Modelos de Regresión Logística a Nivel de Puntuaciones para Incorporar Calidad en la Verificación de Locutor

Sergio Pérez Gómez, Daniel Ramos Castro, Joaquín González Rodríguez y Julián Fierrez*

ATVS - Grupo de Reconocimiento Biométrico
Escuela Politécnica Superior, Universidad Autónoma de Madrid
Calle Francisco Tomás y Valiente 11, 28049, Madrid, España
{sergio.perez,daniel.ramos,joaquin.gonzalez,julian.fierrez}@uam.es
http://atvs.ii.uam.es/

Resumen El presente artículo es un estudio preliminar en el que se evalúan 3 métodos propuestos para compensar la pérdida de rendimiento en los sistemas de verificación de locutor debido a la variabilidad en la calidad de las muestras de audio estudiada. Estos algoritmos están basados en otros definidos en la literatura y serán evaluados mediante la mejora en términos de EER (*Equal Error Rate*) de un sistema GMM-UBM entrenado y testeado mediante la condición *short2-short3* de la evaluación NIST SRE 2008 (*Speaker Recognition Evaluation*) [1]. La mejora relativa de EER más significativa obtenida es del 8,82 %, hallada mediante el algoritmo de *Regresión Logística Lineal de 2 Dimensiones* (2D-LLR) para la condición *tel-mic* de NIST SRE 2008, utilizando UBML (*Universal Background Model Likelihood*) como información de calidad, aunque también se han explorado resultados con la *Relación Señal a Ruido* del mismo modo. Aunque los resultados son preliminares y optimistas, puesto que los algoritmos se han entrenado con los mismos datos que la evaluación, se observa una clara y relevante tendencia de mejora.

Keywords: verificación de locutor, calidad, utilidad, rendimiento, UBML, SNR, regresión logística.

1. Introducción

La idea de que la calidad de una muestra de voz puede afectar al rendimiento de un sistema de reconocimiento automático de locutor es bastante intuitiva [2]. De hecho la medida y compensación de la calidad de una señal de audio ha sido una tarea en el que se ha invertido un gran esfuerzo en el ámbito científico biométrico en los últimos años [3]. Inicialmente este esfuerzo viene por la

* Este trabajo ha sido financiado por el Ministerio de Ciencia e Innovación (TEC2009-14719-C02-01) y la cátedra UAM-Telefónica.

necesidad de controlar la calidad de la voz en las redes telefónicas pero en la actualidad se ha transformado en la definición de medidas de calidad y algoritmos de calibración que permitan predecir el rendimiento de un sistema biométrico.

Si bien es cierto que existen técnicas como *factor analysis* que reducen de forma significativa la variabilidad introducida por el canal [4] estas técnicas dependen en gran medida de la existencia de un corpus apropiado, deseablemente con las mismas condiciones de la voz a reconocer. Sin embargo, estos modelos son dependientes de los datos utilizados para su entrenamiento. La calidad de voz está basada en el conocimiento de la señal de voz mediante la cual se puede predecir tanto el rendimiento de un sistema de verificación de locutor como un posible desalineamiento de las puntuaciones del mismo debido a cambios en dicha calidad. En este trabajo se propone el uso de la información de calidad para ajustar el desalineamiento entre las puntuaciones *target* y *non target* de un sistema de verificación de locutores, mediante modelos de regresión logística.

Esta investigación se inicia en el estudio de distintas medidas de calidad definidas y su utilización para compensar variabilidad intersesión a nivel de locución. De entre todas ellas se ha procedido a elegir 2 que definen de manera más significativa que el resto el poder discriminativo del sistema [5]: la SNR (*Signal To Noise Ratio*) y la UBML (*Universal Background Model Likelihood*). Una vez estudiado el comportamiento del sistema definido en 5.2 frente a estos indicadores de degradación de rendimiento se han diseñado 3 algoritmos basados en un modelo de regresión logística [6]: *Regresión Logística Lineal de 2 Dimensiones* (2D-LLR) y *Regresión Logística Bilineal* (BLR tipo 1 y tipo 2), que evaluarán dicho rendimiento sobre la base de datos de NIST SRE 2008 [1] que presenta un desafío de variabilidad intersesión [7].

