GRABACIÓN DE UNA BASE DE DATOS BILINGÜE EUSKERA/CASTELLANO PARA VERIFICACIÓN DE LOCUTOR

Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sánchez, Igor Odriozola, Juan José Igarza, Inmaculada Hernáez

AhoLab Signal Processing Group.
Departamento de Electrónica y Telecomunicaciones.
Universidad del País Vasco (UPV/EHU).
Alda. Urquijo s/n, 48013 Bilbao.

{ikerl, eva, inaki, ibon, ion, igor, jigarza, inma}@aholab.ehu.es

RESUMEN

Los grupos de investigación de procesado del habla que trabajan con lenguas minoritarias han de afrontar una serie de dificultades a la hora de grabar nuevas bases de datos orales para esas lenguas, tales como la falta de recursos previos, la escasez de personas que hablan el idioma de forma fluida y la dificultad de encontrar financiación para el proyecto. Algunas veces es posible aprovechar campañas de grabación para otros proyectos y extenderlos de tal forma que se incluyan grabaciones en esa lengua minoritaria para cada donante que lo domine. De esta forma se puede grabar una nueva base de datos con poco esfuerzo, ya que la campaña de grabación a sido preparada y financiada de antemano. Usando esta misma técnica se ha creado una nueva base de datos bilingüe euskera/castellano, gracias a la cual se está llevando a cabo un estudio sobre sistemas de verificación bilingües en estos idiomas. En el presente artículo se describe la base de datos resultante así como las dificultades encontradas durante su grabación.

1. INTRODUCCIÓN

Hoy en día muchos sistemas requieren de algún mecanismo de autenticación de usuario para evitar fraudes o accesos no autorizados. La mayoría de estos sistemas utilizan una autenticación basada en claves, pero estas claves se pueden olvidar o robar. Actualmente los métodos de autenticación biométrica son la mejor alternativa, ya que proporcionan una verificación extremadamente segura y precisa [1]. Además, las características biométricas no se pueden perder ni olvidar, y son muy difíciles de imitar. Este tipo de autenticación se utiliza en la actualidad en sistemas como ordenadores portátiles con control de acceso mediante huella digital o acceso a edificios mediante geometría de la mano. La voz es una característica biométrica no intrusiva, que tiene un alto grado de aceptabilidad y que es apropiada para sistemas de verificación a larga distancia sobre redes de datos y voz. Para el desarrollo de estos sistemas de autenticación basados en voz, es necesario contar con bases de datos orales con grabaciones de diferentes locutores.

Como método de autenticación biométrica, la verificación de locutor ha de decidir si una persona es o no quien dice ser, utilizando para ello una o más señales de voz de esta persona [2]. En un sistema de verificación de locutor general se pueden distinguir dos módulos: El módulo de entrenamiento (que genera un modelo para cada usuario del sistema) y el módulo de pruebas (que decide si una señal de voz ha sido producida por un locutor específico) [3][4]. Generalmente se supone que el idioma de las señales de entrenamiento y prueba es el mismo. Pero en entornos multilingües es deseable que los usuarios del sistema de verificación puedan utilizar cualquiera de los idiomas que conozcan para acceder al sistema, sin notar diferencias apreciables en el funcionamiento del mismo. Por ello, en los últimos años, varios grupos de investigación han centrado su atención en sistemas de reconocimiento de locutor en entornos multilingües, donde los modelos pueden ser entrenados utilizando un idioma y las pruebas ser realizadas en otro [5][6].

Este entorno multilingüe añade algunas dificultades al sistema de verificación. Por un lado, la diferencia entre los idiomas de entrenamiento y prueba provoca una reducción de la precisión del sistema [7]. Por otro, las diferencias entre los idiomas del modelo de locutor y el modelo de locutor universal en un sistema de verificación de locutor GMM provoca también un aumento de los errores [8].

El País Vasco es un ejemplo de este tipo de entornos multilingües, en el que conviven dos idiomas oficiales, el euskera y el castellano. El euskera es un idioma minoritario, y por tanto, existe una falta de recursos lingüísticos en este idioma [9]. Concretamente, no existe ninguna base de datos oral pública disponible para el desarrollo de sistemas de verificación en este idioma.

Este artículo presenta el trabajo y las dificultades de grabar una nueva base de datos oral bilingüe en el País Vasco para el desarrollo de sistemas de verificación de locutores bilingües euskera/castellano. La sección 2 analiza los problemas asociados a la grabación de nuevas bases de datos para lenguas minoritarias. La sección 3 describe la base de datos grabada y sus contenidos, mientras que en la sección 4 se describen las dificultades que surgieron durante esta adquisición. Finalmente se comentan algunas conclusiones en la sección 5.

