambda) - \log p(X|\overline{\lambda})$$

donde $X = \{x_1, x_2, ...x_T\}$ es el conjunto de vectores de rasgos extraídos de la expresión de prueba y $p(X|\lambda)$ es la similaridad de observar X dado el modelo λ. H_0 está representado por el modelo λ del locutor C que clama su identidad, λ puede ser entrenado usando GMM o HMM utilizando los rasgos extraídos de las expresiones pronunciadas por el locutor C en la fase de entrenamiento. La representación de H_1 es más delicada porque debe representar potencialmente a todos los locutores alternativos del locutor C.

Para el modelado de $\overline{\lambda}$ se han investigado dos métodos principales. El primero consiste en seleccionar N locutores de cohorte o background y modelarlos individualmente $(\lambda_1, \lambda_2 ... \lambda_{N-1})$ y luego combinar su similaridad en la prueba de la expresión.

El otro método para el modelado de $\overline{\lambda}$ usa las voces de un conjunto de locutores para entrenar un solo modelo, llamado modelo universal de background (UBM, por sus siglas en inglés).

1.4 Normalización de los resultados de la clasificación

Varias técnicas de normalización de la puntuación han sido propuestas con el objetivo de mejorar la discriminación entre locutores clientes e impostores y por consiguiente la robustez del sistema. Las técnicas H-Norm, Z-Norm y T-Norm forman parte de esos métodos.

Estos métodos utilizan la misma función de normalización:

$$LLR_{norm}(X|\lambda_n) = \frac{LLR(X|\lambda_n) - \mu}{\sigma} > \Delta$$

donde Δ es el umbral que se toma para aceptar o rechazar a un cliente, λ_n son los modelos de N locutores y X es un conjunto de vectores extraídos a una locución de prueba, μ y σ son la media y la desviación estándar de una distribución normal de LLR, respectivamente. En la estimación de los parámetros μ y σ es donde se diferencian estos tres métodos.

1.4.1 Z-Norm

Z-Norm es una técnica que normaliza los modelos de los locutores teniendo en cuenta las diferentes condiciones de entrenamiento bajo las que fueron creados antes de la prueba. Se prueba cada modelo del locutor (obtenido en el entrenamiento) contra un conjunto de rasgos de impostores [24], obteniéndose una distribución de puntuación del impostor. A partir de esta distribución, se estiman la media y la varianza, que se utilizan para la normalización de la puntuación:

$$LLR_{ZNORM}(X|\lambda_n) = \frac{LLR(X|\lambda_n) - \mu_I}{\sigma_I}$$

donde μ_I y σ_I son la media y la varianza de la distribución de impostores.

Esta técnica es muy eficiente desde el punto de vista computacional, pues la estimación de los parámetros de normalización puede ser realizada de forma *offline* mientras que el modelo del locutor se entrena.

1.4.2 H-Norm

H-Norm tiene como objetivo normalizar la incompatibilidad entre el canal de entrenamiento y el de la prueba [25]. Los parámetros son estimados enfrentando cada modelo del locutor contra rasgos producidos por impostores, pero dependientes del canal. Durante la prueba, el tipo de canal relacionado con los rasgos de entrada determina el conjunto de parámetros a usar para la normalización de la puntuación. Este método es similar al anterior, pero se conoce el canal por donde se obtienen los rasgos.

1.4.3 T-Norm

La normalización de prueba o T-Norm [26] se aplica de forma extensiva en el reconocimiento del locutor dependiente del texto [27]. Los modelos de los impostores se enfrentan con los rasgos de prueba de los clientes, obteniéndose la distribución de puntuación de los locutores de prueba. A partir de esta distribución, se estiman la media y la varianza que se utilizan para la normalización de la puntuación:

$$LLR_{TNORM}(X|\lambda_n) = \frac{LLR(X|\lambda_n) - \mu_{prueba}}{\sigma_{prueba}}$$

donde μ_{prueba} y σ_{prueba} son la media y la varianza de la distribución de locutores de prueba.

Esta técnica tiene un costo computacional muy alto porque los rasgos de la prueba se enfrentan a todo el conjunto de modelos de los impostores. Si Z-Norm es considerada una normalización dependiente del locutor, T-Norm es dependiente del texto. T-Norm lleva a cabo la estimación sobre los rasgos de prueba, por lo que evita la desigualdad acústica entre la prueba y el entrenamiento presente en Z-Norm. Z-Norm se usa generalmente para compensar la

variabilidad interlocutor y T-Norm para compensar la variabilidad de sesión, y pueden usarse de forma combinada.

1.5 Adaptación de los modelos del locutor

La adaptación es el proceso de extender la sesión de entrenamiento a las sesiones de prueba. Mientras el modelo se entrene con más voces, mejor será la precisión del sistema. Esto debe estar balanceado con los requerimientos de orden práctico donde las sesiones largas para el entrenamiento no son muy aceptadas por los usuarios finales.

La adaptación de los modelos puede ser supervisada o no supervisada. La supervisada, conocida también como adaptación manual necesita de un método de verificación externo para evaluar que el cliente es quien dice ser (legítimo). En la adaptación no supervisada la decisión que toma el sistema de verificación del locutor se usa para decidir si los nuevos datos se usarán para volver a entrenar el modelo de este locutor.

La adaptación supervisada es mejor que la no supervisada, según estudios realizados[28]. La adaptación no supervisada requiere un buen resultado en la comparación entre el modelo del cliente y la expresión de prueba para adaptar el modelo del locutor, debido a que las nuevas expresiones, si se parecen mucho a las modeladas, no aportan una variabilidad representativa para el locutor, el canal de transmisión y el ruido ambiental. El esquema de adaptación supervisado, como no está basado en la expresión que se va a verificar, aportará variabilidad al modelo del locutor de forma natural. Bajos ciertas condiciones, la adaptación supervisada puede reducir en tareas de verificación dependientes del texto el intervalo de error en un factor de 5, luego de 10 -20 iteraciones de adaptación.

2 Retos de la verificación dependiente del texto

El campo de reconocimiento del locutor dependiente del texto se enfrenta actualmente a varios retos tecnológicos relacionados con los algoritmos que se usan para desarrollar los sistemas. Estos retos fueron resumidos en una presentación realizada por [29] en el Taller Internacional sobre reconocimiento del locutor y del lenguaje Odyssey 2004.

2.1 Datos limitados y léxico restringido

El reconocimiento del locutor dependiente del texto está caracterizado por sesiones cortas de entrenamiento y sesiones de prueba. Las sesiones de entrenamiento consisten en varias repeticiones del léxico de entrenamiento, donde el habla total obtenida contiene generalmente de 4 a 8 segundos y los silencios son eliminados. La sesión de prueba consiste en una o dos repeticiones de un subconjunto del léxico de entrenamiento, obteniéndose una señal de 2 o 3 segundos.

El léxico posee una naturaleza restrictiva debido a lo cortas que son las sesiones de entrenamiento. Para obtener buenos resultados con señales cortas en las sesiones de entrenamiento y prueba es necesario restringir el léxico. La tabla 1 muestra varios ejemplos de léxico utilizado por varios sistemas para el entrenamiento. La tabla 2 describe varias estrategias de prueba en dependencia del léxico utilizado en el entrenamiento. En la mayoría de los casos, el

léxico de prueba es exactamente igual que el léxico de entrenamiento. En esquemas de pruebas aleatorios se usa el método 2 por 4 con el objetivo de reducir la carga cognitiva, que consiste en dos repeticiones de una cadena de 4 dígitos. Ya más de 4 dígitos es más difícil de recordar para el usuario.

Abreviación	Descripción
E	Contar del 1 al 9: uno, dos, tres…
T	Número de teléfono de 10 dígitos
S	Número de cuenta de 9 dígitos
N	Nombre y apellidos
MVEMC	Mi voz es mi contraseña

Tabla 1. Ejemplos de léxico para el entrenamiento

Abreviación	Descripción
E	Contar del 1 al 9: uno, dos, tres…
R	Secuencia de dígitos aleatoria: 2 6 8 5 2 6 8 5
pR	Secuencia de dígitos pseudoaleatoria seleccionada de E: 2 3 6 7 2 3 6 7
T	Se usa el mismo número de teléfono de 10 dígitos del el entrenamiento
RT	Secuencia de dígitos aleatoria seleccionada del léxico del entrenamiento
pRT	Secuencia de dígitos pseudoaleatoria seleccionada del léxico del entrenamiento
S	Similar a T pero para un número de cuenta de 9 dígitos
N	Nombre y apellidos
MVEMC	Mi voz es mi contraseña

Tabla 2. Ejemplos de léxico para la prueba. Las abreviaciones están relacionadas con las definidas en la tabla 1

2.2 Uso del Canal

Esta es un área muy importante donde los algoritmos deben ser mejorados, pues en las aplicaciones actuales los usuarios utilizan varios tipos de canales: celulares, teléfonos fijos, etc.

La desigualdad del canal se pone de manifiesto cuando en la sesión de prueba se utiliza un canal diferente al de la sesión de entrenamiento. Las proporciones de llamadas que no son realizadas por el mismo canal que se usa para el entrenamiento están entre el 25-50% de todas las que sí lo usaron. Este problema influye en la exactitud del sistema, pues en vez de duplicar tasa de error [30, 31], la cuadruplican [30].

2.3 Envejecimiento de los modelos del locutor

La exactitud de los sistemas de reconocimiento del locutor ha sido medida en algunas aplicaciones comerciales y en colecciones de datos [32].Los resultados mostraron que dicha exactitud va disminuyendo a medida que pasa el tiempo. En el caso de [32] el intervalo de error aumentó en un 50% en dos meses.

Existen varias fuentes que causan el envejecimiento de los modelos, las principales son el envejecimiento natural, el uso del canal y los cambios en el comportamiento de las personas.

El envejecimiento natural está relacionado con los cambios fisiológicos que ocurren en el aparato fonador con el paso de los años. Los cambios en el uso del canal causan también que el modelo del locutor esté desactualizado respecto al canal que se usa en ese momento. Los cambios en el comportamiento ocurren a la vez que los usuarios utilizan la aplicación, pues la primera vez que la utilizan tienden a cooperar y a hablar lento, pero mientras pasa el tiempo alteran la forma de interactuar con la aplicación.

Todos estos factores afectan los modelos, los resultados de la verificación y por consiguiente, la exactitud del sistema. La forma más común de solucionar este problema es adaptando los modelos del locutor de forma periódica.