Este artículo está organizado de la siguiente manera: la sección 2 presenta la motivación de este trabajo así como la definición de las medidas de calidad que determinan el rendimiento del sistema a mejorar. En la sección 4 se describen los métodos propuestos de compensación aplicados sobre la base de datos y el sistema descrito en la sección 5, cuyos resultados y conclusiones son ampliamente analizados en 6 y 7.

2. Tratamiento e interpretación de medidas de calidad

La calidad de una muestra biométrica viene definida por tres criterios básicos según [8], un borrador estándar de calidad según NIST [9] de dicho tipo de muestras:

- La fidelidad, que se refiere a la exactitud y precisión con la que una muestra biométrica es capturada, procesada y almacenada en el sistema.
- El carácter, entendido como la actitud o predisposición del usuario a que se capture su muestra biométrica (factores conductuales).
- La utilidad, definida como la característica para evaluar y predecir el rendimiento de un sistema de reconocimiento biométrico.

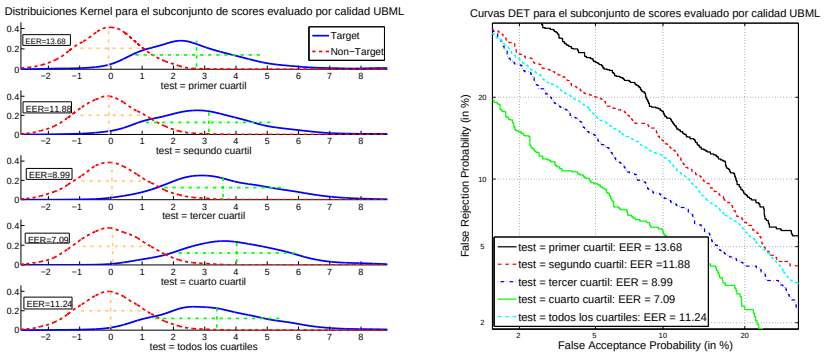

Figura 1. (Izquierda) Representación de las distribuciones *target* y *non target* en función del cuartil elegido para la calidad UBML del fichero de test mostrando su consecuente desalineamiento. Cabe destacar que la calidad del modelo no se tiene en cuenta en esta representación (los subconjuntos de scores han sido agrupados por calidad de test sin importar la del modelo). (Derecha) Curvas DET (*Detection Error Trade-off*) equivalentes a la figura anterior.

Si bien es interesante estudiar estos tres criterios que definen de forma concisa la calidad relativa de una muestra y por lo tanto el rendimiento de un sistema, este trabajo pretende evaluar mediante la *utilidad* [10] el rendimiento global del sistema bajo los diferentes algoritmos propuestos en la sección 4. Por lo tanto, es importante remarcar que los resultados obtenidos fruto de este trabajo (sección 5) dependerán en gran medida de la disponibilidad de unas buenas condiciones de fidelidad (tipo de dispositivo de adquisición, supervisión en la adquisición, tasa de compresión, etc.), del carácter del individuo (cooperación del sujeto, estado emocional, etc.) y de una base de datos apropiada, es decir, de máxima variabilidad y con un número amplio de muestras.

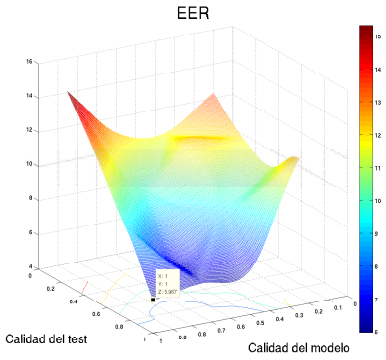

Figura 2. EER en función de la calidad del modelo y del test para el subconjunto de scores dependientes de la calidad UBML condición *tel-mic*.

