Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 101-108 recibido 01-12-2019 revisado 14-01-2020 aceptado 16-01-2020

Uso de técnicas basadas en one-shot learning para la identificación del locutor

Speaker Identification using techniques based on one-shot learning

Juan Chica, Christian Salamea
Universidad Politécnica Salesiana
Grupo de Investigación en Interacción, Robótica y Automática
jchicao@ups.edu.ec, csalamea@ups.edu.ec

Resumen: Un sistema para la identificación de locutor, para ser eficaz requiere una extensa cantidad de muestras de audio por cada locutor que no siempre es fácil de obtener. En contraste, sistemas basados en Meta-learning (en español, aprender a aprender) como one-shot learning utilizan una única muestra para diferenciar entre clases. En este trabajo se evalúa el potencial de un sistema de meta-learning para la identificación del locutor independiente del texto. En la experimentación se utilizan: espectrograma de mel, i-vectores y re muestreo (downsampling) para procesar el audio y obtener un vector de características. Este vector es la entrada de una red neuronal siamesa que se encarga de realizar la identificación. El mejor resultado se obtuvo al diferenciar entre 4 locutores con una exactitud de 0.9. Los resultados mostraron que el uso de técnicas basadas en one-shot learning tiene gran potencial para ser usados en la identificación del locutor y podrían ser muy útiles en ambientes reales como la biometría o ámbitos forenses por su versatilidad.
Palabras clave: Identificación del locutor, Independiente de texto, Meta Learning, N-way clasification, One-Shot learning, Redes Neuronales Siamesas, Voxceleb1

Abstract: A speaker identification system in order to be effective requires a large number of audio samples of each speaker, which are not always accessible or easy to collect. In contrast, systems based on meta-learning like one-shot learning, use a single sample to differentiate between classes. This work evaluates the potential of applying the meta-learning approach to text-independent speaker identification tasks. In the experimentation mel spectrogram, i-vectors and resample (downsampling) are used to both process the audio signal and to obtain a feature vector. This feature vector is the input of a siamese neural network that is responsible for performing the identification task. The best result was obtained by differentiating between 4 speakers with an accuracy of 0.9. The obtained results show that one-shot learning approaches have great potential to be used speaker identification and could be very useful in a real field like biometrics or forensic because of its versatility.
Keywords: Speaker Identification, Text independent, Meta Learning, N-Way clasification, One-Shot learning, Siamese Neural Network, Voxceleb1

1 Introducción

La voz en los seres humanos es producida a través de un largo proceso que se inicia desde el momento en el que se aspira aire, el cual va hacia los pulmones, pasando por la caja torácica y termina en el tracto vocal que es donde se articula la voz. Durante este proceso, la resonancia (formantes) que se genera en el tracto vocal queda registrada el contenido en frecuencia de la señal de la voz. Por otra que, al analizar la forma del espectro de la señal de voz es posible estimar la forma del tracto vocal de una persona y por consiguiente reconocer la identidad de un individuo particular. En el campo de la biometría este tipo de tarea es conocida como reconocimiento de locutor que

ISSN 1135-5948. DOI 10.26342/2020-64-12

incluye la verificación y la identificación de una persona por su voz.

La ventaja de utilizar la señal de la voz para tareas de reconocimiento de locutor recae en el hecho de que la voz es una señal natural emitida por las personas, por lo que la obtención de esta señal no se considera peligrosa ni invasiva (desde un punto de vista físico). Sin embargo, se debe considerar que la señal de audio que se obtiene en situaciones reales puede contener ruido ambiental (sonidos externos del entorno), siendo necesario el utilizar alguna técnica para que en una muestra de audio únicamente se tenga la información de interés (voz), o en su defecto que se enfatice dicha señal de interés. En los sistemas actuales de reconocimiento es necesario utilizar grandes cantidades de información con la finalidad de obtener un buen rendimiento, no obstante, nuevas técnicas desarrolladas basadas en Meta Learning (aprender a aprender), permiten disminuir la cantidad de datos necesarios para obtener un buen resultado.

En este trabajo se propone el uso de un sistema compuesto por dos etapas un Front-End y un Back-End para la identificación del locutor de texto independiente. En el Front End, se realiza un procesamiento de las señales de audio para extraer sus características más relevantes, aquí se utilizan tres enfoques diferenciados por el tipo de procesamiento realizado en la señal de audio. En el primer enfoque, se realiza un re-muestreo de la señal de audio, la cual pasa de 16000 Hercios (Hz) a 1000Hz de frecuencia de muestreo. En el segundo enfoque, se utiliza una técnica basada en coeficientes cepstrales y que ha sido ampliamente usada para el procesamiento del habla, esta técnica se denomina i-vectores y consiste en obtener un vector de características que permite representar la señal original que se encuentra en una alta dimensión, en un formato de baja dimensión sin perder las características de la alta. Por último, en el tercer enfoque se obtiene el espectrograma de mel en escala logarítmica de las señales de audio. La salida de cada enfoque se utiliza como entrada en el Back-End, donde se determina cual es el más óptimo para ser utilizado en el sistema. Por otra parte, en el Back-End se realiza la tarea de clasificación o de identificación al locutor como tal utilizando un modelo de redes neuronales siamesas que a su salida da como resultado una medida de similitud entre dos observaciones.

