Eficacia de las características prosódicas a corto plazo en la verificación de locutor

Iker Luengo[1], Eva Navas[1], Inmaculada Hernáez[1], Jon Sanchez[1], Ibon Saratxaga[1], Iñaki Sainz[1], Juan J. Igarza[1]

[1] Aholab – Escuela Técnica Superior de Ingeniería.
Universidad del País Vasco – Euskal Herriko Unibertsitatea
Alda. Urquijo s/n 48013 Bilbao

ikerl@bips.bi.ehu.es,{eva.navas, inma.hernaez, jon.sanchez, ibon.saratxaga}@ehu.es, inaki@bips.bi.ehu.es, juanjo.igarza@ehu.es

Abstract. En este trabajo combina un sistema tradicional de verificación de locutor basado en MFCC con otro basado en prosodia a fin de determinar si la información prosódica a corto plazo es útil para mejorar el rendimiento de los actuales sistemas automáticos de identificación de locutor. El sistema de verificación tradicional basado en información espectral tiene una tasa de equierror (EER) del 3,85% utilizando 1024 mezclas gaussianas. El sistema basado en prosodia utiliza información de la entonación y energía a corto plazo y produce con 128 mezclas una EER del 23,93%. Tras aplicar LDA y fusionar las tasas de reconocimiento se obtiene una EER final de 3,84%. Este resultado no muestra una mejora significativa respecto al sistema de verificación de locutor tradicional.

1 Introducción

El uso de transacciones a larga distancia se ha extendido en los últimos años: compras a través de Internet, transacciones bancarias basadas en Web, restricción de acceso remoto a zonas seguras en ordenadores… Todos estos sistemas precisan de algún tipo de procedimiento de autenticación de cara a verificar la identidad de los usuarios. La mayoría de ellos recurre a la identificación basada en palabras clave, pero éstas pueden ser olvidadas o robadas.

Hoy por hoy la biometría es la mejor alternativa a este modo de identificación. Las características biométricas no pueden olvidarse o perderse y son difíciles de imitar. Este tipo de autenticación puede ser visto ya en múltiples aplicaciones: los ordenadores portátiles con acceso controlado mediante huella digital o el acceso a áreas restringidas mediante reconocimiento de la geometría de la mano son algunos de los ejemplos más comunes.

El creciente interés en los sistemas de identificación biométricos automáticos se refleja en el incremento de certámenes de competición de sistemas de verificación biométrica tales como el Fingerprint Verification Competition [1] o las Speaker Recognition Evaluations del NIST [2], en los cuales se proponen a evaluación nuevos algoritmos y métodos con el fin de mejorar los resultados actuales. La verificación

automática de locutor no es una excepción. La mayoría de las soluciones actuales utilizan características de la envolvente espectral para parametrizar la voz (MFCC, LPCC...) obteniendo muy buenos resultados [3][4][5]. Las investigaciones recientes intentan incluir información prosódica al objeto de reducir las tasas de error.

La prosodia del habla se refiere a la entonación, energía y velocidad del habla. Es bien sabido que estos rasgos son característicos de cada persona y por tanto aportan información acerca del locutor. Más aún, la prosodia no está correlada con la forma de la envolvente espectral. Por tanto, añadir dicha información a las características espectrales ya empleadas puede llevar a una mejora de los resultados de los sistemas.

La mayoría de los trabajos en este área se centran en el uso de información prosódica a largo plazo [6][7][8] y la fusión de sus resultados con los de los sistemas convencionales. Otros intentan utilizar valores prosódicos tales como la entonación y curvas de potencia muestreados por ventana [9][10]. Este último planteamiento resulta muy interesante ya que las nuevas características pueden ser combinadas fácilmente con los coeficientes cepstrales tradicionales.

Este trabajo se centra en determinar hasta qué punto la información prosódica a corto plazo es útil de cara a mejorar los sistemas automáticos de reconocimiento de locutor actuales, para lo cual, se presenta un nuevo sistema que utiliza tanto características espectrales como prosódicas. El artículo está organizado del siguiente modo: en primer lugar se describe el sistema de verificación desarrollado, a continuación analiza la base datos utilizada en los experimentos, y se finaliza con la descripción de los experimentos y sus resultados.

2 Descripción del sistema de verificación

2.1 El sistema de referencia

El sistema de partida consiste en un sistema tradicional basado en modelos de mezclas gaussianas (*Gaussian Mixture Models*, GMM) combinado con un modelo de locutor universal (*Universal Background Model*, UBM) [11] que utiliza como parámetros los Coeficientes Cepstrales de Frecuencia en escala Mel (*MEL Frequency Cepstral Coefficients*, MFCC). Cada 10 milisegundos se obtiene un vector de 18 parámetros MFCC. Los vectores se amplían con las derivadas de primer y segundo orden de los parámetros. Al objeto de reducir los efectos de canal se aplica resta de la media cepstral (*Cepstral Mean Subtraction*, CMS) [12].

