Reconocimiento del hablante para el español de México usando i-vectors y PLDA

Abel Herrera Camacho
Departamento de Procesamiento de Señales
Facultad de Ingeniería, UNAM
México
abelhc@hotmail.com

Arturo Rivera García
Alumno de la carrera de Ingeniería en Computación
Facultad de Ingeniería, UNAM
México
bassmetal_riga@hotmail.com

Resumen. **En este trabajo se desarrolló un sistema de reconocimiento del hablante basado en la parametrización i-vectors con reducción de parámetros usando PLDA (Probabilistic Linear Discriminant Analysis). Esta combinación de técnicas es considerada muy exitosa. La parametrización con i-vectors es considerada muy completa y robusta, se han diseñado otras parametrizaciones semejantes en tiempos más recientes con resultados ambiguos. Con PLDA se reduce la dimensión de los i-vectors. La combinación de estas técnicas aplicada a señales de voz de poca duración mejora el desempeño respecto a la técnica de Modelos de mezclas gaussianas (GMM, por sus siglas en inglés) con MFCC (mel frequency cepstral coding). Para verificar el desempeño del sistema desarrollado se utilizaron i-vectors con PLDA y PLDA con reducción de la dimensión. El sistema desarrollado se aplicó al español de la Ciudad de México utilizando el corpus "Valquiria"; este corpus fue diseñado y creado en la UNAM. El corpus contiene grabaciones de hombres y mujeres originadas desde teléfonos público, fijo y celular. Las pruebas en grabaciones cortas es relevante por su aplicación al reconocimiento de hablantes en condiciones forenses. Este trabajo adquiere mayor validez porque se aplica a un corpus del español de México, con una técnica muy sólida en que es i-vectors-PDLA.**

Keywords—Reconocimiento del hablante, i-vectors, Modelos de mezclas gaussianas, Probablistic Linear Discriminant Analysis, corpus.

I. INTRODUCCIÓN

Hasta la década de los 80's, un aspecto insustituible en las aplicaciones de voz fue comprimir la señal de voz (la compresión está inserta en la codificación de voz). La señal de voz muestreada contiene muchos bits, que pueden comprimirse al transformar la señal en otro espacio. Otro aspecto crucial fue que la señal de voz (una frase cualquiera) se analizaba por pequeños segmentos llamados tramas ("frames") que iban del inicio al fin de la señal de voz. Las muestras en cada trama constituyen un vector, al comprimirla se obtienen nuevos vectores llamados *vectores de características* ("features")

Lo expresado en el párrafo anterior todavía es válido en aplicaciones de voz que se manejan en tiempo real con pequeños procesadores, como en el análisis por síntesis que realizan los teléfonos celulares. Inclusive es útil en identificación o verificación del hablante en tiempo real. Pero ciertas aplicaciones como el reconocimiento del hablante en condiciones forenses, se pueden tener días o semanas para decidir si una grabación de un sospechoso coincide con las grabaciones que se grabaron previamente. En este caso y otros, dadas las capacidades de almacenamiento y procesamiento actuales ya no son tan relevantes los objetivos de compresión y análisis de voz por tramas.

El primer gran cambio en la codificación de voz en identificación de un hablante se dio usando modelos de mezclas gaussianas (GMM, por sus siglas en inglés), en 1990 por Reynolds [1], que amplió sus resultados en 1995 [2] . Aquí se tienen grandes matrices que representan a cada hablante y se comparan con la voz que se pretende identificar. Una variante llamada "Universal background model" impide que sea una búsqueda muy lenta [3,4]. En este trabajo se usará la técnica GMM como base de comparación para i-vectors/pdla. La técnica GMM a usar se combinará con la codificación MFCC y la clasificación de máximo valor esperado (ME) [5,6].

Los i-vectors es el método a aplicar en este trabajo. Este método fue propuesto por Dehak en 2011 [6], para proporcionar una representación de varios hablantes en varios canales. En fechas más recientes se han aplicado variantes de i-vectors llamados "e-vectors" [7] y "x-vectors" [8], el primero en un sistema de reconocimiento del hablante semejante al de i-vectors y el segundo en un sistema basado en redes neuronales profundas (DNN, por sus siglas en inglés), ambos propuestos en 2018.

La extracción de los tamaños que tendrían los i-vectors fue motivado por la existencia de los vectores basados en "Joint Factor Analysis" (JFA) [9] lo cuales son de gran dimensión. El enfoque de JFA modela al hablante y la variabilidad del canal por separado. los trabaja por separado; mientras que el método de i-vectors modela un solo espacio de baja dimensión que contiene toda la variabilidad.