Por lo tanto, y habiendo enunciado algunos hechos determinantes que influyen en la degradación de la calidad de una muestra de voz, se han seleccionado 2 tipos de indicadores por su impacto o dependencia con el rendimiento del sistema y por su coherencia con diferentes bases de datos y sistemas [5]: la *Relación Señal a Ruido* (SNR) y la *Universal Background Model Likelihood* (UBML) definida recientemente en [5].

El efecto de esta última se puede apreciar en la figura 2, en la que se observa la tendencia del EER en función de la calidad UBML del modelo y del test: a medida que la calidad el sistema aumenta la discriminación del sistema mejora. Como se especificará en 3.2 y se propone en [10] la calidad debe ser mapeada entre 0 y 1: dado que el número de ficheros con una calidad extrema es limitado (existen muy pocos ficheros de voz con muy buena o muy mala calidad) y estos son requeridos para la evaluación del rendimiento de los algoritmos propuestos se ha procedido a homogeneizar el número de ficheros ordenando los archivos por valor de calidad y tomando más o menos ficheros de una calidad determinada hasta alcanzar el porcentaje de ficheros correspondiente al 25 %, del 25 % al 50 %, del 50 % al 75 % y de éste al 100 % (4 cuartiles) de la calidad total. Por lo tanto quedarán definidos 4 bloques de calidad representados como 0,25 para el primer cuartil y 0,50, 0,75 y 1 para los restantes.

Dado que las bases de datos existentes en general no son lo suficientemente ricas en cuanto a variabilidad para implementar técnicas como ésta se recurre a otros métodos de normalización y calibración como la regresión logística, que tratan de calibrar el sistema en función de la calidad del enfrentamiento para reducir el desalineamiento de las distribuciones *target* y *non target* como se representa en la figura 1. Nótese que las distribuciones nombradas hacen referencia a un sistema automático de reconocimiento trabajando en modo verificación en el que se cumple la hipótesis de que el usuario y el fragmento de voz a evaluar son la misma persona y en el que no, de forma respectiva.

3. Medidas de calidad empleadas

3.1. Relación señal a ruido (SNR)

Como su nombre indica la SNR expresa la relación entre la potencia de la señal de voz y la potencia de ruido que la corrompe. Por lo tanto, queda definida mediante la siguiente fórmula:

$$SNR = 10 \log \left(\frac{E_{voz}}{E_{silencio}} \right) \tag{1}$$

siendo E_{voz} y $E_{silencio}$ la energía media de las zonas de voz y silencio respectivamente del fragmento de audio.

El principal problema de utilizar esta medida de calidad es que la fiabilidad de ésta dependerá de la precisión del detector de actividad de voz (VAD) siendo una no muy buena referencia de calidad si el diseño de éste no es el apropiado. No obstante es una medida ampliamente extendida y utilizada.

Siguiendo las recomendaciones de [10] y [11], para trabajar de forma homogénea con cualquier tipo de calidad es necesario expresar todo indicador de degradación en un rango entre 0 y 1 mediante una función de mapeo $Q(x)$, siendo 0 el valor mínimo de calidad y 1 el máximo [5]:

$$Q_{SNR}(x) = \frac{x}{60} \tag{2}$$

donde x corresponde al valor de SNR obtenido en un rango de $(0-60)dB$.

3.2. Similitud a un modelo de habla universal (UBML)

La UBML es una medida de calidad que trata de aproximar la similitud de una locución al modelo de habla universal utilizado para la generación del modelo estadístico de un locutor. Es una medida considerada de forma reciente en [5] que se extrae en los sistemas GMM (*Gaussian Mixture Model*) para calcular la puntuación de similitud:

$$S(O,\lambda_t) = log\left(p(O,\lambda_t)\right) - log\left(p(O,\lambda_{UBM})\right) \tag{3}$$

donde $S(O,\lambda_t)$ es el score o puntuación de similitud entre el modelo del locutor y el modelo universal y $p(O,\lambda_t)$ y $p(O,\lambda_{UBM})$ son las funciones densidad de probabilidad del modelo de usuario y universal respectivamente. Por lo tanto, la UBML queda definida mediante:

$$UBML = log\left(p(O,\lambda_{UBM})\right) \tag{4}$$

y cuya función de mapeo es:

$$Q_{UBML}(x) = \frac{x+13}{8} \tag{5}$$

donde x corresponde al valor de la $UBML$ obtenido en el rango de $(-13,-5)$.