Los modelos del locutor fueron entrenados mediante adaptación máxima a posteriori (*Maximum A Posteriori*, MAP) del UBM previamente entrenado [3]. Sólo se adaptaron las medias, dejando las varianzas y pesos inalterados. Finalmente, como es usual en estos sistemas, se aplicó normalización UBM y HNorm a las verosimilitudes.

2.2 El sistema prosódico

En el sistema prosódico se utiliza información relacionada con la entonación y la energía. Se crearon modelos separados para el modelado prosódico de regiones sonoras y sordas para manejar las discontinuaciones de la curva de entonación. La potencia de la señal se estima cada 10 milisegundos mediante ventanas de Hamming de 30 milisegundos de duración. También se estima el valor de la frecuencia fundamental o F0 cada 10 milisegundos mediante un método basado en la transformación cepstrum y el algoritmo de Viterbi. Este método no sólo calcula el valor de F0, sino que además decide si la trama es sonora o sorda. Una vez estimadas las curvas de potencia y entonación se calculan sus primeras y segundas derivadas para tener en cuenta su dinámica.

Las tramas sonoras y sordas son separadas para obtener dos flujos de vectores de parámetros. Las tramas sonoras son parametrizadas con cinco características (F0 instantánea, su primera y segunda derivada, y la primera y segunda derivada de la potencia), mientras que para las tramas sordas se utilizan sólo dos (la primera y segunda derivada de la potencia). La potencia instantánea se descarta en ambos casos pues su valor está más relacionado con la ganancia del canal que con la identidad del locutor.

Utilizando estos dos flujos de vectores se entrenan dos modelos por locutor utilizando el esquema tradicional GMM-UBM. Primero se desarrollaron dos modelos UBM, uno para tramas sonoras y otro para tramas sordas. Después se crean los modelos a partir de ellos mediante adaptación MAP.

Durante la fase de prueba se calculan dos puntuaciones por grabación, una para los flujos de tramas sonoras y otra para las sordas. A estas puntuaciones se les aplica normalización UBM antes de fusionarlos con la regla del producto. Esto es, la puntuación final de la información prosódica es el producto de las puntuaciones de los flujos de tramas sonoras y sordas.

2.3 Fusión de los dos sistemas

Para combinar los resultados de ambos clasificadores se ha utilizado un esquema de fusión tardía de expertos: en primer lugar se calculan las puntuaciones para el sistema tradicional y para el sistema basado en prosodia, y a continuación se obtiene la puntuación final combinando ambas puntuaciones.

Para ello se seleccionó el algoritmo de discriminación lineal (*Linear Discriminant Algorithm*, LDA) [14]. LDA es capaz de encontrar la combinación lineal de puntuaciones que mejor separa las puntuaciones de usuarios e impostores, pero debe ser entrenado sobre un conjunto de locutores de validación antes de aplicarlo a las pruebas finales.

3 Descripción de la base de datos

Los experimentos llevados acabo han sido realizados sobre la base de datos AHUMADA [15]. Esta base de datos consiste en grabaciones de 103 locutores masculinos españoles y fue grabada específicamente para el desarrollo de sistemas de automáticos de reconocimiento de locutor. De hecho, fue utilizada en las campañas de evaluación de NIST de 2000 y 2001 [2], junto con una ampliación que incluía también locutoras.

Aunque la base de datos completa contiene tanto grabaciones con micrófono de alta calidad como grabaciones telefónicas, sólo se han utilizado estas últimas en los experimentos. Esto permite capturar los efectos que la distorsión de canal tiene sobre el sistema. Las grabaciones telefónicas se llevaron a cabo en tres sesiones, en cada una de las cuales se utilizaron diferentes teléfonos.

- La primera sesión (llamada T1), fue grabada a través de una llamada interna con respuesta en frecuencia plana, por lo tanto, no hay distorsión en las señales.
- Durante la segunda sesión (T2), las grabaciones se realizaron desde el teléfono de la casa del locutor, por lo que el tipo de aparato utilizado se desconoce.
- En la tercera sesión (T3) cada locutor utilizó uno de los nueve terminales telefónicos disponibles en el laboratorio de grabación, y por lo tanto, se conoce el tipo de teléfono utilizado.

