Identificación forense de hablantes en español usando triplet loss

Dr. Iván Vladimir Meza Ruíz
Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas
Universidad Nacional Autónoma de México
Ciudad Universitaria, Ciudad de México
ivanvladimir@turing.iimas.unam.mx

David Emmanuel Maqueda Bojorquez
Facultad de Estudios Superiores Cuautitlán
Universidad Nacional Autónoma de México
Cuautitlán, Estado de México
emmaqueda@comunidad.unam.mx

***Abstract*—Este trabajo propone utilizar la configuración de una red profunda con triplet loss para la identificación forense de hablantes en español. Dentro del marco, entrenamos una red convolucional para producir representaciones vectoriales de cortes de espectrograma. Luego, probamos qué tan similares son estos vectores para un hablante dado y qué tan diferentes son en comparación con otros hablantes. En el futuro, esta será la base para el cálculo de la Radio de probabilidad, que es una piedra angular para la identificación forense.**

Keywords—lingüística forense, triplet loss, identificación de hablantes

I. Introducción

La pérdida de tripletes [1] se introdujo como método para entrenar a una CNN (Convolutional Neural Network) para producir una buena representación vectorial para la tarea de identificación facial. Triplet loss evalúa tres representaciones vectoriales de dos objetos (originalmente una imagen de la cara, en nuestro caso una diapositiva de audio). La primera (ancla) y segunda (positivo) representaciones corresponden a la misma identidad, mientras que la tercera representación corresponde a una segunda identidad (negativo). El objetivo de la pérdida de tripletes es hacer que las dos primeras representaciones sean cercanas en relación con la tercera.

Por otro lado, la identificación forense del hablante se centra en recopilar y cuantificar la evidencia para la identificación de una persona a través de su voz. Sin embargo, no se trata solo de emparejar dos grabaciones por su similitud, sino que en el caso del análisis forense también se tiene que cuantificar las posibilidades de que las grabaciones se confundan dentro de las grabaciones de hablantes de la población (tipicidad). En este trabajo proponemos medir las distancias entre hablantes y hablantes externos como un sistema para cuantificar la similitud y la tipicidad.

II. Justificación

La lingüística forense se ha definido tradicionalmente como la interfaz entre lengua y derecho [2]. Gibbons y Turell [3] proponen a la lingüística forense (visión compartida también por la Asociación Internacional de Lingüistas Forenses) como: el análisis del lenguaje jurídico y judicial, el estudio del lenguaje del procedimiento judicial y el lenguaje como evidencia lingüística (en el ámbito pericial).

Como pruebas judiciales, existen dos clases de grabaciones. Las indubitadas, que son de las que se tiene certeza de su fuente y de su legitimidad (ej. grabaciones tomadas por servicios periciales), y las dubitadas en las que se mantiene alguna duda, aunque parcial o total, sobre su autenticidad, integridad, fecha, formato, identidad de alguno de los hablantes o sobre su legalidad (ej. Intervenciones telefónicas, grabaciones de cámaras de vídeo etc.).

Las controversias pueden llegar a ser muy complejas; dado esto se debe precisar adecuadamente y con alta certeza la diferencia entre las grabaciones dubitables y las indubitables que son las que pueden estudiarse en el marco legal.

En este contexto el presente trabajo pretende encontrar un sistema el cuál mediante muestras de audio pueda encontrar representaciones vectoriales representativas, las cuales permitan determinar la identidad de un cierto hablante con una certeza lo suficientemente válida para que pueda ser usado en el marco legal.

III. Objetivos

A. Generales

- Crear un sistema, el cuál permita la identificación forense de hablantes por medio de muestras de grabaciones de voz, utilizando una CNN y la función de pérdida triplet loss.

B. Particulares

- Crear la infraestructura necesaria para el funcionamiento de la CNN.
- Implementar triplet loss.
- Proyectar datos de salida (Vectores 1D 1024).
- Implementar métodos de evaluación de las salidas (métricas).

IV. Hipótesis

El sistema propuesto será capaz de identificar la relación existente entre las representaciones vectoriales de las muestras de audio tomadas de un espectrograma de grabación completa de una manera eficaz, con un rango de confiabilidad lo suficientemente adecuado para ser usado dentro del marco legal.

V. Metolodología

En este trabajo utilizamos el conjunto de datos en español Voxforge [4]. **Tabla 1.** muestra algunas características de las grabaciones. Este corpus está compuesto por 2.180 hispanohablantes, 21.692 grabaciones que en promedio duran 8.25seg. A partir de esto, hicimos una división del 80% de entrenamiento, el 10% de validación y el 10% de pruebas.

2020/10/21

TABLA 1. VOXFORGE SPANISH CORPUS

Speakers	2,180
Recordings	21,692
Avg. Duration	8.25 s

La entrada de nuestra CNN es una porción de un espectrograma, dicho espectograma esta compuesto por 60ms e información de frecuencias de hasta 21 a 8.5kHz.

Esta configuración crea un parche de 200 × 256 pixeles, con él se alimenta a una capa de cinco capas convolucionales con 32 kernels (2 primeras capas) y 64 kernels (3 últimas capas), cada capa convolucional esta seguida de una capa de normalización por lotes, max pooling (tamaño 2) y una función de activación ReLU.

