Identificación de locutores en entornos multilingües

Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Igor Odriozola, Inmaculada Hernaez

Universidad del País Vasco, Alda. Urquijo s/n,
48013 Bilbao, Spain
{ikerl, eva, inaki, ibon, ion, igor, inma}@aholab.ehu.es

Abstract: Los sistemas de identificación y verificación de locutor tienen resultados pobres cuando el modelo se entrena en un idioma mientras que las pruebas se realizan en otro. Esta situación es bastante común en entornos multilingües, en donde los usuarios deberían poder utilizar el sistema en el idioma que prefieran en cada momento, sin notar una reducción en la fiabilidad del mismo. En este trabajo se estudia la posibilidad de utilizar parámetros derivados de características prosódicas con el objetivo de reforzar la independencia del idioma de estos sistemas. Un análisis previo de las características de los parámetros en términos de variabilidad frente al idioma y la sesión predice un incremento en la robustez frente al idioma cuando parámetros MFCC tradicionales se combinan con valores de energía y entonación extraídos para cada trama. Los resultados experimentales confirman que estos parámetros proporcionan una mejor tasa de reconocimiento de locutor cuando entrenamiento y prueba se realizan con idiomas diferentes.

1 Introducción

En los últimos años varios grupos de investigación han centrado su atención en los sistemas de reconocimiento de locutor en entornos multilingües, donde puede ocurrir que los modelos de locutor se entrenen en un idioma pero que se usen en otro. Trabajos como los realizados por Faúndez y Satué-Villar [1] y Durou [2] demuestran que hay una reducción en la precisión del sistema en estas condiciones, pero no aportan alternativas para aliviar el problema. Otros trabajos como los de Akbacak y Hansen [3] y Ma y Meng [4] proponen algún tipo de solución, pero sus propuestas siempre implican conocer de antemano los posibles idiomas que van a ser utilizados, lo cual no siempre es posible.

En este trabajo se trata de encontrar una solución a nivel de parámetros, es decir, buscar una parametrización que ayude a mantener la tasa de acierto bajo condiciones de desadaptación de idiomas. Al ser una solución a nivel de parámetros, debería ser completamente generalizable a cualquier idioma no visto durante el entrenamiento. Para ello, el presente estudio se centra en el País Vasco, en donde coexisten dos idiomas oficiales: castellano y euskera. Ambos idiomas tienen muy poco en común, ya que el euskera no es un idioma indo-europeo como el castellano. De hecho, el euskera está considerado un idioma aislado, que no tiene ninguna relación con ninguna otra

lengua viva o muerta. Esto proporciona una situación real que puede considerarse como el peor caso para un sistema de reconocimiento de locutor, ya que las diferencias existentes son mucho mayores que las puramente dialectales.

2 Definición del Problema

2.1 Identificación de Locutores en Desadaptación de Idiomas

El método más habitual de diseñar sistemas de reconocimiento de locutor es utilizar modelos de mezclas gaussianas (GMM) [5] para modelar la distribución de parámetros espectrales a corto plazo, tales como MFCC y LPCC [6][7]. Estos parámetros espectrales caracterizan el filtro que modela el tracto vocal de cada locutor en el momento de articulación, capturando por tanto no sólo las características del tracto vocal (permitiendo por tanto la identificación del locutor), sino también las características del tracto vocal para cada fonema. Esto supone que este tipo de parametrizaciones contienen también información acerca del contenido fonético de la locución.

En un sistema de reconocimiento de locutores independiente del texto los problemas surgen cuando, en un entorno multilingüe, el modelo se entrena en un idioma pero las pruebas se realizan en otro. Normalmente el contenido fonético de ambos idiomas no coincide, por lo que los scores de las locuciones de prueba no serán fiables, incrementando la tasa de error del sistema.

2.2 Solución Propuesta

Un método inmediato para reducir la discrepancia entre las locuciones de prueba y el modelo es realizar el entrenamiento con grabaciones en ambos idiomas. De esta forma, es probable que el modelo consiga aprender las características de todos los fonemas. Esta solución es la adoptada por Ma y Meng en su trabajo [4]. Otra posible vía es entrenar un modelo diferente para cada locutor e idioma, y usar un detector de idioma para decidir qué modelo utilizar durante la prueba, tal y como proponen Akbacak y Hansen [3]. Pero este tipo de soluciones requiere conocer de antemano los idiomas que van a ser utilizados, ya que no son generalizables a idiomas no vistos durante la fase de entrenamiento. Por tanto, sería conveniente disponer de una solución más independiente de idioma.