Como los i-vectors están basados en un solo espacio de variabilidad que contiene tanto la información de hablante como la variabilidad de canal, se requiere de técnicas de compensación para limitar los efectos de la alta variabilidad

presentada en el canal de los i-vectors que representan al hablante. La compensación del canal juega un rol importante en los sistemas de verificación del hablante. Cuando el método de los i-vectors fue introducido, también se incluyeron técnicas de compensación del canal. Poco después, Kenny y otros al observar que cada grabación puede ser representada por un i-vector de baja dimensión, introdujeron PLDA [10] en 2011. Estudios preliminares de PDLA exisitieron desde 2007 [11].

Las técnicas basadas en DNN's de reconocimiento del hablante inician en 2014 [12,13]. Destacan los trabajos del grupo de Snyder [14,15,16]. En la evaluación NIST 2018 es ya la técnica predominante [17].

II. LOS I-VECTORS

A. *Concepto de i-vector*

Los i-vectors es una técnica que define un solo espacio que contiene tanto la variabilidad del hablante como la variabilidad de la sesión o del canal, llamado espacio de variabilidad total.

El método para modelar un hablante usando i-vectors se define de la siguiente manera. Dado un conjunto de grabaciones de entrada X_s, su modelo en i-vectors se establece por:

$$M_s = m + Tw \quad (1)$$

La salida M_s es la suma de una componente dependiente del hablante (m) más una componente dependiente del canal (Tw). Este modelo de ruido aditivo es muy conocido en procesamiento de señales [18]. El término m es el supervector del hablante sin considerar el canal, puede ser el vector GMM-UBM. T es una matriz rectangular y w es un vector aleatorio que tiene una distribución gaussiana $N(0,1)$, los componentes del vector w son los factores totales. los cuales son llamado como vectores de identidad o i-vectors..

En este modelo, M se asume normalmente distribuida con un vector de medias m y una matriz de covarianza TT^t.

En la figura 1 se muestra un sistema de identificación/verificación del hablante por i-vectors [19]. Las voces del corpus son necesarias para construir el supervector del hablante sin considerar el canal.

Los bloques de extracción de carácterísticas se refieren a la obtención de coeficientes MFCC o semejantes. Las parametrizaciones de las diferentes voces se refieren a la construcción de supervectores tipo GMM-MFCC. La extracción de i-vectors se mostrará en esta sección.

La manera para entrenar a la matriz T es exactamente la misma a como se entrena la matriz V en el método JFA [20], excepto por una diferencia muy importante, en el entrenamiento de la matriz V la cual contiene los eigenvoices, todas la grabaciones de una persona son consideradas para ser la misma persona, sin embargo aquí en i-vectors se considera todo el conjunto de grabaciones por diferentes hablantes.

El factor total w es una variable oculta, la cual puede ser definida por una distribución posterior condicionada por las estadísticas de Baum Welch para una expresión dada. Esta distribución posterior es una distribución gaussiana y el promedio de esta distribución corresponde exactamente para nosotros los i-vector. De manera similar para las estadísticas de Baum Welch son extraídas usando el UBM [23].

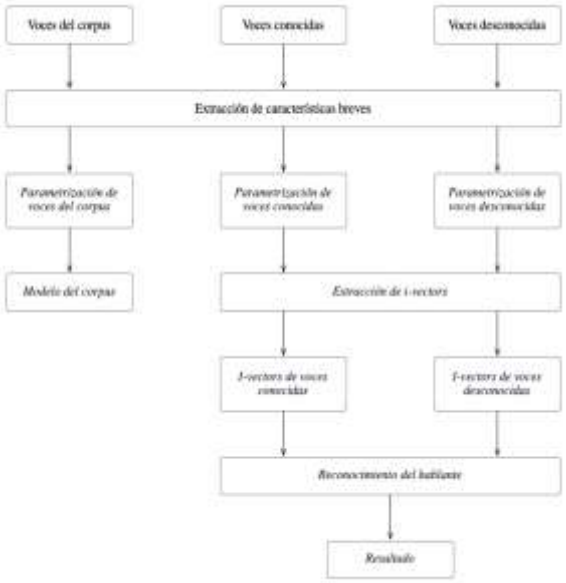

Figura 1. Sistema de reconocimiento del hablante con i-vectors [19].

En la figura 2 se muestra un ejemplo de la obtención de los primeros 5 i-vectors de entre 139 hablantes, cada i-vector contiene 50 características que son las que representan al hablante dentro de todo el espacio de variabilidad, se pueden representar dentro de una matriz de i-vectors donde el número de filas representa el tamaño del i-vector y el número de columnas es igual al número de personas con las que el sistema será entrenado.