4. Algoritmos de compensación propuestos

Los métodos que en esta sección se describen pretenden compensar el rendimiento del sistema a nivel de score dado el desalineamiento de las distribuciones (figura 1) como efecto de la variabilidad de la calidad explicado en la sección anterior. Para ello se han evaluado 3 algoritmos diferentes basados en regresión logística, ya utilizada en reconocimiento de locutor para fusión y calibración [6][12][13].

4.1. Regresión logística lineal de dos dimensiones (2D-LLR)

El método LLR (*Linear Logistic Regression*) es un algoritmo de compensación que transforma el conjunto de puntuaciones generadas bajo unas condiciones de calidad mediante un modelo de regresión lineal en el logaritmo de una relación de verosimilitud [6], el cual puede definirse de la siguiente manera:

$$x_{i,j}^{Norm} = \log \frac{P_{i,j}(x_{i,j}|T)}{P_{i,j}(x_{i,j}|NT)} = \alpha_{i,j} \cdot x_{i,j} + \beta_{i,j} \tag{6}$$

Los pesos $\alpha_{i,j}$ y $\beta_{i,j}$ se obtienen de las puntuaciones de entrenamiento [6][1] de las comparaciones de los modelos del cuartil de calidad i con los ficheros de test de calidad del cuartil j. La puntuación o score a compensar $x_{i,j}$ presenta el mismo cuartil de calidad que las puntuaciones de entrenamiento. Por lo tanto, ya que dicho algoritmo es dependiente de la calidad del modelo y del test se puede decir que la regresión logística seguida presenta dos dimensiones dando de esta manera nombre al algoritmo implementado.

Por último. destacar que este algoritmo se puede generalizar para cualquier valor de calidad realizando algún tipo de interpolación de los pesos $\alpha_{i,j}$ y $\beta_{i,j}$ (por ejemplo cúbica).

4.2. Regresión logística bilineal (BLR)

Este método está basado en [12], donde se tomaba como información complementaria información lingüística del locutor y puede definirse como sigue:

$$x_{i,j}^{Norm} = \alpha \cdot x_{i,j} + \sum_{k=1}^{K} \alpha_k \cdot \lambda_k \cdot x_{i,j} + \beta \tag{7}$$

donde α, α_k y β son ahora pesos fijos para todos los posibles conjuntos de scores dependientes de calidad, y λ_k corresponde a la información de las calidades del modelo y del test. Para la realización de este trabajo se ha definido dicha información de dos maneras diferentes dando lugar a dos algoritmos diferentes:

- BLR tipo 1: $\lambda_1 = Q_m$ y $\lambda_2 = Q_t$, donde la información complementaria corresponde a la calidad mapeada del modelo Q_m junto con la del test Q_t.
- BLR tipo 2: $\lambda_1 = \sqrt{Q_m \times Q_t}$, donde la información complementaria corresponde con la media geométrica de las calidades.

5. Experimentos

5.1. Base de datos y protocolos

El organismo norteamericano NIST (*National Institute of Standards and Technology*) [9] organiza evaluaciones bianuales abiertas de carácter competitivo en el que se elaboran bases de datos y se definen una serie de tareas o protocolos para medir de manera objetiva el rendimiento, bajo las mismas condiciones, de los sistemas presentados. La base de datos utilizada para el desarrollo de este trabajo es la de NIST SRE 2008 [1], la cual ofrece un importante desafío en cuanto a compensación de variabilidad de calidad se refiere. El protocolo seguido para la realización de estos experimentos es el *short2-short3* de la misma evaluación, que comprende archivos de audio conversacionales capturados de un

[1] El toolkit FoCal ha sido usado para el entrenamiento de la LLR. http://sites.google.com/site/nikobrummer/focal

canal telefónico o microfónico de 5 minutos de duración, de los cuales aproximadamente 2.5 minutos son de cada locutor de la conversación una vez suprimidos los silencios, y datos microfónicos en formato *interview* o entrevista en los cuales la mayoría de audio de los 3 minutos de duración corresponde al entrevistado.