En los últimos años se ha propuesto e implementado con éxito el uso de parámetros de alto nivel para problemas de reconocimiento de locutor en entornos monolingües [8]. Entre estos parámetros, las características prosódicas, que están relacionadas con la entonación, la energía y la velocidad del habla, parecen una buena alternativa [9], ya que pueden ser estimados fácilmente mediante algoritmos automáticos de procesado de señal y pueden ser calculados incluso para señales muy cortas. Al igual que las parametrizaciones espectrales, estas características prosódicas contienen información tanto del locutor como del idioma utilizado. En el caso de sistemas multilingües, utilizar estos parámetros prosódicos será conveniente si su variabilidad entre locutores es

mayor que su variabilidad entre idiomas. En este caso es razonable decir que la prosodia es menos dependiente de idioma que dependiente de locutor. Con el objetivo de ver si las características prosódicas pueden ser aplicadas con éxito para reducir la tasa de error en el caso bilingüe castellano-euskera, se han realizado medidas de separabilidad de locutor e idioma tanto para parámetros MFCC como para estas características prosódicas. Estas medidas se detallan en la sección 4.

Por tanto, la solución propuesta utiliza dos tipos de parámetros: Espectrales y prosódicos. Como representativa de la información espectral se ha seleccionado la parametrización MFCC de 18 componentes, junto con sus primeras y segundas diferencias. Se ha calculado un vector MFCC cada 10 ms y se ha aplicado normalización de media y varianza (MVN) sobre cada grabación con el objetivo de reducir efectos de canal.

Las características prosodicas utilizadas han sido los valores de entonación y energía absoluta extraídos cada 10 ms, junto con sus primeras y segundas derivadas. También se ha aplicado MVN con el objetivo de reducir la gran variabilidad entre sesiones que presentan este tipo de parámetros. Esta aproximación hace posible concatenar los vectores MFCC anteriormente calculados y los valores prosódicos, combinando fácilmente ambos tipos de parámetros. Así pues, la parametrización propuesta consiste en vectores MFCC con valores de entonación y energía añadidos. Puesto que no existe información de entonación en las tramas sordas, esta parametrización se realiza utilizando sólo las tramas sonoras.

3 Descripción de la Base de Datos

Para los experimentos se utilizó una nueva base de datos bilingüe castellano-euskera [10]. Esta base de datos contiene grabaciones de 22 locutores bilingües (11 hombres y 11 mujeres) en un entorno semi-silencioso. Las grabaciones se realizaron con un micrófono Plantronics DSP-400, utilizando una frecuencia de muestreo de 44,1 kHz y 16 bits por muestra. Cada locutor realizó cuatro sesiones de grabación espaciadas en el tiempo, con el objetivo de capturar la variación de la voz a lo largo del tiempo. Esta base de datos bilingüe fue adquirida junto con una base de datos biométrica multimodal [11] y el calendario de captura diseñado para esta base de datos biométrica fue utilizado para la nueva base de datos bilingüe. Hay una diferencia de dos semanas entre la primera y segunda sesión, cuatro entre la segunda y la tercera y seis semanas entre la tercera y la cuarta.

Cada locutor grabó 7 secuencias numéricas formadas por 8 dígitos, pudiendo leerlas según su preferencia. Todas las secuencias numéricas son comunes para el castellano y el euskera.

4 Estudio de la Dependencia de los Parámetros con el Idioma

Una parametrización adecuada para el reconocimiento de locutores debe tener una gran variabilidad entre locutores (para permitir discriminarlos) y una reducida variabilidad intra locutor (de forma que la distribución de los parámetros no cambie mucho

entre las condiciones de entrenamiento y prueba). Para verificar que la parametrización propuesta es adecuada, se ha estimado su variabilidad utilizando la divergencia de Kullback-Leibler [12] como medida de distancia entre distribuciones.

Para la variabilidad inter locutor, se ha calculado la divergencia K-L entre todas las posibles parejas de locutores. El valor medio de todas estas medidas es representativo de la divergencia media entre dos locutores cualquiera, y por ello se ha utilizado como estimación de la variabilidad global inter locutor. Este cálculo se ha llevado a cabo de forma separada para el castellano y el euskera.