3 Principales resultados obtenidos

3.1 Rasgos de entrada del reconocedor de habla y del locutor

Los algoritmos más utilizados para la extracción de rasgos en el reconocimiento del habla son MFCC y LPCC. Estos algoritmos han sido desarrollados con el objetivo de clasificar fonemas o palabras (léxico) para tareas independientes del texto. Igualmente, para el reconocimiento del locutor se usan mayormente estos dos algoritmos, aunque el objetivo de este último es la clasificación de locutores, sin importar el contenido léxico. Estos rasgos se usan para ambas tareas por su forma efectiva de representar la señal de voz en general.

Varios estudios han tratado de cambiar el paradigma de reconocimiento para la extracción de rasgos. En [33], una red neuronal con 5 capas es entrenada discriminativamente para maximizar la discriminación del locutor. Luego, las dos últimas capas son descartadas y la capa final

resultante constituye un extractor de rasgos. Los autores reportaron una mejora relativa de un 28 % sobre MFCC en tareas de reconocimiento del locutor independientes del texto.

En [34], las transformadas *wavelet* son usadas para analizar las series de tiempo de la voz en vez del análisis estándar de Fourier. Los autores reportaron una reducción de la tasa de error entre un 15 y 27 % en tareas de reconocimiento del locutor independientes del texto. A pesar de las mejoras reportadas en la literatura, esos algoritmos no logran reemplazar aún a los LPCC o MFCC.

3.2 Exactitud dependiente del Léxico

El tema del contenido léxico del habla es fundamental en el reconocimiento dependiente del texto. Estudios realizados por [32] demostraron que la preservación de la secuencia de dígitos mejora la exactitud del sistema. Los autores reportaron una mejora de más de un 50% cuando la secuencia de dígitos usada en la fase de prueba preserva el orden de la utilizada en el entrenamiento.

En [35], el efecto de la desigualdad en el léxico es comparado con el efecto de la desigualdad debido al cambio en la relación señal-ruido y al cambio en el canal. Se reportó que una desigualdad moderada en el léxico puede afectar más el rendimiento que la variación en la relación señal-ruido y es comparable con la desigualdad del canal.

3.3 Diseño del modelo de background

El diseño del modelo de *background* es fundamental para obtener resultados precisos en un sistema de reconocimiento del locutor. El efecto del léxico también se puede ver en este contexto. Por ejemplo, en un sistema de reconocimiento del locutor dependiente del texto basado en expresiones que son la contraseña, adaptar un modelo de *background* con expresiones exactas a las que dijo el cliente, tiene un impacto muy positivo [36]. En [37], se presenta un algoritmo que selecciona para un cliente varios locutores de *background*. El algoritmo está basado en la similaridad entre las sesiones de entrenamiento de dos usuarios. El contenido léxico no fue objeto de estudio, pero sería interesante investigar si el contenido léxico de cada sesión de entrenamiento influye en la selección de locutores de *background* competitivos, por ejemplo, analizar si locutores similares poseen algún solapamiento léxico significativo.

3.4 Adaptación de los modelos del locutor

La adaptación en línea de los modelos del locutor [38, 39] es un componente fundamental en cualquier aplicación exitosa de reconocimiento del locutor, especialmente en la tarea dependiente del texto debido a las cortas sesiones de entrenamiento.

En [28] se describe un protocolo para evaluar la adaptación . Para cada modelo del locutor, los datos se dividen en tres conjuntos: conjunto de entrenamiento, conjunto de adaptación y conjunto de prueba. El conjunto de adaptación está compuesto por un intento de acceso al sistema por parte del impostor por cada ocho intentos de acceso del cliente (distribuidos aleatoriamente).

Los experimentos fueron diseñados de la siguiente forma. Primero, los modelos de los locutores fueron entrenados y la precisión de los mismos fue medida justo después del

entrenamiento usando el conjunto de prueba. Entonces, una expresión de adaptación fue presentada para cada uno de los modelos del locutor, se decide si se adapta o no el modelo a dicha expresión. Después de esta iteración de adaptación, la precisión fue medida usando el conjunto de prueba. Los pasos de adaptación y prueba son repetidos para cada iteración de adaptación en el conjunto de adaptación. Este protocolo fue diseñado para controlar con gran precisión todos los factores relacionados con el proceso de adaptación.

Actualmente, existen varios algoritmos de adaptación de los modelos del locutor, los más utilizados hasta la fecha son MAP, "*Maximum a Posteriori*" y MLLR, "*Maximum Likehood Linear Regression*", por los buenos resultados que han reportado[40] [41].

3.5 Normalización de la puntuación T-Norm en el contexto del reconocimiento del locutor dependiente del texto

La normalización T-Norm es sensible al léxico usado en las expresiones utilizadas para entrenar los modelos de los impostores obtenidos del modelo de cohorte [42]. En experimentos realizados por [43] se utilizaron 2 tipos de modelos de cohorte y dos tipos de expresiones, que se clasificaron en: léxicamente pobre y léxicamente rico. La riqueza estuvo dada por la variedad de contextos en los que cada dígito puede ser encontrado. Esta riqueza léxica que posee el modelo de cohorte lo hizo robusto ante la variedad de cadenas de dígitos que pueden ser encontradas en la etapa de prueba.

Los resultados obtenidos fueron muy interesantes. El uso de un modelo de cohorte léxicamente pobre en el contexto de expresiones léxicamente ricas degrada significativamente la precisión del sistema, incluso la tasa de error da mayor que si no se aplica T-Norm a la puntuación. Mientras que en todos los otros casos, T-Norm mejora la exactitud del sistema.

4 La verificación texto-dependiente del locutor utilizando modelos acústicos del habla

En la verificación del locutor dependiente del texto se distinguen dos métodos fundamentales: Alineamiento dinámico en el tiempo (DTW) y los modelos ocultos de Markov (HMM). El DTW es descrito aquí por motivos históricos pues se usa muy poco en la actualidad.

Alineamiento dinámico en el tiempo, DTW: Consiste en comparar la locución de entrada con un conjunto de plantillas que representan las expresiones a reconocer. El entrenamiento se basa en almacenar en plantillas las expresiones a reconocer. Esas plantillas son conjuntos de rasgos acústicos ordenados en el tiempo. Para el reconocimiento se debe alinear de manera óptima la secuencia de rasgos de entrada con el modelo de referencia previamente almacenado. Al concluir la comparación, la distancia acumulada entre las dos expresiones es la base de la puntuación [44]. Este método es bastante simple y no requiere muchos recursos computacionales en la fase de entrenamiento. Se puede aplicar en sistemas de control de acceso con contraseña, teniendo previamente las plantillas de todas las posibles contraseñas. Esta es una desventaja de este método pues, al depender de las expresiones de referencia, imposibilita la variabilidad en la señal de voz.

Modelos ocultos de Markov, HMM: Esta técnica de modelado estadístico ha sido muy utilizada en los campos de reconocimiento del habla y reconocimiento del locutor dependiente del texto por su habilidad de modelar adecuadamente la gran variabilidad en el tiempo de la

señal de voz [45]. Este método ha mostrado ser muy efectivo en el modelado y reconocimiento de fonemas, palabras y frases [46]. Las contraseñas, que consisten en secuencias de palabras tales como los dígitos, se utilizan mucho. En ellas, cada palabra está caracterizada por un HMM con un pequeño número de estados, donde cada estado es representado por una densidad de mezcla gausiana. Los parámetros del HMM son entrenados tomando varias repeticiones de la contraseña. De este proceso se obtiene el modelo de la contraseña y con los rasgos de la frase a verificar, se calcula la puntuación que permite decidir si aceptar o rechazar al cliente.

Esta técnica tiene una fundamentación estadística sólida con algoritmos de aprendizaje muy eficientes y además, posee una gran adaptabilidad a la variabilidad de las condiciones de la voz o del canal de transmisión. Dado que los HMM tienen menos cantidad de estados que ventanas en cada expresión, son mejores y más rápidos que los sistemas basados en DTW. Sin embargo, necesitan de muchos datos para el entrenamiento en aras de lograr una buena estimación de los parámetros del modelo [47].

4.1 Métodos de verificación texto-dependiente del locutor utilizando modelos acústicos del habla

Se han propuesto varios métodos y combinaciones de éstos para los sistemas de verificación del locutor dependiente del texto. A continuación se presentará una breve taxonomía de los mismos que muestra cómo se ha desarrollado estos métodos en la última década.

4.1.1 Autentificación de usuario con información verbal

En "*Automatic Verbal Information Verification for User Autentication*" [48] se propone un sistema de autenticación del usuario, que combina la verificación de la información verbal (VIV) con la verificación del locutor.

El usuario es verificado en los primeros cinco accesos mediante la VIV. Este es el proceso de verificar las expresiones que emitió un usuario contra la información almacenada en el perfil del usuario. En estos sistemas, el usuario se identifica y contesta varias preguntas de las cuales se elige una de las expresiones de respuesta como contraseña para la posterior verificación del locutor. En el entrenamiento, el VIV colecciona y verifica la expresión-contraseña, obteniéndose un modelo HMM dependiente del locutor, como se muestra en la figura 3:

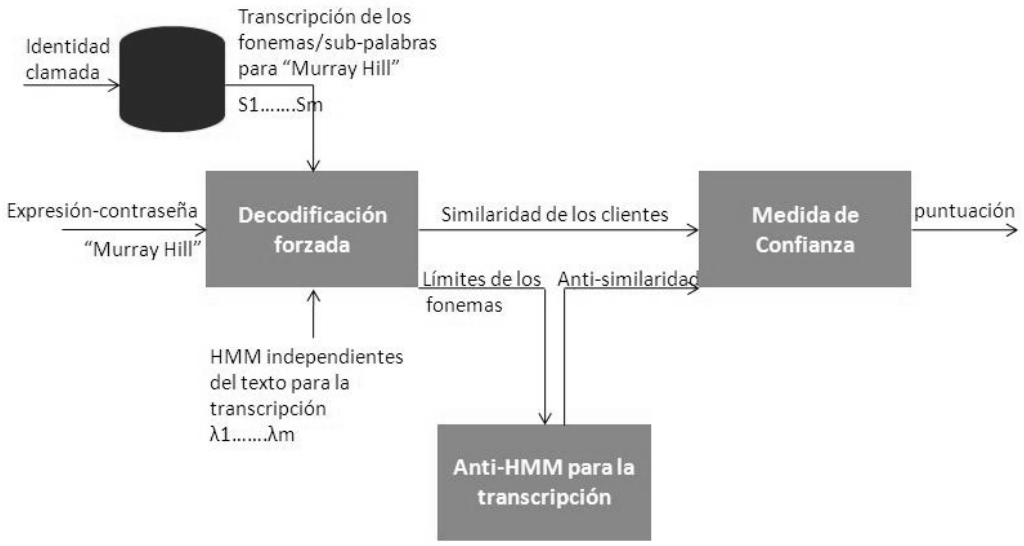

Fig. 3. Obtención de un modelo HMM dependiente del locutor para la expresión-contraseña en el entrenamiento del VIV

La expresión de entrada es alineada con una secuencia de fonemas transcritos de la respuesta correcta usando HMM independientes del texto. Luego, para cada fonema, se calculan las puntuaciones con relación a los HMM independientes del texto y al conjunto de anti-HMM de la frase correspondiente. Con estas puntuaciones y una medida de confianza, se obtiene un valor con el cual se puede realizar la verificación de la expresión. El modelo HMM de dicha expresión con mejor puntuación se almacena como modelo de la contraseña.