El trabajo llevado a cabo presenta 4 condiciones de evaluación o escenarios diferentes:

- *tel-tel*, en el que el modelo de entrenamiento y el fichero de test han sido adquiridos de un canal telefónico.
- *tel-mic*, en el que el fichero de audio con el que se entrena el modelo del usuario a identificar ha sido adquirido a través de un canal telefónico y el fichero de test mediante un micrófono.
- *mic-tel*, en el que el modelo se ha extraído de una grabación con micrófonos y el archivo de enfrentamiento se ha capturado a través de la red telefónica.
- *mic-mic*, en la que el modelo y fichero de test han sido capturados con un dispositivo microfónico.

Nótese que las grabaciones telefónicas presentan diversos factores de degradación de la señal principalmente relacionados con la distorsión del canal de comunicación mientras que las muestras microfónicas se ven más afectados mediante otros parámetros como la fidelidad del dispositivo de captura, la distancia al micrófono, las condiciones ambientales, etc.. Tanto los ficheros de entrenamiento como los de enfrentamiento, presentan un filtrado Wiener ya que se ha demostrado que este tipo de procesado ayuda a mejorar el rendimiento de sistemas que trabajan con este tipo de muestras.

Por último, remarcar que dado el carácter preliminar del estudio los algoritmos propuestos se han entrenado con los mismos datos que la evaluación.

5.2. Sistema de desarrollo

El sistema de desarrollo sobre el cual se ha medido el rendimiento de los 3 algoritmos propuestos es el sistema presentado por el grupo ATVS en la evaluación NIST SRE de 2008. Este sistema está basado en modelo GMM (*Gaussian Mixture Model*) de 1024 mezclas y 19 parámetros MFCC (*Mel Frecuency Cepstral Coefficients*), adaptados de un UBM (*Universal Background Model*) entrenado con una gran cantidad de datos provenientes de las evaluaciones NIST hasta NIST SRE 2006. Dicho sistema también incluye compensación de canal mediante técnicas basadas en *feature warping* y *factor analysis* aplicando adaptación NAP (*Nuisance Attribute Projection*) [14] a los modelos GMM. Por último, destacar el uso de *T-Norm* para normalizar los scores [15].

6. Resultados

El cuadro 1 muestra en términos de EER la mejora de los algoritmos propuestos en función de la condición de NIST SRE 2008 evaluada. El método

2D-LLR que realiza una LLR por subconjunto de scores en función de la calidad del modelo Q_m y la calidad del test Q_t ofrece prácticamente en todos los casos una mejora en términos de EER en cuanto al EER original y al EER mejorado por los otros algoritmos. En cualquier caso, sólo sufre un empeoramiento para la condición *tel-tel* de UBML, el cual puede ser considerado despreciable por su proximidad a 0. No obstante. los métodos BLR-1 y 2 ofrecen peores resultados salvo para esta misma condición, en el que presentan una mejora por encima del resultado obtenido usando 2D-LLR siendo mejor para el segundo caso en el que la información complementaria es la media geométrica de las calidades del modelo y del test.

Los mismos resultados son representados de forma gráfica en la figura 3, en las que se han separado para los 3 algoritmos las 2 condiciones con peor rendimiento, *tel-tel* y *mic-tel* y las 2 con mejor rendimiento, *tel-mic* y *mic-mic*.