Entre los elementos grabados en cada sesión se han seleccionado dos para los experimentos: el texto de lectura común (el mismo texto para todos los locutores y sesiones) y el texto de lectura específico (un texto diferente por locutor y sesión). A partir de aquí estos elementos serán mencionados como C (para Común) y S (para eSpecífico) respectivamente. Todas estas grabaciones fueron muestreadas a 8 Khz. y 16 bits por muestra. La longitud media de los elementos seleccionados es de alrededor de 65 segundos.

Resumiendo, para cada experimento se dispone de tres sesiones telefónicas de 103 locutores varones con dos elementos por sesión. Este sub-corpus fue dividido como sigue: 51 locutores fueron reservados para entrenar los modelos UBM, 26 fueron utilizados para las pruebas de validación y el resto se mantuvieron como usuarios del sistema. Cada usuario fue empleado como impostor para el resto de los usuarios, lo que equivale a disponer de 25 impostores. Los locutores fueron designados aleatoriamente para cada grupo, al objeto de no sesgar los resultados.

El entrenamiento de los modelos de los locutores se realizó con las tareas C y S de la sesión T1. Como esta sesión se grabó sin distorsión espectral los modelos resultantes son independientes de canal. Las tareas C y S de la sesión T3 fueron utilizadas para las pruebas de desarrollo, ya que al conocer el tipo de teléfono usado es posible calcular los parámetros de la normalización HNorm para cada usuario. Finalmente la tarea S de la sesión T2 fue reservada para la prueba final. De este modo, las pruebas finales fueron llevadas a cabo con un terminal desconocido por el sistema, ya que no había sido utilizado con anterioridad, ni durante el entrenamiento, ni en el desarrollo. Además, utilizando sólo la tarea S, también el contenido de las grabaciones es desconocido por el sistema.

Para el entrenamiento se utilizaron todas las grabaciones. Esto genera aproximadamente 130 segundos de material de entrenamiento por locutor. Para las pruebas se separaron nueve extractos o segmentos de cada elemento de unos 10 segundos de duración cada uno, permitiendo solapes entre extractos consecutivos de hasta el 50%. Por tanto, se dispone de 9 segmentos de validación por usuario, lo cual significa que cada usuario fue testeado frente a 25x9=225 segmentos de impostores.

Todo ello hace que el diseño de los experimentos sea muy realista, similar al caso de un sistema en el que los usuarios hacen en el laboratorio las grabaciones para el entrenamiento y desarrollo, pero intentan acceder desde su casa u oficina utilizando su propio teléfono.

4 Resultados

En un sistema GMM el número de mezclas de gaussianas es crítico de cara a la precisión del sistema. Para estimar el número de mezclas para los modelos se han desarrollado GMMs de diferente orden, desde 2 hasta 1024 mezclas, tanto para el sistema MFCC como para los sistemas basados en prosodia. El número final de mezclas fue seleccionado para minimizar la tasa de equierror (*Equal Error Rate*, EER) entre la tasa de falso rechazo (*False Rejection Rate*, FRR) y la tasa de falsa aceptación (*False Aceptation Rate*, FAR). La tabla 1 muestra las tasas ERR de los sistemas entrenados. Como se esperaba, el sistema inicial produce mejores resultados que el prosódico solo. También se puede ver cómo la información de la energía de las ventanas sordas no contribuye positivamente al resultado final del sistema de verificación basado en prosodia.

Tabla 1. Tasas %EER del sistema inicial y de los sistemas basados en prosodia.

Nº mezclas	*2*	*4*	*8*	*16*	*32*	*64*	*128*	*256*	*512*	*1024*
S. Espectral	29.13	20.29	17.18	13.25	10.36	8.12	6.75	5.71	5.21	3.85
Sonoras	35.47	33.21	29.88	28.96	27.78	25.32	23.93	24.03	24.74	23.93
Sordas	46.53	49.45	47.86	45.86	44.44	43.16	42.74	43.18	43.59	44.44
Sonoras+Sordas	35.27	32.48	29.49	28.77	27.33	25.64	23.93	24.89	24.53	24.14

Conforme a las EER obtenidas, se seleccionaron GMMs de 1024 y 128 mezclas para los sistemas espectral y prosódico respectivamente. Tras aplicar LDA y fusionar las puntuaciones, se obtiene una EER final de 3,84%. La figura 1 muestra las curvas DET [16] de los sistemas antes y después de aplicar la fusión. Las curvas DET del sistema inicial y el combinado son bastante similares. De hecho, ambos sistemas presentan la misma tasa de equierror.