En la etapa de reconocimiento, o sea, cuando el usuario clama su identidad, el sistema espera la misma frase obtenida en la sesión de entrenamiento (figura 4).

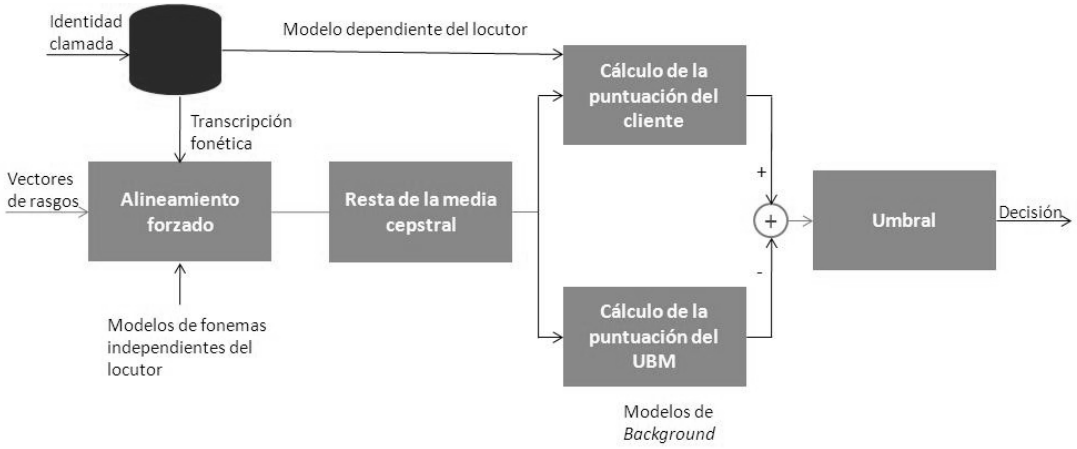

Fig. 4. Etapa de reconocimiento del locutor del VIV con la expresión-contraseña

A la frase dicha por el usuario se le extraen los rasgos y se le hace un alineamiento forzado usando la trascripción fonética de la expresión-contraseña que está almacenada y los modelos

HMM de fonemas independientes del locutor y se le resta la media cepstral. Se calcula la puntuación del locutor usando los rasgos normalizados y el modelo de la expresión-contraseña, mediante el algoritmo Viterbi. La puntuación contra el modelo UBM también es calculada. Para tomar la decisión de aceptar o rechazar el usuario, son restadas las puntuaciones del locutor y del modelo UBM y el valor se compara con un umbral. Si el resultado es mayor se acepta, de lo contrario es rechazado.

Este método brinda más seguridad al usuario y elimina los inconvenientes de un proceso de entrenamiento formal, garantiza la calidad de los datos de entrenamiento para el sistema de verificación mediante la verificación de la información verbal y modela las desigualdades causadas por los diferentes ambientes acústicos entre el entrenamiento y la prueba, pues los datos fueron recolectados de diferentes canales.

4.1.2 Verificación de locutor con reconocimiento de habla continua de gran vocabulario (LVCSR)

En "*Speaker Verification using Text-Constrained Gaussian Mixture Models*" [49] se propone un sistema GMM-UBM restringido en texto, usando segmentaciones de palabras producidas por un sistema LVCSR " *Large Vocabulary Continuous Speech Recognition*", permitiendo al sistema enfocarse en las diferencias de locutores dentro de un conjunto de palabras, como se muestra en la figura 5:

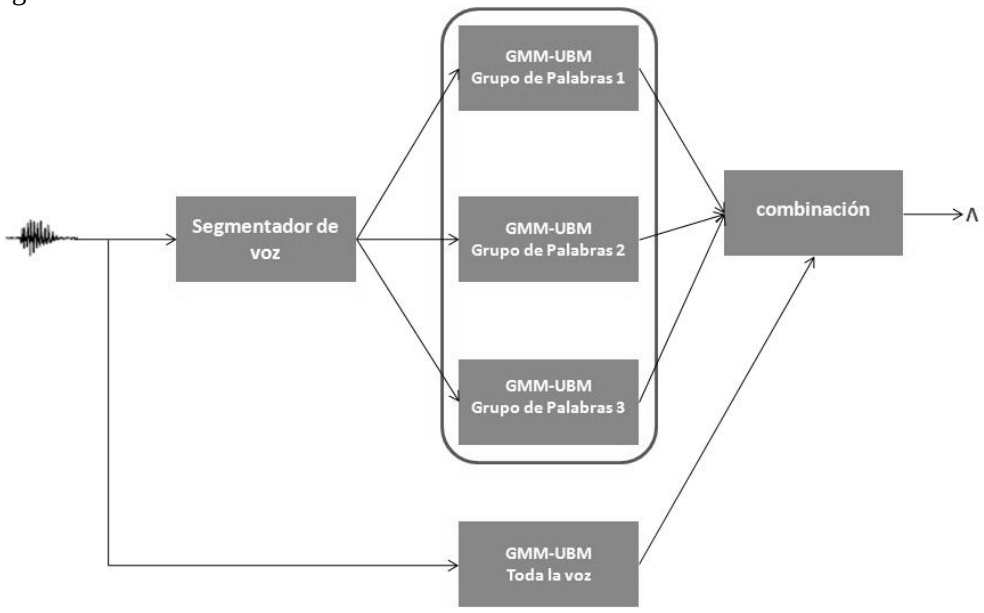

Fig. 5. Sistema GMM-UBM de verificación de locutor restringido en texto

El habla es segmentada en palabras y los verificadores GMM-UBM son entrenados y probados usando solo la voz de ese grupo de palabras. Para ello se usa un segmentador, generalmente un reconocedor del habla LVCSR, para dividir la voz de entrada en palabras. La voz correspondiente a cada palabra es usada para entrenar el sistema GMM-UBM restringido a esas palabras. El UBM se entrena con un gran número de locutores, usando solamente la voz proveniente de ese grupo específico de palabras.

En la verificación, se usa el mismo grupo de palabras para el cálculo de la similaridad. Durante la verificación se usan en paralelo los GMM-UBM restringidos a diferentes palabras. Las puntuaciones combinadas obtenidas se usan para producir el resultado final.

La ventaja de este método viene de restringir el habla y comparar los mismos grupos de palabras pronunciadas por diferentes locutores. Las desventajas están dadas por: primero, la necesidad de tener un buen segmentador, porque, si su efectividad es baja eliminaría la especificidad de los modelos condicionados a las palabras, incumpliéndose el objetivo fundamental de este sistema, y segundo, requiere de grandes cantidades de información para el entrenamiento y la verificación, así como transcripciones del habla de alta calidad.

En "*Text-Constrained Speaker Recognition on a Text-Independent Task*" [50], se propone un método de reconocimiento del locutor en el dominio de conversaciones telefónicas, independiente de lo que se dice, que usa un sistema similar restringido a texto pero usando modelos HMM-UBM. Se toman como palabras de interés aquellas que ocurren con alta frecuencia en el dominio, que serán encontradas en el habla no restringida.

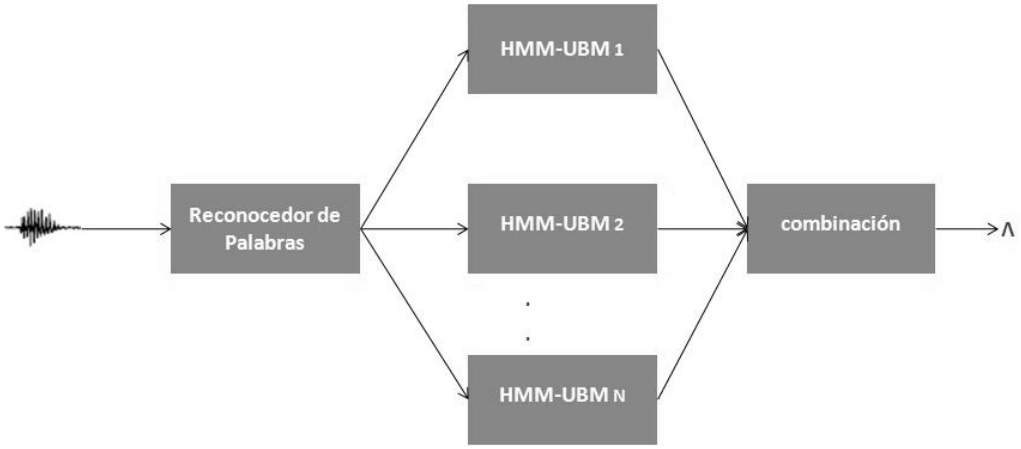

Fig. 6. Sistema HMM-UBM de verificación de locutor restringido en texto

Primeramente, se usa un reconocedor de habla LVCSR para obtener las palabras más frecuentes. A las mismas se le extraen los rasgos y se entrena un UBM independiente del locutor para ellas. Los modelos se construyen a partir de los rasgos y el UBM, usando los HMM y la adaptación MAP [51].

En el proceso de verificación se genera una puntuación para cada experimento modelo-prueba. Para cada una de las palabras seleccionadas en el segmento de prueba, se calcula la puntuación del modelo del locutor, como la puntuación acumulada en la ventana sobre todas las instancias de prueba, cuando el HMM adaptado del locutor se alinea forzadamente a la secuencia de la ventana. Las puntuaciones son combinadas en todas las palabras claves para obtener una puntuación combinada. Se calcula la diferencia entre la puntuación del modelo del locutor y el modelo de background, y el resultado se normaliza.

Este método también restringe el habla y se basa en el reconocimiento previo de palabras, necesitando también de un buen segmentador, lo que es una desventaja. Además, para modelar se usan modelos HMM, que poseen un alto costo computacional.

4.1.3 Verificación de locutor con elección de la contraseña y múltiples modelos de referencia

En "*User-customized password speaker verification using multiple reference and background models*" [52], se propone un método de verificación del locutor que utiliza una combinación de HMM/GMM, donde el usuario puede elegir su contraseña. La contraseña debe ser pronunciada varias veces en la fase de entrenamiento para con ella crear un modelo HMM independiente de ese cliente. En este sentido, lo primero es inferir la topología del HMM, tarea que se realiza usando un híbrido HMM/MLP. Para obtener los parámetros del modelo inferido y efectuar la adaptación, se usan los GMM.