De forma similar, se ha estimado la variabilidad entre sesiones para cada locutor como la divergencia K-L media entre todas las posibles parejas de sesiones disponibles para ese locutor. La variabilidad inter sesión global se ha estimado como la variabilidad media para todos los locutores. Este cálculo también se ha realizado de forma separada para el castellano y el euskera.

Por último para cada locutor se ha calculado la divergencia K-L entre las parametrizaciones de las grabaciones en castellano y euskera. El valor medio entre todos los locutores se ha utilizado otra vez como una medida global de la variabilidad inter idioma.

La relación entre la variabilidad inter locutor e inter idioma puede usarse como medida de la robustez de una parametrización frente al idioma. Similarmente, la relación entre la variabilidad inter locutor e inter sesión puede usarse como medida de la robustez de la parametrización frente a la sesión. Lo ideal es que estas dos medidas sean tan grandes como sea posible.

Los resultados de estas medidas se resumen en la Tabla 1 para parámetros MFCC tradicionales y la parametrización propuesta. Tal y como se esperaba, al añadir los nuevos parámetros prosódicos a los vectores MFCC se incrementan todas las variabilidades, puesto que incrementar el número de dimensiones de los vectores sólo puede aumentar la distancia entre dos distribuciones. Pero mientras que la variabilidad inter locutor se incrementa en un 30%, la variabilidad inter idioma sólo se incrementa un 12%. Tal y como se refleja en la relación entre la variabilidad inter locutor frente a la variabilidad inter idioma, esto supone un incremento de la robustez frente al idioma de un 15%.

Tabla 1: Divergencia K-L frente al locutor, sesión e idioma para MFCC tradicional y con parámetros prosódicos (MFCC+P), para las grabaciones en castellano (C) y euskera (E).

		MFCC	***MFCC+P***	***Ganancia***
locutor	**C**	6.34	8.25	30%
	E	6.82	8.77	29%
sesión	**C**	3.62	4.81	33%
	E	3.52	4.64	32%
idioma	**-**	4.09	4.61	12%
locutor/idioma	**C**	1.55	1.79	15%
	E	1.67	1.90	14%
locutor/sesión	**C**	1.75	1.72	-2%
	E	1.94	1.89	-3%

La otra cara de la moneda es que las medidas de entonación y energía tienen una gran variabilidad inter sesión, por lo que al menos parte de la ganancia obtenida en robustez frente al idioma se perderá debido a la sensibilidad frente a la sesión. La relación entre la variabilidad inter locutor frente a variabilidad inter sesión se reduce alrededor de un 2-3%. Esto significa que cuando las pruebas se realicen en el mismo idioma que el entrenamiento (es decir, no hay variabilidad inter idioma), es previsible que los resultados sean un poco peores con la parametrización propuesta que utilizando únicamente parámetros MFCC tradicionales.

5 Condición de los Experimentos

Se han utilizado modelos GMM entrenados mediante el algoritmo EM [13], tanto para la parametrización MFCC tradicional como para los vectores MFCC con características prosódicas a corto plazo añadidas. Los parámetros MFCC tradicionales se han entrenado usando tanto tramas sordas como sonoras, mientras que los modelos de MFCC con características prosódicas utilizan sólo las tramas sonoras. También se han evaluado modelos MFCC con sólo tramas sonoras para facilitar la comparación.

Las grabaciones se han diezmado a 8kHz. Se ha utilizado un detector de actividad vocal (VAD) basado en la desviación espectral a largo plazo [14] con el objeto de eliminar las regiones de silencio de las grabaciones antes de aplicar la parametrización. Las cuatro sesiones disponibles en la base de datos se han utilizado en los experimentos en un esquema leave-one-out. El modelo de cada locutor se ha entrenado utilizando dos sesiones completas (aproximadamente 45 segundos de voz), mientras que una tercera sesión se ha utilizado para pruebas de desarrollo, con el objetivo de estimar los meta-parámetros del modelo (en este caso, el número de componentes gaussianas). La cuarta sesión se ha reservado para las pruebas finales. Este procedimiento se ha repetido cuatro veces, cambiando en cada caso la función de cada sesión. Por último, se ha calculado la precisión global del sistema como la precisión media de todas las iteraciones. En los casos en los que se ha realizado un entrenamiento bilingüe (utilizando ambos idiomas en el entrenamiento), se ha tomado una sesión de entrenamiento en castellano y la otra en euskera, de forma que se siguen utilizando dos sesiones de entrenamiento. Esto permite una comparación directa entre los sistemas, ya que todos ellos han sido entrenados utilizando aproximadamente la misma cantidad de voz.