			UBML		**SNR**	
Compensación	Condición	EER	EER_{norm}	Mejora EER	EER_{norm}	Mejora EER
2D-LLR	tel-tel	7, 80	7, 82	−0, 35	7, 64	1, 97
	tel-mic	11, 24	10, 24	**8,82**	10, 45	**7,01**
	mic-tel	11, 68	11, 66	**0,16**	11, 60	**0,72**
	mic-mic	8, 51	8, 19	**3,76**	8, 43	0, 94
BLR tipo 1	tel-tel	7, 80	7, 68	1, 51	7, 69	**1,35**
	tel-mic	11, 24	10, 47	6, 79	10, 46	6, 94
	mic-tel	11, 68	11, 77	−0, 81	12, 08	−3, 46
	mic-mic	8, 51	8, 35	1, 93	8, 40	**1,3**
BLR tipo 2	tel-tel	7, 80	7, 63	**2,08**	7, 69	**1,35**
	tel-mic	11, 24	10, 50	6, 56	10, 50	6, 56
	mic-tel	11, 68	11, 80	−1, 05	12, 07	−3, 38
	mic-mic	8, 51	8, 32	2, 3	8, 40	**1,3**

Cuadro 1. Tabla resumen en la que se muestra la mejora en términos de EER del sistema en tanto por ciento de los algoritmos propuestos para las 4 condiciones y los subconjuntos dependientes de calidad.

7. Conclusiones y trabajo futuro

En este artículo se ha estudiado el rendimiento del sistema, en cuanto a términos de EER se refiere, en función de los subconjuntos de scores dependientes de las calidades UBML y SNR respaldadas en estudios anteriores como buenos indicadores de degradación. Para ello se han implementado varios algoritmos probados sobre NIST SRE 2008 con resultados prometedores.

Resumiendo los resultados obtenidos mediante esta base de datos cabe destacar que la mejora del 8,82 % evaluando los datos a partir de la calidad UBML en la condición *tel-mic* mediante el método 2D-LLR es considerable, lo cual incita a seguir realizando estudios en este sentido, ya que la fiabilidad y carácter de las muestras telefónicas son diferentes a las microfónicas produciendo un mayor desajuste en las distribuciones (ver media de las distribuciones *target* en figura 1). Otra resultado a destacar es que el método BLR-2 mejora en un 2,08 % en la condición *tel-tel*, donde el método 2D-LLR no obtiene buen rendimiento para el subconjunto de scores obtenidos a través de la calidad UBML. Dicho método es una variación elegante del método 2D-LLR aunque con mayores problemas en

cuanto a coste computacional y divergencia. Por último, en cuanto a conclusiones extraídas se refiere, destacar que para un sistema GMM obtener la calidad UBML, en la cual se obtienen mejoras ligeramente superiores a usar la SNR, no implica ningún coste adicional ya que es un hecho que debe realizarse de forma imperativa si se pretende lanzar un score como resultado final a evaluar.

Como trabajo futuro hay que remarcar que sería interesante entrenar los pesos de compensación de las regresiones logísticas mediante un conjunto de scores diferente al de test, no como en este estudio, en el que dichos parámetros contenían información de los datos a normalizar debido a que se ha utilizado la base de datos de 2008 a posteriori. Estudios en esta línea se están llevando a cabo en el ATVS. Otra línea de investigación a seguir sería evaluar el rendimiento de estos algoritmos bajo otras técnicas de normalización como ZT-Norm que trata de normalizar las distribuciones *target* y *non target* a través de los modelos de entrenamiento, o el uso de cohortes dependientes de la calidad a la hora de normalizar.

Referencias

1. NIST, "2008 speaker recognition evaluation plan: http://www.nist.gov/speech/tests/sre/2008/index.htm," 2008.
2. F. Alonso-Fernandez *et al.*, "A comparative study of fingerprint image-quality estimation methods," *IEEE Trans. on Information Forensics and Security*, vol. 2, no. 4, pp. 734–743, December 2007.
3. V.Grancharov *et al.*, *Speech Quality Assessment Chapter 5*, Springer, 2007.
4. P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.
5. Alberto Harriero *et al.*, "Analysis of the utility of classical and novel speech quality measures for speaker verification," in *Proceedings of International Conference on Biometrics*. June 2009, vol. 5558 of *LNCS*, pp. 434–442, Springer.
6. Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
7. M. A. Przybocki, A. F. Martin, and A.N. Le, "NIST speaker recognition evaluations utilizing the mixer corpora-2004, 2005, 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.
8. NIST, "Biometric sample quality standard draft (revision 4), 6.," 2008.
9. "NIST speech group website: http://www.nist.gov/speech," .
10. Patrick Grother and Elham Tabassi, "Performance of biometric quality measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 531–543, 2007.
11. D. Garcia-Romero *et al.*, "Using quality measures for multilevel speaker recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 192–209, 2006.
12. L. Ferrer *et al.*, "System combination using auxiliary information for speaker verification," in *Proc. of ICASSP*, Las Vegas, Nevada, USA, 2008, pp. 4853–4856.
13. N. Brümmer and J. du Preez, "Application independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.
14. A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. of ICASSP*, 2005, pp. 629–632.
15. R. Auckenthaler *et al.*, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