El uso de múltiples modelos de referencia para el modelado acústico y múltiples modelos de background para la normalización de la similaridad, fueron las principales contribuciones de este trabajo. Además, se indaga en varias técnicas de cálculo de puntuación tales como Selección Dinámica del Modelo (DMS, por sus siglas en inglés) y técnicas de fusión de clasificadores.

Los resultados obtenidos empleando dos protocolos diferentes mostraron que un criterio de selección apropiado para los modelos del cliente y el de background pueden mejorar significativamente el rendimiento del sistema de verificación del locutor, con selección de la contraseña, haciéndolo tan competitivo como un sistema dependiente de texto.

Este método posee un alto costo computacional, pues combina varios métodos de modelado y clasificación complejos (HMM, GMM y Redes Neuronales) con el objetivo de obtener un modelo HMM dependiente del locutor. Además, posee varios problemas a los que se enfrentan los que trabajan en este campo: inferencia de la topología del HMM, adaptación del locutor y normalización eficiente de la similaridad.

4.1.4 Verificación de locutor con modelo del habla HMM independiente del locutor combinado con modelo del locutor GMM/UBM

En "*On combining classifiers for speaker authentication*"[53], se presenta un sistema de autenticación de locutor basado en la combinación de varios clasificadores. Se diseñan dos clasificadores individuales: el verificador de la expresión y el verificador del locutor, tomando en cuenta aspectos prácticos de implementación como la complejidad, tiempo de entrenamiento, etc. Para modelar los locutores y realizar la verificación, utilizaron el modelo GMM-UBM. Para modelar la expresión, usaron modelos HMM de sub-palabras y para el modelo de background, utilizaron los HMM de cadenas de sub-palabras más cercanas. Se evalúan reglas de combinación como la de la suma, la del producto, la del mínimo y la del máximo, así como las redes neuronales.

Se investigó para el verificador del locutor la relación entre el número de mezclas gausianas del GMM-UBM y el rendimiento. La conclusión principal obtenida de los resultados es que con 16 mezclas, existe un buen compromiso entre rendimiento y complejidad.

Para propósitos de autenticación del locutor, el verificador de expresiones trabaja peor que el del locutor. Sin embargo, resultados experimentales mostraron una mejora en el rendimiento cuando los dos verificadores se combinan.

En lo referente a la mejor forma de combinación de los clasificadores, experimentos mostraron que las redes neuronales funcionan mejor que los otros métodos debido a su capacidad de aprender el punto óptimo de operación con los resultados de los clasificadores.

En "*Reinforced Temporal Structure Information For Embedded Utterance-Based Speaker Recognition*" [54], se propone un método diseñado para aplicaciones embebidas (contienen una

cantidad limitada de recursos computacionales). Utiliza las ventajas del GMM/UBM independiente del texto y del reconocedor de habla HMM. Además, proponen reforzar el modelado de la información temporal, que permite mejorar el rendimiento de los sistemas de reconocimiento del locutor particularmente, cuando se cuenta con pocos datos.

La figura 7 muestra la arquitectura del sistema embebido. Los nodos de esta estructura son modelos GMM. La capa superior es la menos especializada, es un UBM clásico que modela el espacio acústico. La capa media contiene modelos del locutor independientes del texto. Los mismos son obtenidos mediante un proceso de adaptación de los modelos, se usó la adaptación MAP. Sólo se adaptó la media, los demás parámetros son los mismos que posee el UBM. La capa final es un SCHMM (HMM semi-continuo), con topología de derecha a izquierda que captura la información temporal de las expresiones que el usuario usa para autenticarse. Cada uno de los estados SCHMM es un GMM que se derivó de la capa media.

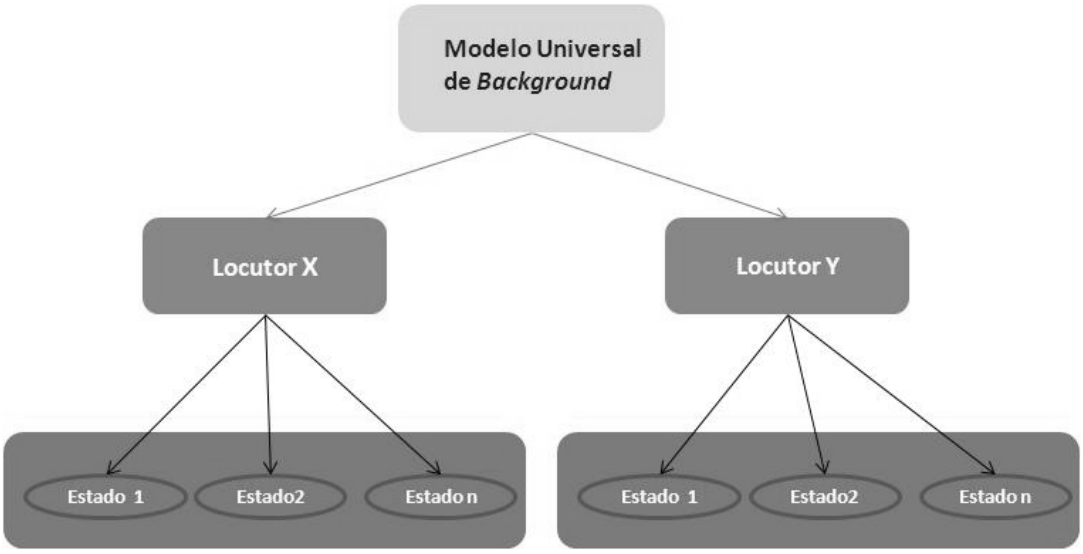

Fig. 7. Vista general de la arquitectura del modelo embebido

Se hicieron experimentos donde los impostores conocían y desconocían la expresión que el cliente utiliza para autenticarse. El rendimiento del método propuesto es equivalente al de un GMM/UBM cuando los impostores conocen las expresiones que usan los clientes (EER que se obtuvo con GMM fue de 4,49 y con el sistema propuesto 4,61). Sin embargo, cuando los impostores no conocen la expresión de los clientes, el método propuesto supera al GMM (ERR del sistema propuesto 0,56 y de GMM 0,87). Este nuevo método utiliza las ventajas del contenido lingüístico de las expresiones del cliente, pero no toma en cuenta la variabilidad de las expresiones ni la duración que pudiera poseer la misma.

4.1.5 Verificación de locutor con modelos HMM adaptados al locutor (MLLR)

La adaptación MLLR (*Maximum-Likelihood Linear Regression*) [55, 56],consiste en transformar matrices de medias y opcionalmente matrices de covarianza de un modelo HMM mediante una

transformación afín que maximice la función de similaridad dados los nuevos datos de adaptación y el modelo:

$$\hat{\mu} = A\mu + b$$
$$\hat{\sum} = H \sum H^T$$

donde μ es el vector de medias en el modelo, $\sum$ es la matriz de covarianza, $\hat{\mu}$ y $\hat{\sum}$ las matrices de media y covarianza adaptadas respectivamente, (A, b) es la transformación afín para la adaptación de la media y H la matriz de transformación para la adaptación de la covarianza. Para encontrar los parámetros óptimos, se usa generalmente el método de maximización de la expectancia (EM, por sus siglas en inglés) en dos pasos: estimar la transformación de la media dados A y b y luego estimar la transformación H de la covarianza.

La adaptación MLLR es una técnica especialmente desarrollada para la adaptación de modelos HMM independientes del locutor, a la voz de un locutor en particular a partir de un número limitado de expresiones y por lo tanto, se puede aplicar al problema de reconocimiento del locutor dependiente de texto. Además, en los casos en los que se cuenta con pocas expresiones para realizar la adaptación, la adaptación MLLR consigue mejores resultados [57] si se agrupan en clases y se transforman las medias de toda la clase, utilizando para ello una misma transformación lineal. De esta manera, la adaptación MLLR reduce el número de parámetros a entrenar, pasando de depender linealmente del número de gausianas a depender linealmente del número de clases. Esto la convierte en una técnica muy robusta de adaptación de modelos HMM a un locutor, incluso cuando se utilizan modelos más complejos.

En la tesis “Reconocimiento de Locutor dependiente de texto mediante adaptación de modelos ocultos de Markov fonéticos” [57], se estudian los sistemas de reconocimiento del locutor dependiente del texto que usan como herramienta de modelado y clasificación los HMM. Se comparan dos métodos de adaptación: el Baum-Welch y el MLLR. Los resultados experimentales mostraron que la adaptación MLLR tiene un mejor rendimiento que la Baum-Welch, incluso si se aumenta la cantidad de datos del entrenamiento.

En “*MLLR Transforms as features in Speaker Recognition*”[40], se explora el uso de la adaptación MLLR del sistema de reconocimiento del habla como rasgos para el reconocimiento del locutor.

El sistema de reconocimiento del habla realiza una primera descodificación usando los coeficientes MFCC y un modelo del lenguaje bi-grama. Las hipótesis resultantes son usadas para adaptar un segundo conjunto de modelos basados en rasgos PLP. Estos modelos adaptados son usados en un segundo paso de descodificación que está restringido por tri-gramas que generan las listas de los N-mejores. Estos son recalculados por un modelo del lenguaje cuatri-grama y por modelos prosódicos, hasta llegar a la palabra final.

La transformada MLLR se aplica en los dos pasos del reconocimiento. En el primero, se basa en un modelo de fonemas de referencia, usando tres transformadas: para no-voz, para fonemas sonoros y no sonoros. El segundo paso de descodificación está basado en las palabras de referencia generadas por el primer paso, y se aplican nueve tipos de transformadas diferentes a las clases: no voz, vocales altas/bajas, consonantes sonoras, explosivas sonoras/no sonoras, fricativas sonoras/no sonoras y nasales.

Los coeficientes de una o más transformadas de adaptación son concatenados en un vector de rasgos y modelados usando SVM. El SVM es entrenado para cada locutor usando los rasgos

de un conjunto de entrenamiento de *background* como muestras negativas y los datos de los locutores como muestras positivas.

Además, el rango dinámico del vector de coeficientes es normalizado, que reemplaza cada valor del rasgo por su rango en la distribución de *background*, esta normalización realiza un re-escalado adaptado de los rasgos para obtener una distribución aproximadamente uniforme.

El método propuesto está compuesto por los siguientes componentes: extracción de vectores de rasgos a partir de muestras de habla de locutores por medio de una etapa de adaptación de un reconocedor del habla y construcción de una función discriminante del locutor mediante SVM. Este método de adaptación de rasgos ha dado muy buenos resultados, estando a la altura e incluso superando a los métodos cepstrales y fue evaluado como el mejor rasgo en un estudio del estado del arte de los rasgos más utilizados para el reconocimiento del locutor [58].