El objetivo de este trabajo es evaluar el potencial uso de técnicas basadas en one-shot learning en tareas de identificación de locutor de texto independiente. Nos hemos decantado por el uso de esta técnica debido a que una vez se ha entrenado el modelo, se puede identificar locutores que incluso no han sido presentados al modelo durante la fase de entrenamiento, lo cual es una gran ventaja en aplicaciones reales. Por otra parte, para evaluar el sistema en un ambiente real se seleccionó la base de datos de Voxceleb1, ya que contiene audios grabados en condiciones reales (diferentes ambientes) y las personas están utilizando su lenguaje natural (independencia del texto). Los resultados obtenidos muestran que este tipo de técnicas tienen gran potencial para ser utilizadas en la identificación de locutores, particularmente el mejor rendimiento se alcanza al diferenciar un locutor de entre 4 pues se obtiene alrededor de 0.9 (90%) de exactitud al identificarlo.

2 Fundamento Teoríco

2.1 Estado del Arte

Los primeros trabajos sobre reconocimiento de locutor se remontan a los años 50s, donde se ejecutaron los primeros intentos por construir una máquina para tareas de reconocimiento en base a los principios fundamentales de la fonética. En 1952, Davis, Biggulph y Balashek desarrollaron el primer sistema documentado que permitía identificar el habla aislada pronunciada por un solo locutor. Este sistema estaba basado en filtros analógicos que extraían medidas de las resonancias espectrales de tracto vocal por cada digito. En la década de los 70s las variaciones intra-hablantes comienzan a ser investigadas por Endres (Endres et al., 1971) et.al. y Furui (Furui, 1981). Siendo este último quien propuso usar los coeficientes espectrales como características fundamentales, aunque no fueron consideradas relevantes en principio, años después a finales de los 80 aumentó su uso e incluso continúan siendo utilizadas en la actualidad en casi todos los sistemas de reconocimiento de voz (Donoso García del Castillo, 2014). Durante la época de los 80s, en los sistemas de reconocimiento de habla se comenzaron a utilizar técnicas de programación dinámica, pero fueron reemplazadas posteriormente por los modelos ocultos de Markov (HMM del inglés Hidden Markov Models), y a su vez, por su buen rendimiento en

estas tareas también fueron utilizados en los primeros sistemas de reconocimiento de locutor (Univaso, 2017). Sin embargo, no fue sino hasta principios de los 90s, que se introduce el concepto del modelo de Mezclas Gaussianas (GMM) el cual luego fue empleado en la mayoría de los sistemas de reconocimiento. Por otra parte, en esta misma época también se profundizó en los sistemas robustos a variables no deseadas como el **ruido**, las diferencias de canal y la variabilidad inter locutor. A finales de esta década, se inician evaluaciones de sistemas de reconocimiento de locutor (SRE) llevadas a cabo por el Instituto Nacional de Estándares y Tecnología (NIST) que buscan determinar el estado del arte, y continúan hasta la actualidad.

En los años 2000, se comienzan a introducir diversas técnicas de normalización en los modelos y también se inicia el uso de máquinas de soporte vectorial (SVM) como clasificadores (Wan and Campbell, 2000). Finalmente, en la década del 2010 es donde se han dado grandes avances y se han propuesto técnicas que son investigadas y utilizadas en la actualidad. A principios de esta época, en busca de remover o atenuar las características que no son propias del hablante, se inician los estudios sobre la representación del habla en subespacios vectoriales. Así, se plantean técnicas como el análisis factorial conjunto (Joint Factor Analysis, JFA) (Kenny, 2006) o técnicas basadas en el espacio de variabilidad total como los i-vectores (Dehak et al., 2011), que son consideradas el estado del arte según las evaluaciones de NIST (Univaso, 2017).

2.2 One-Shot Learning

En tareas tradicionales de clasificación, a partir de una observación (representada por un vector de características) un modelo obtiene a su salida la probabilidad de que esta observación pertenezca a una clase particular. Por ejemplo, si se tiene un modelo entrenado y se quiere determinar si una imagen particular es de un gato, un perro u otro animal, el modelo genera una salida con 3 valores (correspondientes a cada una de las 3 clases), siendo el mayor de ellos el correspondiente a la clase más probable. En este tipo de sistemas resulta complejo añadir una nueva clase cuando un modelo ya ha sido entrenado, pues se requiere grandes cantidades de observaciones de cada clase para aprender sus características particulares y poder diferenciarlas. En contraste, técnicas basadas en el concepto de one-shot learning (utilizar una única observación para diferenciar una clase) pueden mejorar la flexibilidad de las técnicas tradicionales de clasificación (Li Fe-Fei et al., 2003). En este tipo de técnicas, en lugar de evaluar directamente una observación para determinar a qué clase pertenece (como en un sistema de clasificación tradicional), se toma una observación adicional que sirve como referencia para obtener un valor de similitud entre esta y una nueva observación que se presente, siendo esta última de la que se quiere determinar a qué clase pertenece. Algo importante a notar es que, en este caso el modelo no aprende directamente a clasificar, sino que está aprendiendo a determinar qué tan similares son dos observaciones (1 similar, 0 no hay similitud) y a través de esto clasificar.