La salida de la red CNN es un vector de dimensión 1D 1024 que representa el segmento de audio de voz. En el caso de la función de pérdida (triplet loss), aplicamos tres márgenes diferentes de 0.2, 0.5 y 0.8, esto significa que una muestra de dos hablantes diferentes debe estar separada al menos por el margen establecido.

VI. RESULTADOS

Para medir el rendimiento de la red, se han calculado hasta el momento tres métricas: distancia promedio interna para muestras (**IAD – Inner Average Distance**), distancia promedio externa entre hablantes y centroides de muestra de otros hablantes (**OAD – Outter Average Distance**)[5].

Con estas dos métricas, se propone calcular un símil de la Razón de Verosimilitud (**LR – Likelihood Radio**). También calculamos el coeficiente de silueta medio (**MSC – Mean Silhoutte Coefficient**) que varía de -1 (peor resultado) a 1 (mejor resultado).

Tabla 2. muestra las puntuaciones principales de los diferentes márgenes. Como puede verse, cuanto menor sea el margen, mayor será la diferencia entre la distancia interior y exterior. Aunque para un margen pequeño las muestras son las más cercanas (**IAD**), para un margen mayor estas están más alejadas (**OA**).

TABLA 2. MÉTRICAS

Margen	IAD	OAD	LR	MSC
0.2	**0.449**	3.16	0.142	0.248
0.5	0.891	6.221	0.141	0225
0.8	2.04	**12.32**	**0.166**	**0.259**

En particular, nuestra propuesta beneficia a LR para mayores márgenes. Por otro lado, podemos observar que este comportamiento es confirmado por MSC que es mayor para un margen mayor. **Figura 1.** muestra una proyección 2D de las muestras de 'test', podemos observar que las muestras del mismo hablante están agrupadas juntas, pero hay un espacio donde las agrupaciones se tocan o se superponen.

FIGURA 1. PROYECCIÓN VECTORIAL 2D DE MUESTRAS DE 218 HABLANTES PARA PRUEBA (MISMO COLOR, MISMO HABLANTE, MARGEN 0.8).

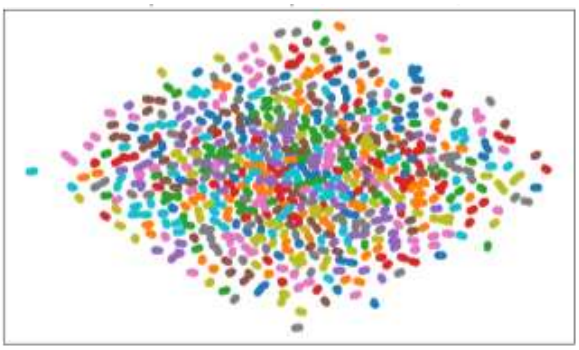
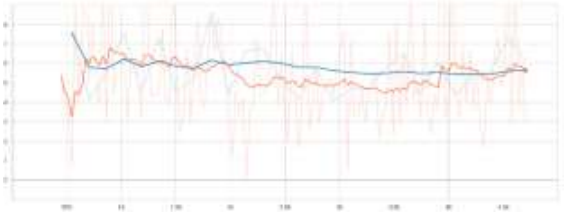


Figura 2. muestra la pérdida de la CNN para el conjunto de datos de entrenamiento y de validación de la misma, la cual muestra resultados satisfactorios mas no óptimos del entrenamiento y comportamiento de la CNN.

Cabe destacar que, aunque mejorables, los datos muestran concordancia con el entrenamiento de la CNN y el fin para la cual es usada, permitiendo una clasificación adecuada.

FIGURA 2. PÉRDIDA DE 'TRAINING' (NARANJA) Y 'VALIDATION' (AZUL)



VII. CONCLUSIONES

En este trabajo hemos explorado el uso de una configuración de red de triplet loss para la identificación forense de hablantes. Hemos establecido que un margen mayor para la pérdida da mejores resultados en términos de qué tan cerca están las muestras de un hablante y qué tan lejos están de los otros hablantes. Sin embargo, nuestros experimentos muestran que hay margen de mejora, ya que se puede notar la superposición entre los hablantes.

El trabajo futuro se centrará en la introducción de nuevos métodos para entrenar la pérdida de triplet loss, y añadir más métricas que permitan evaluar su desempeño.

VIII. REFERENCIAS

[1] Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N. (2016). Person re-identification by multichannel parts-based cnn with improved triplet loss function. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1335-1344.

[2] Cicres, Jordi. (2014). Comparación forense de voces mediante el análisis multidimensional de las pausas llenas. Revista signos, 47(86), pp. 365-384.

[3] Gibbons, J. & Turell, M. T. (Eds.) (2008). Dimensions of Forensic Linguistics. Ámsterdam/Filadelfia: John Benjamins.

[4] Hernández-Mena, C. VoxForge Spanish Corpus.

[5] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster 54 analysis. Journal of computational and applied mathematics, 20, pp. 53-65.