En "*Constrained MLLR for Speaker Recognition*" [59], se propone un nuevo método de extracción de rasgos para el reconocimiento del locutor basado en la adaptación CMLLR "*Constrained Maximum Likehood Linear Regression*", el cual evita el uso de las transcripciones para obtener los modelos de las palabras. El método CMLLR aplicado a los sistemas de reconocimiento del locutor permite extraer rasgos que están más enfocados a las características relacionadas con el locutor.

El método presentado posee dos etapas. En la primera, se construye un modelo universal de *background* GMM/UBM a partir de los rasgos cepstrales del locutor. En la segunda, se estiman las transformadas CMLLR para cada locutor de interés usando el UBM creado, obteniéndose un vector de rasgos de alta dimensión por locutor, el cual se modela usando las SVM.

La ventaja de esta técnica sobre la propuesta por [40] es que el proceso de entrenamiento no depende de la trascripción ni del lenguaje, y así captura las diferencias entre los rasgos acústicos independientes del locutor y dependientes del locutor. Sin embargo, dado que se usa un modelo GMM para estimar la transformada CMLLR, esta es menos precisa y probablemente más dependiente del mensaje. Este método combinado con sistemas MFCC-SVM y MFCC-GMM tiene rendimientos muy significativos.

4.1.6 Verificación de locutor con modelos HMM adaptados al locutor (MAP)

La adaptación MAP es un proceso de estimación en dos pasos. En el primer paso se estiman los estadísticos de los datos de entrenamiento para cada mezcla del UBM. En el segundo paso, se combinan estos nuevos estadísticos con los estadísticos de los parámetros del UBM.

De manera más especifica:

Dado un UBM y un vector de entrenamiento perteneciente al locutor que se desea adaptar $X = \{x_1, x_2, x_3 ... x_T\}$, es necesario, primero, determinar el alineamiento probabilístico que existe entre el vector de entrenamiento y las mezclas que componen el UBM.

Esto es, para la mezcla i del UBM, se calcula:

$$P(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{j=1}^{M} w_j p_j(x_t)}$$

A partir de ese término y de x_t, se calculan los estadísticos necesarios para calcular a su vez los pesos, medias y varianzas:

$$n_i = \sum_{t=1}^{T} P(i|x_t)$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} P(i|x_t) x_t$$

$$E_i\left\{(x-\mu)^2\right\} = \frac{1}{n_i} \sum_{t=1}^{T} P(i|x_t) x_t^2$$

Finalmente, con estos nuevos estadísticos calculados de los datos de entrenamiento, se actualizan los estadísticos antiguos del UBM para cada mezcla i para obtener los parámetros adaptados:

$$\hat{w}_i = \left[\alpha_i^w n_i / T + \left(1-\alpha_i^w\right) w_i\right] \gamma$$

$$\hat{\mu}_i = \alpha_i^m E_i(x) + \left(1-\alpha_i^m\right) \mu_i$$

$$\hat{\sigma}_i^2 = \alpha_i^v E_i(x^2) + \left(1-\alpha_i^v\right)\left(\hat{\sigma}_i^2 + \hat{\mu}_i^2\right) - \hat{\mu}_i^2$$

$\left\{\alpha_i^w, \alpha_i^m, \alpha_i^v\right\}$ son los coeficientes que controlan el balance entre las estimaciones antiguas y nuevas de los pesos, medias y varianzas, respectivamente. Estos coeficientes se definen como sigue:

$$\alpha_i^p = \frac{n_i}{n_i + r^p}, p \in \left\{\alpha_i^w, \alpha_i^m, \alpha_i^v\right\}$$

siendo r^p un factor fijo de relevancia para el parámetro p.

Además, γ es un factor de escala que se calcula sobre todos los pesos adaptados para asegurar que éstos suman uno.

La adaptación MAP, al combinar los modelos GMM de los locutores con los modelos universales de *background* (UBM), ha sido clave en la mejora de los clasificadores, alcanzando el estado del arte de dichos modelos durante la evaluación del *National Institute for Standards and Technology*, NIST del 2004.

En "*Unsupervised Learning of HMM Topology for Text-dependent Speaker Verification*" [47], se hace un análisis del problema de los pocos datos de entrenamiento que existe en el reconocimiento dependiente del texto. La adaptación MAP "*Maximum a Posteriori*" ayuda a solucionar este problema introduciendo el modelo UBM. Sin embargo, se plantea que, para diferentes locutores, la topología del modelo HMM tiende a ser la misma debido a la forma de

adaptación. Por ello, en este trabajo se proponen dos métodos no supervisados: UBM local y UBM global, capaces de obtener la topología de un HMM para cada locutor y así aumentar la capacidad discriminativa de cada modelo, lo que hace más robusta la verificación. Los resultados experimentales mostraron que ambos son efectivos para trabajar con pocas observaciones. El modelo UBM local trabaja muy bien bajo diferentes condiciones de entrenamiento. El modelo UBM global puede superar al HMM adaptado si los datos del entrenamiento tienen un volumen moderado. Una vez obtenida la topología del HMM por cualquiera de estos métodos, se realiza la adaptación MAP para refinar los parámetros del modelo.

En "*MAP and Sub-Word Level T-Norm for Text-Dependent Speaker Recognition*" [41], se presentan mejoras en el reconocimiento del locutor dependiente del texto utilizando como método de modelado los HMM y la adaptación MAP, y además, se proponen dos métodos para normalizar la puntuación en este tipo de sistema.

A partir de los coeficientes MFCC extraídos, se crean los modelos para un conjunto de fonemas, con el léxico fonético y la transcripción ortográfica. Se entrenaron previamente 39 modelos fonéticos HMM independientes del texto.

El modelo de la expresión dependiente del texto tienen la misma estructura que los HMM independientes del texto pero se pueden obtener de tres maneras diferentes: con la re-estimación Baulm-Welch de los HMM fonéticos independientes del texto, adaptando los mismos mediante MLLR o realizando la adaptación MLLR seguida de una adaptación MAP. En este sentido, la adaptación MAP introduce mejoras de un 22.6% en el EER.

Teniendo estos dos modelos, la expresión a verificar se alinea a ellos mediante el algoritmo Viterbi, obteniéndose puntuaciones para cada ventana. La puntuación final es la proporción entre el promedio de puntuación del modelo dependiente del texto obtenido por ventanas y el promedio de puntuación del modelo independiente del texto obtenido por ventanas.

Para normalizar la puntuación final se propusieron dos métodos nuevos: T-Norm a nivel de fonemas y T-Norm a nivel de estados, que normalizan las puntuaciones de segmentos similares antes de promediar las mismas. Estos segmentos pueden ser fonemas o estados del HMM, como se intuye de los nombres de los métodos.

Los experimentos demostraron que estos dos métodos trabajan mejor que el método clásico T-Norm a nivel de expresión, pues este último sufre de desigualdad en el léxico, cosa que no eliminan estas nuevas técnicas pero que sí reducen considerablemente. Se introducen en este aspecto mejoras de un 20 %en el EER.

5 Problemas por resolver aún en los métodos, detectados durante la realización del estudio del estado del arte

Durante el estudio realizado del "Estado del arte de los métodos de verificación texto-dependiente del locutor utilizando modelos acústicos del habla", se detectaron un grupo de problemas, aún no resueltos.

5.1 Gran volumen de datos a procesar (especialmente los LVCSR)

Los sistemas de reconocimiento del habla para vocabularios grandes (LVCSR) [60] han sido usados en tareas independientes del texto. Se reconocen de una conversación, las palabras más

frecuentes dichas por cada locutor y a estas se le aplican técnicas del reconocimiento del locutor dependiente del texto. Esto implica una mejora en los sistemas de reconocimiento del locutor porque se restringen a un grupo específico de unidades. Pero, a su vez, aumenta el costo computacional de ese sistema por la existencia del LVCSR, que debe procesar largas conversaciones.

5.2 Complejidad computacional

En la verificación del locutor dependiente del texto, el método que más se usa es el de modelado HMM, por su habilidad de modelar adecuadamente la gran variabilidad en el tiempo de la señal de voz. Esta técnica necesita de grandes cantidades de datos para el entrenamiento, en aras de lograr una buena estimación de los parámetros del modelo y un buen rendimiento por parte del reconocedor, además de que el diseño de los HMM puede ser muy complejo en dependencia del sistema. Por todo lo anteriormente dicho, posee un alto costo computacional.

5.3 Diferencias en el léxico entre entrenamiento y prueba

Este tema ha sido poco investigado hasta la fecha. La desigualdad en el léxico aparece cuando se usan diferentes expresiones para el entrenamiento y la prueba, por ejemplo, cadenas de dígitos diferentes. Estudios realizados [32, 42] en sistemas de verificación del locutor independiente y dependiente del texto mostraron que tanto el uso de N-gramas como las distribuciones de puntuación de los impostores proporcionan buen rendimiento en dichos sistemas.

En [35], se analiza el impacto de la desigualdad en el léxico, cuantificándolo mediante la distancia de Levenstein [61] entre la expresión de prueba y el modelo del locutor. Mientras mayor sea la distancia, mayor será el grado de la desigualdad en el léxico. Los resultados que se obtuvieron mostraron que este tipo de desigualdad es la causa fundamental de degradación de los sistemas de verificación del locutor.

Un reto a enfrentar es la reducción de la desigualdad en el léxico con medidas que establezcan el grado real del mismo.

5.4 Pocos datos para entrenar: Adaptación

La adaptación en línea del modelo del locutor [38, 39] es un componente central en cualquier aplicación exitosa de reconocimiento del locutor, especialmente en la tarea dependiente del texto debido a las cortas sesiones de entrenamiento.

Varios estudios del reconocimiento del locutor dependiente del texto [38, 39] han demostrado la efectividad de esta técnica. La adaptación es el proceso extendido de la sesión de entrenamiento a la sesión de prueba. La adaptación del modelo se puede hacer de dos formas: supervisada y no supervisada. La supervisada no utiliza información falsa, por lo que no hay posibilidad de que se corrompa el modelo del locutor. Permite la utilización de expresiones del mismo locutor que hayan obtenido malas puntuaciones, lo que tiene como ventaja el añadir al modelo nueva información. En el modo no supervisado, la decisión de utilizar la expresión de la voz de prueba para la adaptación es del sistema de verificación. El sistema decide basándose en la puntuación de la expresión de prueba [24, 62]. La desventaja de este modo es que pudiera existir algún impostor cuya puntuación sea alta y entonces el sistema actualizaría el modelo con

una información que no es del locutor. La cantidad de información desconocida del locutor se reduce debido a que las puntuaciones de esas expresiones poseen una puntuación baja y el sistema tiende a rechazar la actualización del modelo con dichas expresiones.