2. ADQUISICIÓN DE NUEVAS BASES DE DATOS ORALES PARA LENGUAS MINORITARIAS

Aunque se consideran de alto interés social, las lenguas minoritarias no son económicamente interesantes. Puesto que hay pocos hablantes, no compensa invertir una gran cantidad de dinero para la investigación y desarrollo de nuevos recursos orales como bases de datos. Esto significa que generalmente es difícil encontrar fuentes de financiación para estos proyectos y que en los casos en los que se consigue, la financiación lograda es escasa. Además generalmente no hay muchos grupos de investigación trabajando en estos idiomas, por lo que tampoco es sencillo buscar colaboraciones ente grupos para repartir la carga de trabajo y los costes.

En el caso de bases de datos de verificación de locutor, el proceso se complica aun más debido a los requerimientos específicos de estas bases de datos. Por un lado, las grabaciones deben realizarse a lo largo de un período de tiempo suficientemente largo como para recoger la variabilidad natural de la voz [10]. Esto significa que cada locutor debe ser grabado más de una vez, en diferentes sesiones, lo que hace que el proceso de grabación sea más largo y caro. Por otro lado, es interesante que la distribución de sexo y edad de los locutores se aproxime a la verdadera distribución de los usuarios potenciales del sistema. Esta restricción, junto con el hecho de que en una lengua minoritaria hay pocos hablantes, hace que el reclutamiento de los donantes sea más complicado.

Para hacer factible la grabación de bases de datos en una lengua minoritaria puede ser interesante aprovechar las campañas de grabación organizadas para otros proyectos e incluir algunas grabaciones adicionales en esta lengua, aunque éste no sea el objetivo principal de la campaña. De esta forma se pueden obtener nuevas bases de datos con poco esfuerzo, dado que la campaña de grabación ya ha sido preparada de antemano.

En el laboratorio de procesado de señal Aholab de la Universidad del País Vasco se llevan a cabo diferentes investigaciones en tecnologías de la voz para el euskera, principalmente en los campos de conversión de texto a habla (CTH), reconocimiento automático del habla (RAH) y verificación de locutor. Para el desarrollo de estas investigaciones se necesitan bases de datos orales en euskera. Algunas veces, cuando se está llevando a cabo una campaña de grabación para otros proyectos (principalmente en castellano), se pide a los donantes que realicen algunas grabaciones extra en euskera, para completar una base de datos paralela en esta lengua.

3. DESCRIPCIÓN DE LA BASE DE DATOS

La nueva base de datos euskera/castellano se grabó junto con una base de datos biométrica multimodal adquirida en cinco universidades de España, incluyendo la Universidad del País Vasco [11]. En esta base de datos se adquirieron diferentes características biométricas, tales como huella dactilar, firma, escritura manuscrita, iris o habla (en castellano). Aprovechando la oportunidad también se grabó en euskera a aquellos donantes reclutados en la Universidad del País Vasco que eran hablantes fluidos en este idioma. De esta forma se consiguió construir una pequeña base de datos bilingüe para verificación de locutor con poco esfuerzo.

3.1. Diseño de la base de datos

El protocolo de adquisición incluyó cuatro sesiones distribuidas en el tiempo para capturar la variabilidad intra-locutor. Hay una diferencia de dos semanas entre la primera y la segunda sesión, cuatro ente la segunda y la tercera y seis semanas ente la tercera y la cuarta sesión.

En cada sesión se grabaron una serie de frases aisladas y unas secuencias numéricas. El conjunto de frases es el mismo para todos los locutores, aunque cambian de una sesión a otra. La primera sesión consta de cuatro frases para cada idioma, mientas que en las demás sesiones sólo se grabaron dos en cada idioma. Por tanto, el corpus contiene 10 frases en castellano y otras 10 en euskera. Se trata de frases fonéticamente ricas y equilibradas, seleccionadas mediante la herramienta CorpusCRT a partir de dos grandes corpus textuales, uno para cada idioma. Esta herramienta desarrollada por el grupo TALP de la UPC[1] proporciona un conjunto de frases reducido manteniendo, en la medida de lo posible, la frecuencia de aparición de los fonemas del corpus original.

Las secuencias numéricas están formadas por 8 dígitos que el locutor podía leer como prefiriera. Cada locutor tiene una secuencia numérica única que se repite cuatro veces en cada sesión. Además, cada locutor también grababa la secuencia numérica asignada a otros tres locutores, diferentes cada vez, con el objetivo de utilizarlas como pruebas de impostor. En total se graban 7 secuencias numéricas por cada sesión, locutor e idioma.

[1] Universidad Politécnica de Catalunya. www.talp.upc.es

3.2. Datos adicionales

Las grabaciones de la base de datos se procesaron para extraer la información de actividad vocal y las curvas de entonación.