6 Resultados Experimentales

Como referencia la Tabla 2 muestra las tasas de acierto de los modelos GMM con parametrización MFCC tradicional y utilizando el mismo idioma tanto para entrenamiento como para prueba (C=castellano, E=euskera). La precisión del sistema disminuye para un número de componentes gaussianas mayor que 64, a causa del sobreentrenamiento de los modelos debido a la reducida cantidad de material de entrenamiento disponible. La Tabla 3 muestra las tasas de acierto de los modelos de 64 componentes cuando el entrenamiento y las pruebas se realizan con idiomas diferen-

tes. Como puede apreciarse, bajo estas condiciones la precisión se reduce significativamente. También se aprecia que si se realiza un entrenamiento bilingüe, los resultados vuelven a estar cerca de los obtenidos en el caso de un único idioma.

Tabla 2: Tasa de identificación correcta para diferente número de componentes gaussianas utilizando parámetros espectrales con entrenamiento y pruebas realizadas con el mismo idioma

# mix	C-train; C-test	E-train, E-test
2	81.12	79.25
4	89.29	87.93
8	94.05	92.35
16	96.60	95.24
32	97.62	95.41
64	98.34	97.29

Tabla 3: Tasas de identificación correcta para los modelos de 64 componentes gaussianas con entrenamiento y prueba en diferente idioma. CE significa entrenamiento bilingüe.

C-E	E-C	CE-C	CE-E
63.55	67.34	96.77	95.58

Sin embargo, esta solución de entrenamiento bilingüe no es generalizable a idiomas no vistos durante el entrenamiento, y sería preferible utilizar una parametrización más robusta frente al cambio de idioma. La Tabla 4 muestra los resultados obtenidos con la parametrización propuesta, usando sólo tramas sonoras. Con el objetivo de facilitar la comparación, también se muestran los resultados de un sistema MFCC tradicional usando sólo las tramas sonoras. Si se comparan los resultados de la Tabla 2 y la Tabla 3 para modelos de 64 componentes con los valores de la Tabla 4 para parámetros MFCC tradicionales, puede comprobarse que el hecho de descartar las tramas sordas tiene poca influencia en los resultados finales para esta parametrización.

Cuando se añaden los parámetros prosódicos a corto plazo, la precisión del sistema en condiciones de idioma único se reduce ligeramente, tal y como predicen las medidas de variabilidad de la sección 4. Sin embargo las tasas de acierto aumentan significativamente en el caso de usar un idioma diferente para el entrenamiento y las pruebas, debido a que la mejora en la robustez frente al idioma es mayor que la pérdida de robustez frente a la sesión.

Tabla 4: Tasa de identificación correcta para modelos de 64 componentes usando sólo tramas sonoras, para parametrización MFCC y MFCC con parámetros prosódicos a corto plazo. También se detalla el incremento de precisión obtenido al añadir los parámetros prosódicos.

	C-C	E-E	C-E	E-C	CE-C	CE-E
MFCC	97.6	96.8	62.6	67.0	96.6	95.6
MFCC+P	97.1	96.3	71.0	73.0	96.1	94.4
Ganancia (%)	-0.5	-0.5	13.4	8.9	-0.5	-1.3

Cuando se realiza un entrenamiento bilingüe de los modelos la precisión también se reduce un poco en el caso de los parámetros prosódicos añadidos. Estos modelos ya han adquirido una robustez frente al idioma gracias a este entrenamiento bilingüe. Sin embargo, esta robustez sólo es válida para los dos idiomas considerados en el entrenamiento (castellano y euskera), y la precisión del sistema volvería a caer si se utilizara un tercer idioma para las pruebas.

7 Conclusiones

En este trabajo se han estudiado las ventajas de añadir información de energía y entonación a corto plazo a parámetros MFCC para obtener una parametrización más robusta frente al idioma en sistemas de reconocimiento de locutores. En una primera etapa se han estimado las variabilidades frente a locutor, sesión e idioma de estos parámetros. Estas medidas han permitido prever una mejora en la precisión del reconocimiento cuando la prueba se realiza un idioma no visto durante el entrenamiento. Los resultados experimentales confirman esta predicción, mostrando una mejora significativa de la tasa de acierto bajo condiciones de desadaptación de idiomas.