Un reto en este sentido sería definir un método de adaptación no supervisada en la que, a pesar de que la puntuación sea baja, pueda adaptarse el modelo del locutor con nueva información desconocida.

5.5 Segmentación automática en presencia de ruido: Voice Activity Detection

El rendimiento de los sistemas de reconocimiento del habla y del locutor se ve afectado actualmente por el ruido. La aplicación de algoritmos que detectan la actividad de voz "Voice Activity Detectors" (VAD) mejora el rendimiento de estos sistemas ante ambientes ruidosos. Los VAD detectan la presencia o ausencia de voz en la señal y la separan en segmentos de voz y silencio como se muestra en la figura 8. Construir un VAD es algo complejo y su dificultad aumenta en la medida que aumenta el ruido en la señal a analizar.

Fig. 8. Esquema de un VAD

Ante buenas condiciones del ambiente y el canal, se usan métodos como el cálculo del promedio de la energía de la señal en periodos cortos y el cálculo de la tasa de cruces por cero, que muestran muy buenos resultados [63]. Sin embargo, bajo condiciones ruidosas, los VAD tradicionales no son robustos dado que la señal está muy contaminada. Un reto a enfrentar es designar un VAD robusto a los ambientes ruidosos.

Son problemas aún no resueltos en el diseño de los VAD:

- Selección de características robustas al ruido.
- Estimación de la estadística del ruido.
- Selección del método de clasificación.
- Definición de la regla de decisión.
- Suavizado de la decisión final.

5.6 Diferencias de canal entre las etapas de entrenamiento y prueba: Normalización

Los sistemas de reconocimiento del locutor sufren severas degradaciones debido a las diferentes condiciones existentes en las etapas de entrenamiento y prueba. Estas diferencias pueden estar dadas por dos factores fundamentales: las variaciones intra-locutor (emociones, edad, salud) y las condiciones del ambiente de donde fue obtenida la señal (entorno, canal de transmisión).

La diferencia entre el entrenamiento y la prueba implica que las características acústicas de ambas señales sean diferentes, que existe una distorsión de la distribución a corto plazo de los rasgos y que los rasgos pudieran dañarse bajo estas condiciones. Además, ocasiona variabilidad en las puntuaciones que se utilizan para tomar la decisión de aceptar o no un cliente.

Para solucionar este problema se han propuesto varios métodos de compensación a las desigualdades del canal [64]:

Normalización de los rasgos:

- Sustracción de la Media Cepstral (CMS, por sus siglas en inglés)
- Filtro RASTA(versión de CMS que varía en el tiempo)
- Sustracción de la media y varianza cepstral (CMVN, por sus siglas en inglés)

Normalización de la puntuación:

- Z-Norm
- T-Norm
- H-Norm

Compensación de los rasgos:

- Mapeo de rasgos

Compensación del modelo:

- Síntesis de los modelos del locutor

Como se observa, hay muchos métodos propuestos para enfrentar las desigualdades de canal, cada uno de ellos enfocado hacia la normalización de algún resultado del proceso. Esto indica que el enfrentamiento a las desigualdades del canal es un reto a enfrentar en cualquier sistema de reconocimiento del locutor.

6 Conclusiones

La verificación del locutor dependiente del texto con rasgos acústicos es un tema que ha recibido mucha atención en el campo del reconocimiento del locutor, por la variedad de aplicaciones que posee. En este trabajo se realizó un estudio exhaustivo de la literatura relacionada con esta temática, y se estableció una clasificación de los métodos más utilizados. De ellos, los que mejores resultados han obtenido son los que modelan al locutor mediante HMM adaptados al locutor, ya sea por adaptación MAP o MLLR. Resultados obtenidos que se expusieron en este trabajo ponen en evidencia la superioridad de la adaptación MLLR sobre MAP, en tareas dependientes del texto, por su capacidad de adaptar el HMM a un locutor con pocos datos.

En este campo se ha investigado bastante y se han propuesto muchos métodos con el objetivo de lograr una verificación del locutor precisa, pero aún existen problemas abiertos que no tienen la solución más óptima, por ejemplo: los grandes volúmenes de datos a procesar, la complejidad computacional de los algoritmos de entrenamiento y prueba , las diferencias en el léxico en los datos de entrenamiento y los de la prueba, los pocos datos disponibles para el entrenamiento, las diferencias en los canales de entrenamiento y prueba y la detección de tramas con voz en presencia de ruido.

Por lo tanto, queda mucho que hacer para lograr un sistema capaz de reconocer una persona por su voz, que sea tan competente como el cerebro de una persona.

Referencias bibliográficas

1. Martin, A., et al., "The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspectives". Speech Commun, 2000. 31: p. 225–254.

2. Sturim, D.E., et al., "Speaker verification using text-constrained gaussian mixture models". Proc. IEEE ICASSP, 2002. 2002(1): p. 677–680.
3. Boakye, K. and B. Peskin, "Text-constrained speaker recognition on a text-independent task". Proc. Odyssey Speaker Recognition Workshop, 2004. 2004.
4. Heck, L.P. and M. Weintraub, "Handset dependent background models for robust text-independent speaker recognition". Proc. IEEE ICASSP, 1997. **1997**(2): p. 1037- 1040.
5. Higgins, A., L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting". Digit. Signal Process, 1991. **1**: p. 89-106.
6. Rosenberg, A.E. and S. Parthasarathy, "The use of cohort normalized scores for speaker recognition". Proc. IEEE ICASSP 1996. 1: p. 81-84.
7. Liu, Y., M. Russell, and M. Carey, "The role of dynamic features in text-dependent and -independent speaker verification". Proc. IEEE ICASSP 2006. 1: p. 669- 672.
8. Reynolds, D., "Channel robust speaker verification via feature mapping ". Proc. IEEE ICASSP 2003(2), 2003: p. 53- 56.
9. Teunen, R., B. Shahshahani, and L.P. Heck, "A model-based transformational approach to robust speaker recognition". Proc. ICSLP 2000(2), 2000: p. 495-498.
10. Duda, R.O., P.E. Hart, and D.G. Stork, "Pattern Classification". 2nd edn. ed. 2001: Wiley, New York.
11. Kato, T. and T. Shimizu, " Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence-preserving connected digit patterns". Proc. IEEE ICASSP 2003. 2: p. 57-60.
12. Matsui, T. and S. Furui, "Concatenated phoneme models for text-variable speaker recognition". Proc. IEEE ICASSP, 1993. 2: p. 391-394.
13. Che, C.W., Q. Lin, and D.S. Yuk, "An HMM approach to text-prompted speaker verification". Proc. IEEE ICASSP 1996. 2: p. 673-676.
14. Hébert, M. and L.P. Heck, "Phonetic class-based speaker verification". Proc. Eurospeech, 2003. Vol. 2003: p. 1665-1668.
15. Reynolds, D.A., T.F. Quatieri, and R. B.Dunn, "Speaker verification using adapted gaussian mixture models". Digit. Signal Process, 2000. 10: p. 19-41.
16. Gauvain, J.-L. and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains". IEEE T. Speech Audi. Process, 1994. **2**: p. 291-298.
17. Schmidt, M. and H. Gish, "Speaker identification via support vector classifiers". Proc. IEEE ICASSP, 1996. 1996(1): p. 105-108.
18. Campbel, W.M., et al., "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation". Proc. IEEE ICASSP 2006. 2006(1): p. 97-100.
19. Krause, N. and R. Gazit, "SVM-based speaker classification in the GMM model space". Proc. Odyssey Speaker Recognition Workshop, 2006. Vol. 2006.
20. Fine, S., J. Navratil, and R.A. Gopinath, "A hybrid GMM/SVM approach to speaker identification". Proc. IEEE ICASSP, 2001. 2001(1): p. 417- 420
21. Campbell, W.M., "A SVM/HMM system for speaker recognition". Proc. IEEE ICASSP 2003. 2003(2) p. 209-212.
22. Sankar, A. and R.J. Mammone, "Growing and pruning neural tree networks", in *IEEE Trans. Comput.* . 1993. p. 272-299.
23. Farrell, K.R., "Speaker verification with data fusion and model adaptation". Proc. ICSLP, 2002. 2002(2): p. 585-588.
24. Mirghafori, N. and L.P. Heck, "An adaptive speaker verification system with speaker dependent a priori decision thresholds ". Proc. ICSLP, 2002. 2002(2): p. 589-592.
25. Reynolds, D.A., "Comparison of background normalization methods for text-independent speaker verification". Proc. EuroSpeech 1997. 1997(2): p. 963-966.
26. Auckenthaler, R., M.J. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems". Digit. Signal Process. , 2000. 10: p. 42-54.

27. Hébert, M. and D. Boies, "T-Norm for text-dependent commercial speaker verification applications: effect of lexical mismatch.", in *Proc. IEEE ICASSP* 2005. p. 729–732.
28. Heck, L.P. and N. Mirghafori, "Online unsupervised adaptation in speaker verification". Proc. ICSLP, 2000. Vol. 2000.
29. Heck, L.P. "On the deployment of speaker recognition for commercial applications". in *Proc. Odyssey Speaker Recognition Workshop*. 2004.
30. Wadhwa, K., "Voice verification: technology overview anf accuracy testing results". Proc. Biometrics Conference, 2004. Vol. 2004.
31. Boies, D., M. Hébert, and L.P. Heck, "Study of the effect of lexical mismatch in text-dependent speaker verification". Proc. Odyssey Speaker Recognition Workshop, 2004. Vol. 2004.
32. Kato, T. and T. Shimizu, " Improved speaker verification over the cellular phone network using phoneme balanced and digit-sequence-preserving connected digit patterns". Proc. IEEE ICASSP 2003(2), 2003: p. 57–60.
33. Heck, L.P., et al., "Robustness to telephone handset distortion in speaker recognition by discriminative feature design". Speech Commun, 2000. 31: p. 181–192.
34. Siafarikas, M., et al., "Overlapping wavelet packet features for speaker verification ". Proc. EuroSpeech, 2005. 2005.
35. Boies, D., M. Hébert, and L.P. Heck, " Study of the effect of lexical mismatch in text-dependent speaker verification". Proc. Odyssey Speaker Recognition Workshop, 2004. Vol. 2004.
36. Hébert, M. and L.P. Heck, "Phonetic class-based speaker verification". Proc. Eurospeech, 2003. Vol. 2003: p. 1665–1668.
37. Reynolds, D., "Speaker identification and verification using Gaussian mixture speaker models". Speech Commun, 1995. 17: p. 91–108.
38. Fredouille, C., et al., "Behavior of a bayesian adaptation method for incremental enrollment in speaker verification". Proc. IEEE ICASSP, 2000. 2000.
39. Heck, L.P. and N. Mirghafori, "Online unsupervised adaptation in speaker verification,". Proc. ICSLP, 2000. 2000.
40. Stolcke, A., et al., "MLLR Transforms as Features in Speaker Recognition". *Proceedings of Eurospeech*, 2005: p. 2425-2428.
41. Toledano, D.T., et al., "MAP and sub-word level T-norm for text-dependent speaker recognition". Interspeech 2008, 2008.
42. Hébert, M. and N. Mirghafori, "Desperately seeking impostors: data-mining for competitive impostor testing in a text-dependent speaker verification system.". ICASSP, 2004.
43. Teunen, R., B. Shahshahani, and L.P. Heck, "A model-based transformational approach to robust speaker recognition". Proc. ICSLP 2000(2), 2000: p. 495–498.
44. Campbell, J.P., "Speaker Recognition: A Tutorial.". Proceedings of the IEEE, 1997. 85, No. 9,: p. 1437- 1462.
45. Rabiner, L.R. and B.-H. Juang, "*Fundamentals of Speech Recognition*". 1993: Prentice-Hall, Englewood Cliffs.
46. Rosenberg, A.E. and S. Parthasarathy, "Speaker background models for connected digit password speaker verification". Proc. ICASSP, 1996: p. 81-84.
47. Liu, M. and T.S. Huang, "Unsupervised learning of hmm topology for text-dependent speaker verification". Proc. of the International Conference on Spoken Language Processing, 2006.
48. Li, Q., et al., "Automatic Verbal Information Verification for User Authentication". IEEE Transactions on Speech and Audio Processing, 2000. Vol. 8: p. 585-596.
49. Sturim, D.E., et al., "Speaker verification using text-constrained Gaussian mixture models". Proceedings ICASSP '02 2002. Vol.1: p. 677-680.
50. Boakye, K. and B. Peskin, "Text-Constrained Speaker Recognition on a Text-Independent Task ". Odyssey 2004, 2004: p. 129-134.
51. Boakye, K., "Speaker Recognition in the Text-Independent Domain Using Keyword Hidden Markov Models". 2005, University of California at Berkeley.