La estimación de la actividad vocal es necesaria para descartar tramas en las que no hay información del habla, de forma que el nivel de ruido existente durante los silencios no corrompa los parámetros calculados para el sistema de verificación. La detección de actividad vocal utilizada se basa en la desviación espectral a largo plazo (LTSD) tal y como se propone en [12].

Para el cálculo de las curvas de entonación se utilizó una herramienta desarrollada por el mismo grupo Aholab [13], que utiliza programación dinámica y coeficientes cepstrales para estimar la curva de pitch.

4. DIFICULTADES ENCONTRADAS

4.1. Escasez de hablantes bilingües

Recoger una base de datos en euskera no es fácil, ya que no hay mucha gente que lo hable de forma fluida, ni siquiera en el propio País Vasco. La Tabla 1 presenta el número de hablantes de euskera en la Comunidad Autónoma del País Vasco en 2001, según edades[2]. En esta tabla se consideran hablantes bilingües tanto activos como pasivos, es decir, incluye a aquellos locutores cuyo primer idioma no es euskera, y por tanto, cuyo dominio del idioma no es siempre bueno (algunos no son hablantes fluidos).

Edad	Total	Porcentaje
16-24	170 453	23.1%
25-34	171 608	23.3%
35-49	175 522	23.8%
50-64	104 055	14.1%
>=65	115 442	15.7%
TOTAL	737 080	100.0%

Tabla 1: *Distribución de hablantes bilingües activos y pasivos por edades en la Comunidad Autónoma del País Vasco en 2001.*

El conocimiento y uso del euskera varía según el rango de edad. La Tabla 2 muestra el porcentaje de personas monolingües y bilingües para cada rango de edad. Como puede verse la proporción de hablantes de euskera es mayor entre las personas jóvenes.

Además de la falta de hablantes, el hecho de que la base de datos bilingüe se grabara como una extensión de otra base de datos biométrica perteneciente a otro proyecto también acarrea problemas. Las especificaciones principales de la base de datos biométrica, como por ejemplo el número de voluntarios, su distribución de edad y los plazos de entrega tenían que ser respetados. Puesto que las grabaciones en euskera no formaban parte de la base de datos principal, el conjunto de especificaciones para la base de datos biométrica no tenía en cuenta los requerimientos especiales necesarios para una base de datos bilingüe. Por ejemplo, no era posible rechazar a un voluntario por el hecho de no hablar euskera, ya que esto hubiera supuesto dificultar el reclutado y extender los plazos de entrega de la base de datos principal. Esta es la razón por la que, aunque 55 voluntarios fueron grabados en castellano, sólo 30 de ellos se grabaron en euskera, al ser los únicos realmente bilingües.

Edad	Monolingües	Bilingües
16-24	31.4%	68.6%
25-34	50.5%	49.5%
35-49	63.3%	36.7%
50-64	72.5%	27.5%
>=65	67.4%	32.6%

Tabla 2: *Porcentaje de hablantes monolingües y bilingües por edades en la Comunidad Autónoma del País Vasco en 2001.*

4.2. Desviación de la distribución de edad

En una base de datos de verificación de locutor la población debe estar correctamente representada. Es importante que la base de datos incluya ejemplos representativos de todos los potenciales usuarios del sistema. Esta es la razón por la que este tipo de bases de datos suelen estar equilibrados en sexo y rangos de edad. Para lograr este equilibrio se propone una distribución objetivo para los locutores, según la distribución real de los usuarios potenciales, y se seleccionan los donantes de acuerdo a este objetivo. La Tabla 3 muestra la distribución objetivo de rangos de edad y la distribución de los locutores grabados en castellano y euskera.

Edad	Objetivo	Castellano	Euskera
18 – 25	30%	32.7%	33.3%
25 – 35	20%	40.0%	53.3%
35 – 45	20%	12.7%	10.0%
45 – 55	20%	7.3%	3.3%
>= 55	10%	7.3%	0.0%

Tabla 3: *Distribución de edad en los locutores de la base de datos en castellano y euskera.*

El reclutamiento de los locutores se realizó principalmente entre los estudiantes y el personal de la Escuela de Ingeniería de la Universidad. La media de edad en este colectivo es relativamente baja, tal y como

[2] Fuente: EAS (Sistema Indicador de Lengua del País Vasco). http://www1.euskadi.net/euskara_adierazleak/zerrenda.apl?hizk=i&gaia=25&sel=64

se refleja en la desviación del objetivo en los rangos de 25 a 35 años y de más de 45, tanto para castellano como para euskera. Además, es muy difícil reclutar donantes bilingües mayores de 35 años, ya que la mayoría de las personas en este rango de edad no hablan euskera, tal y como se refleja en la Tabla 1. Esta es la razón por la que hay tan pocos locutores grabados en euskera en rangos de edad altos.