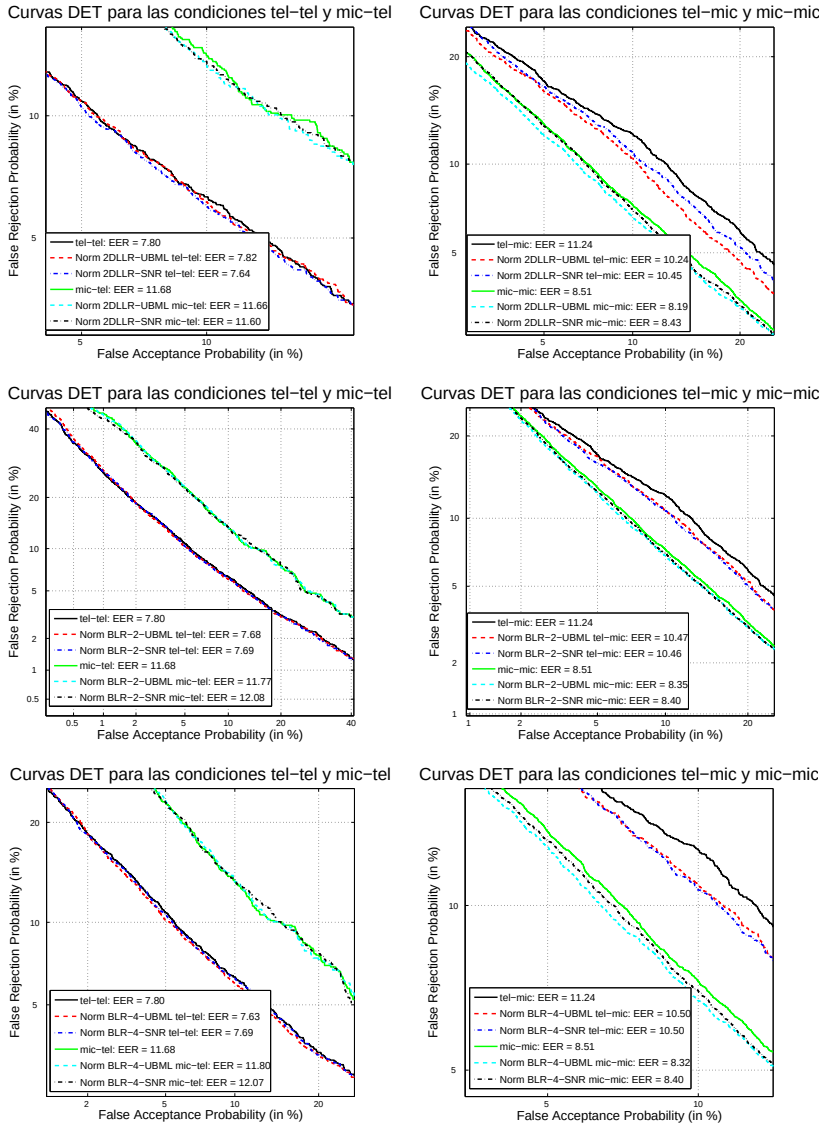

Figura 3. Rendimiento de los algoritmos mostrado en formato curva DET (*Detection Error Trade-off*). A la izquierda se muestran las curvas para las condiciones *tel-tel* y *mic-tel* que corresponden a las de peor mejora y a la derecha las de *tel-mic* y *mic-mic* con resultados prometedores. Notar que la escala en cada gráfica es diferente.