Eficiencia Computacional y Alto Rendimiento en Reconocimiento Automático de Locutor: el Sistema ATVS-UAM en NIST SRE 2010

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Javier Franco-Pedroso, Daniel Ramos, Doroteo T. Toledano y Joaquin Gonzalez-Rodriguez

ATVS – Biometric Recognition Group.
Escuela Politecnica Superior, Universidad Autonoma de Madrid.
C. Francisco Tomás y Valiente 11, 28049 Madrid, Spain.
javier.gonzalez@uam.es

1 Resumen

En esta contribución se presenta el sistema presentado por el grupo de reconocimiento biométrico ATVS a la evaluación NIST Speaker Recognition Evaluation 2010 (NIST SRE 2010) [1]. En esta evaluación, se ha intentado evitar el envío de un sistema complejo y pesado computacionalmente mediante la combinación múltiples sistemas para intentar maximizar la figura de mérito del Equal Error Rate (EER) o Detection Cost Function (DCF). En su lugar, ATVS ha enviado un único sistema simple y eficiente computacionalmente.

El sistema presentado está basado en el uso de información a nivel espectral, mediante la compensación de variabilidad intersesión en el espacio de variabilidad total, también conocido como espacio *i-vector* [2]. Se compone de los siguientes módulos:

- Fase de desarrollo (*development* o *dev*):
  - Detección de actividad de voz (VAD) basada tanto en energía como en modelado dinámico del ruido mediante las herramienteas del paquete Sound eXchange (SOX) [3].
  - Extracción de características. 38 coeficientes MFCC con compensación de variabilidad intersesión basada en *feature mapping* [4].
  - Entrenamiento de GMM para cada locución del conjunto de entrenamiento, adaptados desde un modelo universal UBM [5].
  - Reducción de dimensionalidad mediante PCA, obteniendo la matriz de transformación T del subespacio de variabilidad total [2].
  - Entrenamiento de los subespacios de locutor y canal en el espacio de variabilidad total, mediante el uso de *Linear Discriminant Analysis* (LDA) y *Within-Class Covariance Normalization* (WCCN)
- Fase de comparación (*testing*):

- Detección de actividad, extracción de características y entrenamiento de modelos.
- Transformación al espacio de variabilidad total.
- Compensación de variabilidad intersesión en dicho espacio.
- Cálculo de la puntuación de similitud (*score*) mediante distancia coseno [2].
- Normalización de la puntuación mediante Z-Norm seguido de T-Norm [6].
- Calibración, mediante regresión logística dependiente de género y del canal de procedencia de cada fichero en la comparación [7].

La Figura 1 esquematiza el funcionamiento del sistema.

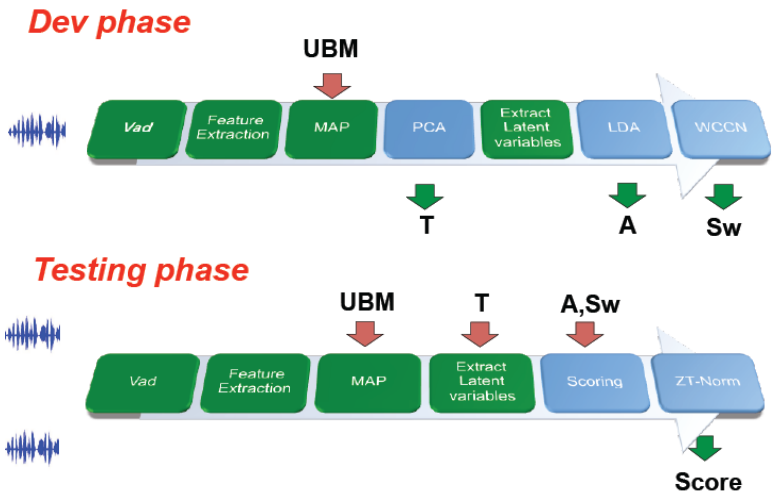

Fig. 1. Funcionamiento esquemático del sistema ATVS en NIST SRE 2010**.**

El esquema general del sistema varía para cada uno de los tipos de ficheros de habla. En NIST 2010 existen dos tipos básicos de ficheros: telefónicos (habla conversacional grabada utilizando diversos tipos de teléfonos) y microfónicos, siendo estos de tipo *phonecall-mic* (habla procedente de conversaciones telefónicas grabada con micrófonos) y de tipo *interview* (habla procedente de entrevistas grabada con micrófonos). Ello daba lugar a cuatro tipo de comparaciones, en concreto *tel-tel* (habla telefónica comparada con habla telefónica), *tel-mic* y *mic-tel* (habla telefónica comparada con habla microfónica) y *mic-mic* (habla microfónica comparada con habla microfónica). Para el entrenamiento de subespacios de variabilidad total y compensación de variabilidad intersesión, se utilizó habla telefónica para las comparaciones *tel-tel* y habla telefónica con habla microfónica para el resto de comparaciones. Se utilizaron para ello bases de datos procedentes de evaluaciones NIST pasadas y bases de datos públicas.