Estos resultados experimentales también muestran una pequeña pérdida de precisión cuando el entrenamiento y las pruebas se realizan utilizando un único idioma, debido a la gran variabilidad inter sesión de los parámetros prosódicos. En cualquier caso, esta pérdida puede ser perfectamente asumible cuando el sistema es utilizado en un entorno multilingüe y no puede realizarse un entrenamiento con varios idiomas, o cuando no es posible conocer de antemano el idioma que el locutor va a utilizar al usar el sistema.

Aunque las características prosódicas a corto plazo mejoran la robustez frente al idioma de los sistemas de reconocimiento de locutor, los resultados todavía están lejos de ser totalmente independientes del idioma. Se necesitan nuevos parámetros o nuevas técnicas de normalización de idioma para poder construir un sistema que mantenga una precisión similar independientemente de los idiomas de entrenamiento y prueba.

Agradecimientos: Este trabajo ha sido financiado parcialmente por el Gobierno Vasco bajo la subvención IE06-185 (proyecto ANHITZ, http://www.anhitz.com) y por la Universidad del País Vasco y EJIE S.A. bajo la subvención EJIE07/02 (proyecto MULTILOK).

Referencias

1. Faundez, M., Satue-Villar, A.: Speaker recognition experiments on a bilingual database. In: IV Jornadas en Tecnologias del Habla (4JTH), pp. 261--264 (2006).
2. Durou, D.: Multilingual text-independent speaker identification. In: Multi-lingual Interoperability in Speech Technology (MIST), pp. 115--118 (1999).

3. Akbacak, M., Hansen, J. H. L.: Language normalization for bilingual speaker recognition systems. In: International Conference on Acoustics, Speech, and Signal Processing (ICCASP), pp. 257--260 (2007).
4. Ma, B., Meng, H.: English-Chinese bilingual text-independent speaker verification. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP'04), pp. 293--296 (2004).
5. Paalanen, P., Kamarainen, J. K., Ilonen, J., Kälviäinen, H.: Feature representation and discrimination based on Gaussian mixture model probability densities – Practices and algorithms. Pattern Recognition 39, 1346--1358 (2006).
6. Young,S.: Large vocabulary speech recognition: A review. In: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 3--28 (1995).
7. Reynolds, D. A., Rose, R. C.: Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models. IEEE transactions on Speech and Audio Processing 3, 72--83 (1995).
8. Reynolds, D. A., Campbell, J. P., Dunn, R. B., Gleason, T., Jones, D., Quatieri, T. F., Carl, Q., Sturim, D., Torres-Carrasquillo, P.: Beyond Cepstra: Exploiting High-Level Information in Speaker Recognition. In: Workshop on Multimodal User Authentication, pp. 223--229 (2003).
9. Dehak, N., Dumouchel, P., Kenny, P.: Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification. IEEE Transactions On Audio, Speech, And Language Processing 15, 2095--2103 (2007).
10. Luengo, I., Navas, E., Sainz, I., Saratxaga, I., Sanchez, J., Odriozola, I., Igarza J.J., Hernaez, I.: Building a Basque/Spanish bilingual database for speaker verification. In Workshop Collaboration: interoperability between people in the creation of language resources for less-resourced languages, pp. 23--26, (2008).
11. Galbally, J., Fierrez, J., Ortega-Garcia, J., Freire, M. R., Alonso-Fernandez, F., Siguenza, J.A., Garrido-Salas, J., Anguiano-Rey, E., Gonzalez-de-Rivera, G., Ribalda, R., Faundez-Zanuy, M., Ortega, J.A., Cardeñoso-Payo, V., Viloria, A., Vivaracho, C. E., Moro, Q. I., Igarza, J.J. Sanchez, J., Hernaez I., Orrite-Uruñuela, C.: BiosecurID: a Multimodal Biometric Database. In: MADRINET Workshop, pp. 68--76 (2007).
12. Kullback, S., Leibler, R. A.: On information and sufficiency. Annal of Mathematical Statistics 22, 79--86 (1951).
13. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification. Wiley, John and Sons (2001).
14. Ramirez, J., Segura, J. C., Benitez, C., de la Torre, A., Rubio, A.: Efficient Voice Activity Detection Algorithms Using Long Term Speech Information. Speech Communication 42, 271--287 (2004).