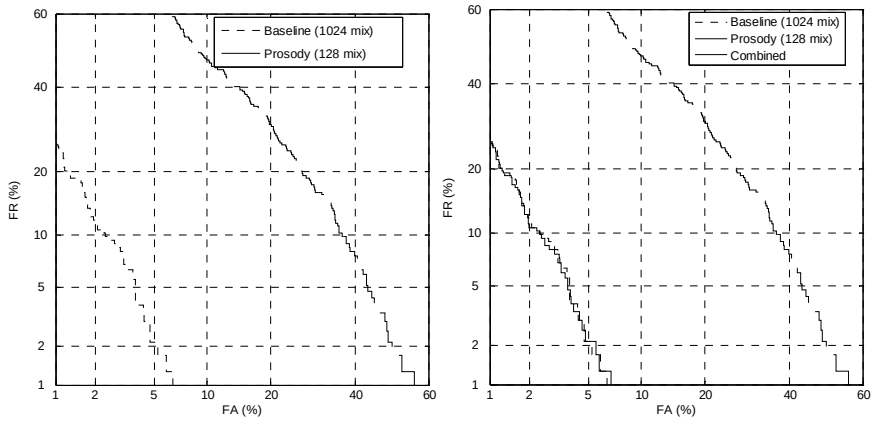

Figura. 1. Curvas DET del sistema inicial y del sistema basado en prosodia (izquierda) y sistema final combinado (derecha).

5 Conclusiones

Como resultado de estos experimentos se observa que la inclusión de características sencillas relativas a información de entonación y energía a corto plazo no mejoran los resultados de los sistemas actuales. Debido al gran avance experimentado en los últimos años en el campo de la verificación automática de locutor (como las técnicas de normalización de los aparatos telefónicos y de verosimilitud) los sistemas actuales basados en la envolvente espectral consiguen resultados muy superiores a los basados en prosodia (3,85% frente a 23,93% EER en estos experimentos).

Esto no quiere decir que la prosodia no sea útil para los sistemas automáticos de reconocimiento. Tal como las investigaciones recientes apuntan se observa cierta mejora [9][10], pero los sistemas basados en prosodia se encuentran aún muy lejos de los basados en el espectro. Por ejemplo, mientras en un sistema tradicional la normalización HNorm es suficiente para resolver la variabilidad de terminales telefónicos, en prosodia es necesario resolver la variabilidad intersesión. Todavía queda por explorar una gran tarea en este interesante campo de investigación.

6 Agradecimientos

Este trabajo ha sido financiado parcialmente por del Ministerio de Ciencia y Tecnología (TIC2003-08382-C05-03).

7 Referencias

1. FVC2006 web site: http://bias.csr.unibo.it/fvc2006/
2. NIST Speaker Recognition Evaluations' web site: http://www.nist.gov/speech/tests/spk/
3. D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.
4. M. A. Przybocki and A. F. Martin, "NIST Speaker Recognition Evaluation Chronicles," presented at Odyssey, Toledo, España, 2004.
5. A. F. Martin and M. A. Przybocki, "The NIST Speaker Recognition Evaluations: 1996-2001," presented at Speaker Odyssey 2001, Creta, Grecia, 2001.
6. M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust Prosodic Features for Speaker Identification," presented at ICSLP, Philadelphia, EEUU, 1996.
7. A. G. Adami, R. Michaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," presented at ICASSP, Hong Kong, 2003.
8. B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. A. Reynolds, and B. Xiang, "Using Prosodic and Conversational Features for High Performance Speaker Recognition," presented at ICASSP, Hong Kong, 2003.
9. D. A. Reynolds, W. Andrews, J. P. Campbell, J. Navratil, B. Peskin, A. G. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Michaescu, J. J. Godfrey, J. Douglas, and B. Xiang, "SuperSID project final report," SuperSID project (http://www.clsp.jhu.edu/ws2002/groups/supersid/) 2002.
10. M. Arcienega and A. Drygajlo, "A Bayesian Network Approach for Combining Pitch and Spectral Envelope Features for Speaker Verification," presented at COST275 workshop, Roma, Italia, 2002.
11. F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," EURASIP Journal on Applied Signal Processing, vol. 4, pp. 430-451, 2004.
12. C. Barras and J. L. Gauvain, "Feature And Score Normalization for Speaker Verification of Cellular Data," presented at ICASSP, Hong Kong, 2003.
13. R. Auckenthaler, M. J. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," Digital Signal Processing, vol. 10, pp. 42-54, 2000.
14. J. B. Kennedy and A. M. Neville, Basic Statistical Methods for Engineers and Scientists, Harper & Row, New York, 1986.
15. J. Ortega-García, J. González-Rodríguez, and V. Marrero-Aguiar, "AHUMADA: A Large Speech Corpus in Spanish for Speaker Characterization and Identification," Speech Communication, vol. 31, pp. 255-264, 2000.
16. A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET Curve in Assessment of Detection Task Performance," presented at Eurospeech, Rhodes, Grecia, 1997.