La desviación de la distribución de edades es mayor para las grabaciones en euskera que para el castellano. Otra vez, la razón principal es que durante el reclutamiento era prioritario mantener la distribución de edad para la base de datos principal, en la que se incluían las grabaciones en castellano. Pero al descartar a los no bilingües, la nueva distribución de edades para el euskera no coincidía con el objetivo.

El equilibrio de sexos fue más sencillo de conseguir. En la Tabla 4 se muestra la distribución objetivo junto con las obtenidas para el castellano y euskera. Como se aprecia, las distribuciones logradas no difieren significativamente entre ambos idiomas.

Sexo	Objetivo	Castellano	Euskera
Hombre	50%	47.3%	43.3%
Mujer	50%	52.7%	56.7%

Tabla 4: *Distribución del sexo de los locutores en la base de datos en castellano y euskera.*

5. CONCLUSIONES

Teniendo en cuenta que el euskera es una lengua minoritaria, el desarrollo de nuevos recursos orales para este idioma es difícil y la financiación escasa. Bajo estas circunstancias, la adquisición de una base de datos en una lengua mayoritaria representa una oportunidad que puede aprovecharse para construir otra base de datos en la lengua minoritaria. Haciendo uso de esta estrategia se ha creado una nueva base de datos para verificación de locutores bilingües en euskera y castellano. Sus características no son ideales, ya que el proceso de adquisición no fue diseñado explícitamente para ella, pero sigue siendo un recurso útil, con el que ya se han realizado algunos experimentos de verificación de locutor [14].

6. AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado 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).

Los autores también quieren agradecer su participación a todos los voluntarios que tomaron parte en la adquisición de la base de datos biométrica.

7. BIBLIOGRAFÍA

[1] A.K. Jain, A. Ross, S. Pankanti, *Biometrics: a tool for information security*, IEEE Transactions on Information Forensics and Security, vol. 1, pp. 125—143, 2006.

[2] J.P. Campbell, S*peaker Recognition: A tutorial,* In Proceedings of the IEEE, vol. 85, pp. 1437—1462, 1997.

[3] J.M. Naik, *Speaker verification: a tutorial*. IEEE Communications Magazine, vol. 28, pp. 42—48, 1990.

[4] F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, D.A. Reynolds, *A Tutorial on Text-Independent Speaker Verification* EURASIP Journal on Applied Signal Processing, vol. 4, pp. 430—451, 2004.

[5] T. Nordstrom, H. Melin, J. Lindberg, *Comparative Study of Speaker Verification Systems using the Polycost Database,* In Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP), pp. 1359—1362.

[6] M. Faundez-Zanuy, A. Satue-Villar, *Speaker Recognition Experiments on a Bilingual Database*, In Proceedings of the 14th European Conference on Signal Processing (EUSIPCO), 2006.

[7] B. Ma, H. Meng, *English-Chinese bilingual text-independent speaker verification* In Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), vol. 5, pp. 293—296, 2004.

[8] R. Auckenthaler, M.J. Carey, J.S.D. Mason, *Language dependency in text-independent speaker verification* In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) vol. 1, pp. 441—444, 2001.

[9] A. Díaz de Ilarraza, K. Sarasola, A. Gurrutxaga, I. Hernaez, N. Lopez de Gereñu, *HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities,* In Proceedings of the Workshop on NLP of Minority Languages and Small Languages, 2003.

[10] P. Kenny, P. Dumouchel, *Disentangling speaker and channel effects in speaker verification*, In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 37—40, 2004.

[11] J. Galbally, J. Fierrez, J. Ortega-Garcia et. al., *BiosecurID: a Multimodal Biometric Database*, In Proceedings of the User-Centric Technologies and Applications Workshop, pp. 68—76, 2007.

[12] J. Ramirez, J. Segura, C. Benitez, A. de la Torre, A. Rubio, *Efficient Voice Activity Detection Algorithms Using Long Term Speech Information*, Speech Communication, vol. 42, pp. 271—287, 2004.

[13] I. Luengo, I. Saratxaga, E. Navas, I. Hernáez, J. Sanchez, I. Sainz, *Evaluation Of Pitch Detection Algorithms Under Real Conditions*. In Proceeding of 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1057—1060, 2007.

[14] Luengo, I., Navas, E., Sainz, I., Saratxaga, I., Sanchez, J., Odriozola, I., Hernaez, I. Text independent speaker identification in multilingual environments. Proc. of the Sixth International Language Resources and Evaluation (LREC'08), paper 461, 2008.