B. *Entrenamiento de i-vectors*

La matriz de variabilidad total T se entrena con un proceso iterativo de tipo EM [6,20]. Los datos de entrenamiento consisten en grabaciones de todos los hablantes que participarán de la verificación o Identificación.

Luego de inicializar la matriz T aleatoriamente, se calculan las estadísticas de Baum-Welch de orden cero y uno para cada grabación u de entrenamiento. Se calculan las estadísticas para cada grabación porque para el entrenamiento de la matriz T.

Luego, para una grabación u definida por una secuencia de N vectores de características $x_1, \ldots, x_N$ pertenecientes al hablante s donde cada vector es de tamaño D, y un UBM $\Lambda^{(b)} = \left\{\lambda_i^{(b)}, \mu_i^{(b)}, \Sigma_i^{(b)}\right\}_{i=1}^{C}$ con C componentes gaussianos, se calculan las estadísticas de Baum-Welch de orden cero y de primer orden.

$$N_c = \sum_t x_k \; ; \qquad F_c = \sum_t (x_k) * x_k \quad (2)$$

Se procede al cálculo de las estadísticas de Baum-Welch, estas estadísticas son el resultado de un algoritmo iterativo donde en cada iteración se aumente la probabilidad de generar una nueva una expresión. El algoritmo de Baum-Welch pertenece a la familia de métodos EM, los cálculos de valores esperados y maximización son realizados de forma simultánea.

$$N_c\,(u) = \sum_{k\,=\,1}^{N} \gamma_k(c) \tag{3}$$

$$F_c\,(u) = \sum_{k\,=\,1}^{N} \gamma_k(x_k) \tag{4}$$

$$\vec{F}_s\,(u) = \sum_{k\,=\,1}^{N} \gamma_c(x_k)(x_k - \mu_c) = F_c\,(u) - N_c\,(u) * \mu_c \tag{5}$$

donde las estadísticas de orden cero N_c nos indican la probabilidad posterior $(\gamma_k(c))$, y las estadísticas de primer orden F_c nos indican la probabilidad posterior de que x_t sea generado por un componente de mezcla de $i = 1, \ldots, C$.

La ecuación para extraer un i-vector que represente a un hablante *h*, puede obtenerse de la siguiente ecuación:

$$w(h) = (I + T^t \Sigma^{-1} N_h T)^{-1}\, T^t \Sigma^{-1} F_h \tag{6}$$

De esta última ecuación, I es la matriz identidad, N_h es una matriz diagonal donde sus elementos de la diagonal son bloques dados por $N_c(u)I\,(c = 1, \ldots, C)$, F_h es un vector formado por la concatenación de las estadísticas de primer orden centradas de Baum-Welch para una repetición dada, Σ es la matriz de covarianza que modela la variabilidad residual no capturada por la matriz de variabilidad total *T*. En la práctica se sustituye esta matriz por las matrices de covarianza del modelo UBM.

En la figura 2 está la representación de una matriz de variabilidad total la cual contiene tanto las características de cada persona como las características de las fuentes no deseadas, así como el ruido del canal. El número de filas es igual al tamaño del i-vector y las columnas de la matriz es igual al número de coeficientes MFCC trabajados por el numero de componentes gaussianas implementadas en el modelo UBM, para este ejemplo son 13 coeficientes MFCC y 512 gaussianas.

III. PDLA

A. *Concepto*

Después de haber realizado la extracción de los i-vector se aplica PDLA, Los objetivos son complementar la información de las personas debido a que podría no tenerse suficientes datos y manejar las variabilidades en la voz y el canal. Se aplica PDLA después de haber usado Linear Discriminant Analysis (LDA).

LDA ayuda a cubrir dos aspectos importantes del sistema que son, la reducción de la dimensión de los datos y normalización de los datos. El objetivo de LDA es buscar dentro de las matrices de dispersión unos nuevos ejes que nos indican la posición del i-vector maximizando la variabilidad dentro de las grabaciones de la persona y minimizando la variabilidad en el canal, de esta manera lograr tener una menor dimensión en los vectores resultantes que nos representen mejor a cada clase.