52. BenZeghiba, M.F. and H. Bourlard, "User-customized password speaker verification using multiple reference and background models"*.* Speech communication 2006. Vol. 48, no9: p. 1200-1213.
53. Rodríguez, L., C. García, and J.L. Alba, "On combining classifiers for speaker authentication", in *Pattern Recognition*. 2003 p. 347-359.
54. Larcher, A., J.-F. Bonastre, and J.S.D. Mason, "Reinforced Temporal Structure Information For Embedded Utterance-Based Speaker Recognition"*.* Interspeech 2008, 2008.
55. Leggetter, C.J. and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models"*. Computer Speech and Language*, 1995. Vol. 9: p. 171-185.
56. Gales, M.J.F. and P.C. Woodland, "Mean and Variance Adaptation within the MLLR Framework"*. Computer Speech and Language*, 1996. Vol. 10(4): p. 249-264.
57. Elizalde, C.E., "Reconocimiento de Locutor dependiente de texto mediante adaptación de modelos ocultos de Markov fonéticos". 2007, Universidad Autónoma de Madrid.
58. Stolcke, A., et al., "Speech Recognition as Feature Extraction for Speaker Recognition"*.* Signal Processing Applications for Public Security and Forensics, 2007: p. 1-5
59. Ferras, M., et al., "Constrained MLLR for Speaker Recognition "*.* Proceedings of ICASSP, 2007 p. 53-56.
60. Moreno, P.J., et al., "Continuous Recognition of Large-Vocabulary Telephone-Quality Speech"*.* Proceedings of the Spoken Language Systems Technology Workshop, 1995: p. 70-73.
61. Levenshtein, V.I., "Binary codes capable of correcting deletions, insertions and reversals. "*.* Doklady Akademii Nauk SSSR, 1965. Vol. 163(4) p. 845–848.
62. Mirghafori, N. and M. Hébert, "Parametrization of the score threshold for a text-dependent adaptive speaker verification system"*.* Proc. IEEE ICASSP, 2004. Vol. 2004(1): p. 361–364.
63. ITU, I.-T.R., "A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70". 1996.
64. Wenjie, Z., "Cross-channel Text-independent and Text-dependent Speaker Verification", in *NeGSST 2006 Summer Seminar, 2006/8/17-18* 2006: National Taipei University of Technology.

Anexo 1 Desarrollo experimental de un método de identificación de palabras aisladas utilizando los HMM para su posterior utilización en un método de verificación dependiente del texto.

Se construyó un reconocedor de palabras aisladas, en este caso, los dígitos del 0 al 9 usando la herramienta HTK.

Pasos para construir el reconocedor:

1. Se creó la base de datos de entrenamiento: de la base de datos Ahumada se tomaron voces de 15 locutores de una sesión telefónica, donde cada uno repite dos veces y aleatoriamente los números del 0 al 9. Se obtuvieron 30 muestras de cada número para el entrenamiento. Las señales fueron etiquetadas según la palabra que contenía y los silencios, de la siguiente forma: sil + palabra + sil.
2. Se extrajeron los rasgos de dichas señales: Para cada trama se extraen los 12 primeros coeficientes MFCC, el coeficiente c0, que es proporcional a la energía total en la trama, 13 coeficientes delta y 13 coeficientes de aceleración, en total se obtuvieron 39 coeficientes por trama.
3. Se definió un prototipo de HMM izquierda-derecha de 6 estados, para cada elemento del vocabulario. Se modela cada palabra (uno, dos, tres, cuatro, cinco, seis, siete, ocho. nueve, cero y sil) con un HMM. El prototipo se muestra a continuación:

```
~o <VecSize> 39 <MFCC_0_D_A>
~h "cero"
<BeginHMM>
 <NumStates> 6
 <State> 2
  <Mean> 39
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  <Variance> 39
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
 <State> 3
  <Mean> 39
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  <Variance> 39
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
 <State> 4
  <Mean> 39
```

```
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  <Variance> 39
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
 <State> 5
  <Mean> 39
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
  <Variance> 39
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
  1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
 <TransP> 6
 0.0 0.5 0.5 0.0 0.0 0.0
 0.0 0.4 0.3 0.3 0.0 0.0
 0.0 0.0 0.4 0.3 0.3 0.0
 0.0 0.0 0.0 0.4 0.3 0.3
 0.0 0.0 0.0 0.0 0.5 0.5
 0.0 0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

A continuación se describe el significado de los parámetros del prototipo:
~o <VecSize> 39 <MFCC_0_D_A> : Encabezamiento del fichero, tamaño del vector de rasgos de cada trama y tipo de rasgo: MFCC, c0, delta y aceleración
~h "cero" <BeginHMM> (...)<EndHMM>: Contiene la descripción del modelo HMM llamado “cero”.
<NumStates> 6: Número total de estados en el HMM, incluyendo los dos estados que no emiten observaciones, el 1 y el 6.
<State> 2, 3, 4, 5: Introduce la descripción de la función de observación de cada estado. Se usan funciones de observación gausianas simples, con matrices diagonales. Tal función es descrita con un vector de media y otro de varianza. Los estados 1 y 6 no se describen porque no tiene función de observación.
<Mean> 39 0.0 0.0 (...) 0.0 (x 39): Vector de media de la función de observación actual. Cada elemento es inicializado arbitrariamente en 0, esto será modificado luego durante el proceso de entrenamiento.

<Variance> 39 1.0 1.0 (...) 1.0 (x 39): Vector de varianza de la función de observación actual. Cada elemento es inicializado arbitrariamente a 1,esto será modificado luego durante el proceso de entrenamiento.
<TransP> 6: Matriz de transición de 6 x 6 de cada estado del modelo HMM. Los valores nulos indican que esa transición no está permitida. Los otros valores se inicializan arbitrariamente, pero cada columna de la matriz debe sumar 1, esto será modificado luego durante el proceso de entrenamiento.

A partir de este prototipo se inicializó cada modelo HMM utilizando el algoritmo de alineamiento en el tiempo, Viterbi. Después se re-estimaron los valores óptimos de los parámetros del modelo inicializado. Este proceso se repitió tres veces para obtener un modelo HMM robusto de la palabra.

4. Se definió la gramática del reconocedor (descripción de lo que se va a reconocer) y el diccionario. La gramática quedó definida de la siguiente forma:

```
$WORD = UNO | DOS | TRES | CUATRO | CINCO | SEIS | SIETE | OCHO |
NUEVE | CERO;
( { START_SIL } [ $WORD ] { END_SIL } )
```

y el diccionario:

```
UNO       [uno]    uno
DOS       [dos]    dos
TRES      [tres]   tres
CUATRO    [cuatro] cuatro
CINCO     [cinco]  cinco
SEIS      [seis]   seis
SIETE     [siete]  siete
OCHO      [ocho]   ocho
NUEVE     [nueve]  nueve
CERO      [cero]   cero
START_SIL [sil]    sil
END_SIL   [sil]    sil
```

Los elementos de la izquierda se refieren a los nombres de las variables de la gramática. Los elementos de la derecha se refieren a los nombres de los modelos HMM. Los elementos entre corchetes indican los símbolos que serán la salida del reconocedor: los nombres de las etiquetas.

5. Se realizó el reconocimiento con datos de prueba.

Para la prueba se seleccionaron de la Base de datos Ahumada 5 muestras de voz por cada número del 0 al 9 de la misma sesión telefónica. A los mismos se le extrajeron los rasgos y se realizó la comparación con cada uno de los 10 modelos. El reconocedor sólo se confundió en una palabra de un conjunto de prueba de 50 muestras, obteniéndose un error de identificación del 2%.

Anexo 2 Principales instituciones, grupos de trabajo e investigadores que trabajan en estos métodos

1. Institución: International Computer Science Institute http://www.icsi.berkeley.edu/
- Grupo de trabajo: Speech Group, http://www.icsi.berkeley.edu/Speech/
- Investigadores: Kofi Boakye, fkaboakye@icsi.berkeley.edu
  Barbara Peskin, barbarag@icsi.berkeley.edu

2. Institución: MIT Massachusetts Institute of Technology
- Grupo de trabajo: Information Systems Technology, Lincoln Laboratory, http://www.ll.mit.edu/IST/
- Investigadores: Joseph. P. Campbell, jpc@ll.mit.edu
  William M. Campbell, wcampbell@ll.mit.edu
  Robert B. Dunn, rbd@ll.mit.edu
  Douglas A. Jones, daj@ll.mit.edu
  Douglas Reynolds, dar@sst.ll.mit.edu
  Pedro A Torres Carrasquillo, ptorres@sst.ll.mit.edu
  Thomas F. Quatieri, tfq@sst.ll.mit.edu
  Douglas Sturim , sturim@sst.ll.mit.edu
3. Institución: University of Avignon, http://www.univ-avignon.fr/
- Grupo de trabajo: Laboratorie Informatique Avignon , LIA, http://www.lia.univ-avignon.fr
- Investigadores: Jean-François Bonastre, jean-francois.bonastre@liauniv-avignon.fr
  - Corinne Fredouille, corinne.fredouille@lia.univ-avignon.fr
4. Institución: DalleMolleInstitut e for Perceptual Artificial Intelligence, IDIAP, http://www.idiap.ch
- Investigadores: Herve Bourlard, bourlard@idiap.ch
  Johnny Mariethoz, marietho@idiap.ch
5. Institución: Eurecom Institute
- Grupo de trabajo: Deparment of Multimedia Communications
- Investigador: Mohamed Faouzi BenZeghiba.
6. Institución: University of Illinois at Urbana-Champaign, Beckman Institute http://beckman.uiuc.edu/index.aspx
- Investigadores: Ming Liu, mingliu1@ifp.uiuc.edu
  Thomas Huang, huang@ifp.uiuc.edu
7. Institución: Universidad Autónoma de Madrid
- Grupo de trabajo: Área de Tratamiento de Voz y señales http://atvs.ii.uam.es/
- Investigadores: Doroteo T. Toledano, doroteo.torres@uam.es
  Daniel Hernandez-Lopez, daniel.hernandez@uam.es
  Cristina Esteve-Elizalde, cristina.esteve@uam.es
  Ruben Fernandez Pozo, ruben.fernandez@uam.es
  Luis Hernandez Gomez, luis.hernadez@uam.es
8. Institución: SRI International
- Grupo de trabajo: Speech Technology and Research Laboratory

- Investigadores: Andreas Stolcke, fstolcke@speech.sri.com
  Luciana Ferrer, lferrer@speech.sri.com
  Sachin Kajarekar, sachin@speech.sri.com
  Elizabeth Shriberg, ees@speech.sri.com
  Anand Venkataraman, anandg@speech.sri.com

9. Institución: Network ASR Core Technology Nuance Communications
- Investigador: Matthieu Hébert, hebert@nuance.com
10. Institución: Tokyo Institute of Technology Street
- Grupo de trabajo: Department of Computer Science
- Investigador: Sadaoki Furui, furui@cs.titech.ac.jp

Anexo 3 Relación de las principales publicaciones: revistas, páginas web y libros que tratan el tema

Revistas

1. Speech Communication, www.elsevier.nl/locate/specom
2. Journal of the Acoustical Society of America, http://asa.aip.org/jasa.html
3. IEEE Transactions on Speech and Audio Processing, http://www.ieee.org/portal/pages/pubs/transactions/tsap.html
4. IEEE Transactions on Audio, Speech & Language Processing, http://www.ewh.ieee.org/soc/sps/tap/index.html
5. Computer, Speech & Language, http://www.sciencedirect.com/science/journal/08852308
6. Digital Signal Processing http://www.idealibrary.com/links/toc/dspr/10/1/0
7. Pattern Recognition, http://www.sciencedirect.com/science/journal/00313203
8. Pattern Recognition Letters, http://www.sciencedirect.com/science/journal/01678655
9. International Journal of Pattern Recognition and Artificial Intelligence, http://www.worldscinet.com/ijprai/ijprai.shtml

Páginas web

1. NIST Speaker Recognition Benchmarks, http://www.nist.gov/speech/tests/spk/index.htm
2. Joe Campbell's Site for Speaker Recognition Speech Corpora, http://www.apl.jhu.edu/Classes/Notes/Campbell/SpkrRec/
3. The Biometric Consortium, http://www.biometrics.org/
4. Comp.Speech FAQ on speaker recognition, http://www.speech.cs.cmu.edu/comp.speech/Section6/Q6.6.html
5. International Speech Communication Association, ISCA, http://www.isca-speech.org
6. ISCA on-line archive, http://www.isca-speech.org/archive.html
7. IEEE Signal Processing Society, http://www.ewh.ieee.org/soc/sps/

Libros

1. Survey of the State of the Art of Human Language Technology. Ronald A. Cole, http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html,
2. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Xuedong Huang, Hsiao-Wuen Hon, Prentice Hall 2001, ISBN: 0130226165
3. Automatic Speech and Speaker Recognition: Advanced Topics. Chin-Hui Lee, Frank K. Soong, Kuldip K. Paliwal, KLUWER 1996, ISBN:0-7923-9706-1
4. Discrete-Time Speech Signal Processing: Principles and Practice. Thomas F. Quatieri, Prentice Hall 2002, ISBN: 0-13-242942-X
5. Discrete-Time Processing of Speech Signals. John R. Deller. John H. L. Hansen, John G. Proakis

6. The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department,2005
7. The Scientist and Engineer's Guide to Digital Signal Processing. Steven W. Smith, California Technical Publishing 1997, ISBN 0-9660176-3-3
8. Pattern Recognition in Speech and Language Processing. Wu Chou, Biing Hwang Juang, Georgia Institute of Technology, CRC Press
9. Pattern Classsification, Richard O. Duda, PeterE. Hart,David G.Stork, Wiley-Interscience, ISBN 0-471-05669-3
10. Fundamentals of Speech Recognition. Lawrence Rabiner, Biing Hwang Juang. Prentice-Hall International, Inc
11. Springer Handbook of Speech Processing. Jacob Benesty, M. Mohan Sondhi, Yiteng Huang, Springer-Verlag Berlin Heidelberg 2008, ISBN: 978-3-540-49125-5

Anexo 4 Herramientas de Software que tratan el tema

1. CSLU Toolkit

Conjunto de herramientas para la exploración, el conocimiento y la investigación en la voz y la interacción " human-computer", http://cslu.cse.ogi.edu/toolkit/

2. LNKnet MIT Lincoln Laboratories Pattern Classification Software

LNKnet es un paquete de software que integra más de 22 redes neuronales, clasificadores con máquinas de aprendizaje, métodos de agrupamientos y algoritmos de selección de rasgos, máquinas de soportes vectoriales y clasificadores bayesianos sencillos. Tiene una versión para Linux y una para Windows usando el ambiente de Cygwin, las herramientas están creadas en lenguaje C, http://www.ll.mit.edu/IST/lnknet/.

3. Becars Library and Tools for Speaker Verification (version 1.1.9)

Desarrollada por la Universidad de Balamand (Líbano) y la École Nationale Supérieure des Télécommunications (GET-ENST Paris, France).

Sistema de verificación del locutor, que provee una librería desarrollada en C, así como varias herramientas que permiten obtener un GMM, realizar la clasificación y obtener la puntuación correspondiente a la comparación. Incluye una implementación del algoritmo EM con diferentes tipos de criterios: Maximización de la Verosimilitud (ML), Maximización a Posteriori (MAP), y Regresión Lineal de Máxima Verosimilitud (MLLR).

Ha participado exitosamente en varias evaluaciones NIST.

Plataformas: puede ser compilado sobre UNIX/Linux y Windows.

Principales desventajas: Su poca actualidad (abril 2005) y no contiene los códigos fuentes.

4. HTK Versión 3.4

Paquete de herramientas y funciones para el desarrollo de aplicaciones de procesamiento del habla, especialmente de reconocimiento, a través del uso de los modelos ocultos de Markov (HMM). El sistema está conformado por dos etapas fundamentales. En la primera de entrenamiento, se estiman los parámetros del conjunto de modelos HMM usando expresiones de entrenamiento y sus correspondientes transcripciones. En la segunda etapa, las expresiones desconocidas son transcritas usando las herramientas de reconocimiento del HTK. http://htk.eng.cam.ac.uk/. y http://mi.eng.cam.ac.uk/~sjy/software.htm.

Esta herramienta cuenta con las siguientes ventajas:

- Código fuente disponible (C++).
- Documentación tanto teórica como de la aplicación en sí, lo cual permite su uso tanto en un proyecto teórico como aplicado.
- Esquema completo de reconocimiento del habla.
- Actualidad (2006).

Desventajas: No incluye una implementación para el reconocimiento del locutor.

5. Speaker Recognition Tools

Este paquete contiene un conjunto de utilidades para la investigación en el campo del reconocimiento del locutor, incluyendo clasificación con GMM, Cuantización Vectorial (VQ), y redes neuronales MLP. Puede usarse también como un clasificador general.

Desarrollado en C++. Plataforma: SUN SPARC (SunOS), PC (MSDOS).

Esta herramienta puede ser de utilidad en desarrollos teóricos, no así en aplicaciones, donde los algoritmos deben ser visibles.

6. CMU Sphinx

El Grupo Sphinx de la Universidad Carnegie Mellon comprende un amplio proyecto con el objetivo de estimular la creación de herramientas y aplicaciones relacionadas con el habla. Sus líneas de trabajo principales son: reconocimiento del habla, sistemas de diálogo, y síntesis de voz. No presenta restricciones respecto a su uso comercial o redistribución.

Plataformas: GNU/Linux, variantes de Unix, y Windows NT o posterior.

http://cmusphinx.org/

7. Alize

Alize es una plataforma de software dirigida a facilitar el desarrollo de aplicaciones en el área del reconocimiento tanto del habla como del locutor. Esta ha sido desarrollada en el Laboratorio de Informática de Avignon (LIA) desde 2003. Alize está compuesta de dos niveles distintos:

- Nivel base, donde se encapsula la complejidad técnica de los módulos (adquisición de datos, cálculo, almacenamiento, etc). Este nivel evita que el usuario tenga que administrar la memoria directamente.
- En un segundo nivel se incluyen las utilidades y algoritmos que se manipulan por el usuario (listas de administración, inicialización de modelos, algoritmos MAP, etc.).

Alize presenta una documentación muy bien detallada para su uso, también tiene las siguientes características:

- Tiene un nivel de funcionamiento que corresponde con el estado del arte actual, en términos de error pero también en términos de recursos de cómputo necesarios.
- Facilita el desarrollo de demostraciones y aplicaciones prácticas.
- Está programada en C++.

http://www.lia.univ-avignon.fr/heberges/ALIZE/

8. MASV - Munich Automatic Speaker Verification System

Sistema diseñado para la verificación del locutor. Implementa distintos tipos de modelos que incluyen HMM y GMM. Diferentes técnicas de normalización de la puntuación están implementadas: UBM, normalización simple de cohorte y normalización H-Norm.

Principal Desventaja: La programación del sistema sobre Perl y Matlab, lo cual limita la escalabilidad y las prestaciones del sistema.

RT_009, noviembre 2009
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: DCG Matilde Galindo Sánchez
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. No. 21812 e/218 y 222, Rpto. Siboney, Playa;
Ciudad de La Habana. Cuba. C.P. 12200
Impreso en Cuba