La Tabla 1 muestra un resumen de la carga computacional del sistema presentado. Se observan figuras de mérito muy relevantes, como un tiempo de comparación del orden de un microsegundo. En total, en una CPU de 2,2 GHz con 1024 MB de cache y 4 GB de memoria RAM, el reconocimiento completo de dos ficheros se efectúa 77 veces más rápido que tiempo real, lo cual constituye una muy apropiada carga computacional para aplicaciones reales.

Tabla1: Carga computacional desglosada del sistema ATVS en NIST SRE 2010.

	GMM-FA
Development	
Entrenamiento de UBM	• Tel: UBM, 4M vectores MFCC: 10h (hombres), 10h (mujeres) • Tel+Mic: UBM, 5M vectores MFCC: 12h (hombres), 12h (mujeres)
Entrenamiento del subespacio de variabilidad total	• Tel: T: 30m (hombres), 30m (mujeres), LDA: 8m (h), 8m (m), WCCN: 6m (h), 6m (m) • Tel+Mic: T: 45m (h), 1h 10m (m), LDA: 10m (h), 12m (m), WCCN: 8m (h), 10m (m)
Extracción de características (por cada fichero de 265s)	
MFCC	2s
VAD	1.57s
Entrenamiento de modelos (por cada fichero de 265s)	
Variables ocultas del subespacio de variabilidad total	0.05s
Total (entrenamiento)	3.62s
Tanto por uno de tiempo real en la CPU utilizada	0.013 TR
Comparación (por cada par de ficheros de 265s)	
Variables ocultas del subespacio de variabilidad total	0.05s
Puntuación	1e-6 s
Z-norm	0.02s (~300 de tamaño de la cohorte)
T-norm	0.02s (~300 de tamaño de la cohorte)
Total	3.66s
Tanto por uno de tiempo real en la CPU utilizada	0.013 TR

En cuanto al rendimiento en discriminación del sistema en la evaluación real, la Figura 2 presenta las curvas DET del mismo para los sistemas primario y secundario. Ambos sistemas difieren el uno del otro del reconocedor en el VAD utilizado (el secundario utiliza el VAD cuyas etiquetas fueron distribuidas por el Instituto Tecnológico de Brno (BUT). Se observa que esa diferencia es sustancial, debido a la gran cantidad de ficheros de habla microfónica existentes en la evaluación en los que

aparece habla de entrevistador, para los cuales la detección de la actividad del locutor de interés resulta crítica. Por otro lado, se puede ver en la Figura 2 que el sistema primario presenta un rendimiento más que aceptable para la complejidad de la tarea, y considerando sus prestaciones a nivel computacional podemos concluir que el sistema presentado es altamente adecuado para aplicaciones reales.

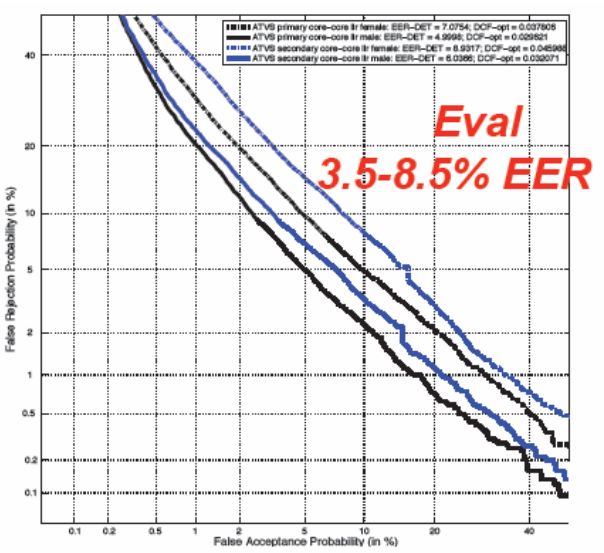

Fig. 2. Rendimiento del sistema ATVS en NIST SRE 2010. En azul curva correspondiente a mujeres, en negro a hombres. En sólido, curva correspondiente al sistema primario, en punteado al secundario**.**

Referencias

1. Página de la evaluación NIST SRE 2010: http://www.itl.nist.gov/iad/mig//tests/sre/.
2. Dehak, N., Dehak, R., Kenny, P., Brummer, N., Ouellet, P and Dumouchel, P.: Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In Proc Interspeech 2009, Brighton, UK (2009).
3. "Sound Exchange" software, Available at http://sox.sourceforge.net/ (accessed 28/06/2010).
4. Pelecanos, J., Sridharan, S.: Feature Warping for Robust Speaker Verification. In Proc Odyssey 2001, Crete, Greece (2001).
5. Reynolds D.A.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, vol. 10, pp. 19--41 (2000).
6. Auckenthaler R., Carey M., Lloyd-Tomas H.: Score Normalization for Text-Independent Speaker Verification Systems. Digital Signal Processing, vol. 10, pp. 42--54, (2000).
7. Brummer, N., Burget, L. et al.: Fusion of Heterogenous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. IEEE Trans. On Audio, Speech and Language Processing, vol.15, no.7 (2007).