50x6656 single

	1	2	3	4	5	6	7	8
1	1.2986e+03	304.3454	482.2730	-450.1388	2.5091e+03	1.6371e+03	925.3516	-2.5767e+03
2	-2.2594e+03	2.0181e+03	-2.0221e+03	-1.0075e+03	-325.7390	-637.2549	-704.2054	3.7384e+03
3	107.7325	293.6382	-495.1253	-279.9082	1.1055e+03	-614.4387	-232.7239	-590.6366
4	232.4364	-1.6373e+03	1.4847e+03	-282.5307	421.1766	-520.6458	-130.8683	-1.0864e+03
5	547.6380	-1.9757e+03	-226.6517	939.8602	2.3722e+03	1.0711e+03	1.4484e+03	-754.9155
6	35.8838	296.5356	-681.2608	-187.7014	-507.7151	-372.5364	-188.0252	145.1895
7	-597.7451	-957.7278	719.1306	-946.6612	3.3825e+03	-755.8773	614.7466	82.6918
8	-312.9457	-791.7435	-437.4660	505.4991	2.8066e+03	80.8423	-66.2261	993.4421
9	811.0988	-1.4177e+03	-284.4875	3.1892e+03	-621.3872	825.9953	-364.6298	-1.3057e+03
10	34.1245	-2.0976e+03	295.4169	-1.7331e+03	677.2914	-365.7682	-550.8615	300.4455
11	1.2385e+03	696.1030	344.8557	2.0244e+03	-3.4723e+03	1.3271e+03	-22.4182	-1.6238e+03
12	482.4647	301.6622	1.2305e+03	1.0507e+03	-2.4190e+03	-589.4385	-815.5968	-614.7121
13	555.9335	-907.1628	-1.3490e+03	1.0679e+03	377.1248	2.3540e+03	707.7141	-718.0334
14	-877.5187	1.5299e+03	1.0785e+03	-2.0920e+03	2.2447e+03	-1.3019e+03	-53.5703	340.3623
15	-438.3336	433.4825	-548.5379	-1.4573e+03	1.1031e+03	327.4803	154.2576	205.7384
16	1.1012e+03	-3.9596e+03	2.0743e+03	-322.6945	-1.2305e+03	-78.8346	651.8399	-2.2738e+03
17	-383.5022	2.8259e+03	-462.9336	-1.0883e+03	-1.3240e+03	177.9285	800.0921	1.0088e+03
18	36.2419	682.0635	-180.4642	-194.1682	-2.5457e+03	35.4790	-390.9987	-38.2686
19	-415.0982	-447.1557	761.8611	-1.1500e+03	-1.5562e+03	-1.5044e+03	-508.9327	536.8196
20	-974.0378	-2.7732e+03	190.6489	-1.6524e+03	1.4943e+03	-147.2917	-40.1475	2.4879e+03

Figura 2. Fragmento de la matriz de variabilidad T.

B. *LDA*

Como primer paso consiste entonces en el cálculo de encontrar las matrices S_B y S_W, donde S_B es una matriz de "dispersión" que nos indica la variabilidad que existe entre cada individuo tratando de maximizar esa diferencia entre los hablantes. S_W es la otra matriz de dispersión que trabaja con los datos presentes dentro de cada clase tratando de minimizar la separación entre ellos de tal manera que representen mejor al individuo.

Para poder obtener la matriz S_B se debe realizar el cálculo del vector de media de cada clase y generar con los vectores de medias un vector de media global, para después calcular la separación entre clases. Para S_W el proceso es semejante ya que se debe calcular la distancia entre la media de una clase y los ejemplos que existen en esa clase, y como tercer paso se debe construir un nuevo espacio de baja dimensión que nos va a permitir maximizar la varianza entre las clases y minimizar la varianza dentro de la clase [21,22].

C. *LDA probabilístico*

Existe varios enfoques del algoritmo PLDA como el estándar, simplificado o gaussiano y de dos colas. En este trabajo se hace uso únicamente del enfoque simplificado o también llamado PLDA gaussiano.

El aplicar esta técnica junto a los i-vector nos permite discriminar mejor las voces de las personas descomponiendo los i-vectors en 3 partes que son, un vector de medias global, una matriz que contiene el espacio de variabilidad entre las clases y una matriz de residuo la cual contiene la variabilidadno capturada, estos tres componentes son el

modelo generativo que PLDA obtiene al trabajar con los i-vectors.

El método PLDA ayuda notablemente a la separación de las características de las personas de toda información de fuentes no deseadas.

En nuestros experimentos se hace un uso notable de la técnica de reducción de la dimensionalidad, al pasar de 138 a 40 dimensiones. Esto ayuda incluso a obtener mejores resultados.

IV. Experimentos y Resultados

Se recolectó una muestra con audios que cumplieran con una duración mayor a 3 minutos dentro del corpus Valquiria, con la finalidad de observar el desempeño los sistemas con audios de mayor duración con tan solo un entrenamiento de 10 segundos. En totalidad, se usaron 22 audios para mujeres (taba y 17 para hombres, se usaron 140 audios en entrenamiento, 50 i-vectors, 512 gaussianas, 13 MFCC y y una reducción a 40 diemensiones.

Se muestran los resultados en las tablas I y II.

TABLA I. Resultados para mujeres

GMM	Precisión GMM	PLDA-RD	Precisión PLDA-RD
18/22	73%	22/22	100%

TABLA II. Resultados para hombres

GMM	Precisión GMM	PLDA-RD	Precisión PLDA-RD
13/17	76%	17/17	100%

De las tablas se observa una mejora de la precisión PLDA respecto a GMM como se había esperado y que coincide con experimentos de otros autores.

Agradecimiento

Los autores expresan su agradecimiento a la DGAPA de la UNAM por su apoyo en los proyectos IG101804 de 2021 e IT101818 de 2020.

Referencias

[1] R. Rose y D.A. Reynolds, “Text independent speaker identification using automatic acoustic segmentation,” International Conference on Acoustics, Speech, and Signal Processing IEEE, pp. 293-296, 1990.

[2] D.A. Reynolds y R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, Jan 1995.

[3] D.A. Reynolds, “Comparison of background normalization methods for text-independent speaker verification,” European Conference on Speech Communication and Technology, pp. 963–966, 1997.

[4] D. A. Reynolds, T. F. Quatieri, y R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, vol. 10, no. 1–3, pp. 19–41, 2000.

[5] A. Herrera and A. Zúñiga, “Reconocimiento automático de hablantes en el ámbito forense usando MFCC’sy GMM’s.” In Memorias de la 26° Reunión de Otoño de Comunicaciones, Computación, Electrónica y Exposición Industrial, ROC&C’2016 de la IEEE Sección México, memoria USB, 2016.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” Transactions on Audio, Speech, and Language Processing, vol. 19, no 4, p. 788-798, 2011.

[7] S. Cumani and P. Laface. “Speaker recognition using e–vectors,” IEEE Transactions on Audio, Speech, and Language Processing 26 (4), 736-748, 2018.

[8] D Snyder, D Garcia-Romero, G Sell, D Povey, S Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition”. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5329-5333, 2018.

[9] P. Kenny, “Joint factor analysis of speaker and session variability: Theory and algorithms,” Technical Report CRIM-06/08-13, CRIM, Montreal, Quebec, Canada, 2005.

[10] M. Senoussaoui, P. Kenny, N. Brummer, E.D. Villiers, y P. Dumouchel, “Mixture of PLDA Models in I-Vector Space for Gender Independent Speaker Recognition,” Interspeech 2011, pp. 1–19, 2011.

[11] S.J.D. Prince and J.H. Elder, “Probabilistic Linear Discriminant Analysis for Inferences About Identity,” IEEE International Conference on Computer Vision, , No. iii, pp. 1–8, 2007.

[12] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, “Deep neural networks for extracting Baum-Welch statistics for speaker recognition,” in Proc. Odyssey, 2014.

[13] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 1695– 1699.

[14] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, “Improving speaker recognition performance in the domain adaptation challenge using deep neural networks,” in Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014, pp. 378– 383.

[15] D. Snyder, D. Garcia-Romero, and D. Povey, “Time delay deep neural network-based universal background models for speaker recognition,” in 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, 2015, pp. 92–97.

[16] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification.” in Proc. INTERSPEECH, Stockholm, Sweden, August 2017, pp. 999–1003.

[17] S.O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, y J. Hernandez-Cordero. The 2018 NIST Speaker Recognition Evaluation. INTERSPEECH 2019, pp. 1483-1487,2019

[18] B. Widrow and S. Stearns, *Adaptive Signal Processing*, Prentice-Hall, 1985.

[19] L. Lei y S. Kun. “Speaker Recognition Using Wavelet Packet Entropy, I-Vector, and Cosine Distance Scoring”. *Journal of Electrical and Computer Engineering*, vol. 2017, Article ID 1735698, 2017.

[20] J.A. Fredes Sandoval. “Estudio comparativo de técnicas para robustez de sistemas de verificación de locutor texto independiente”. Tesis de licenciatura. Universidad de Chile, 2015.

[21] P. Rajan, A. Afanasyeva , V. Hautamaki , y T. Kinnunen “From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification”. *Digital Signal Processing*, vol. 31, p. 93-101, 2014.

[22] A. Tharwat, T. Gaber, A. Ibrahim y A.E. Hassanien, “Linear discriminant analysis: A detailed tutorial”. *AI communications*, vol. 30, no 2, p. 169-190, 2017