Existen diferentes arquitecturas que permiten realizar tareas de clasificación utilizando one-shot learning (Koch et al., n.d.; Li Fe-Fei et al., 2003; Santoro et al., 2016; Sung et al., 2018; Vinyals et al., 2016). Entre ellas, el método basado en redes siamesas (siamese neural networks) ha sido uno de los que ha obtenido muy buenos resultados en tareas de clasificación. En la figura 1 se muestra un diagrama general de una red siamesa, se puede observar que está formada por dos redes neuronales convolucionales (CNN) las cuales no son diferentes, sino que son dos copias de la misma red, básicamente comparten sus parámetros (son dos redes gemelas de ahí la terminología de siamesas). Las entradas son una observación de referencia de la clase denominada “Anchor” y una observación de entrada de la cual se quiere determinar a qué clase pertenece. Ambas pasan a través de las capas convolucionales y suponiendo que el modelo fue entrenado correctamente se tendrán dos casos posibles: si las dos entradas pertenecen a la misma clase, su vector de características (encoding) debe ser similar, mientras que, si son de una clase diferente, su vector de características será diferente. De igual forma, si se obtiene la diferencia absoluta (Absolute Difference) entre los dos vectores de características, este resultado será diferente dependiendo del caso que se presente. Finalmente, el resultado de la diferencia de los vectores, ingresa a una última capa que contiene

únicamente una función sigmoide que da a su salida un valor de 1 si los vectores son similares y un valor de 0 si son diferentes.

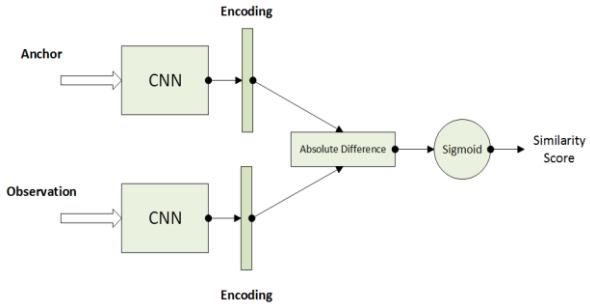

Fig. 1. Arquitectura general de una red siamesa

Se puede notar que tal como en el concepto de one-shot learning, utilizando redes siamesas lo que se busca es aprender a clasificar por medio de una función de similitud, es decir lo que el modelo aprende es la **tarea** de distinguir la similitud más no a clasificar entre clases directamente.

3 Metodología Experimental

3.1 Sistema Propuesto

En este trabajo se propone el uso de un sistema compuesto por un Front-End que se encarga del procesamiento de la señal de audio y un Back-End donde se realiza la tarea de la identificación del locutor como tal. Por otra parte, para evaluar el sistema se seleccionó la base de datos de Voxceleb1 y la división realizada en la misma; siendo 1211 locutores para entrenamiento y 40 locutores para testeo. Además, debido a que en la base de datos existían audios de diferentes duraciones se utilizaron únicamente 20 audios de 5 segundos de duración por cada locutor (los audios de menor duración se descartaron y los de mayor duración se recortaron a 5 segundos), con la finalidad de valorar la capacidad del sistema para aprender utilizando una menor cantidad de datos que sistemas tradicionales.

Partiendo de la base de datos, en el Front-End, se utilizan tres enfoques diferentes para disminuir el tamaño del vector de entrada y a su vez adecuar la señal de audio acorde a la arquitectura del modelo implementado en el Back-End. En el primer enfoque, el audio fue re-muestreado (downsampling) a una frecuencia mucho menor; de 16000 Hz se pasó a 1000 Hz y luego este vector unidimensional se transformó en un vector bidimensional de 70x70, el cual es la entrada del Back-End. Cabe destacar que, únicamente en este enfoque se elimina un segundo de la muestra para regularizar el tamaño del vector (al re-muestrear la señal se redondea hacia abajo quedando un vector bidimensional de 70x70) y esta metodología se acopla al objetivo del primer enfoque pues se busca evaluar la capacidad del sistema para identificar un locutor cuando existe pérdida de información. En el segundo enfoque, se utilizaron los i-vectores de las señales de audio. Para ello, en primer lugar, se obtuvieron los coeficientes cepstrales en escala de mel (MFCC) de los audios utilizando un tamaño de ventana de 25 milisegundos (ms), espaciado cada 10 ms y 26 filtros triangulares. Luego, utilizando estos MFCC se entrenó un UBM (Universal Background Model) el cual fue utilizado para obtener 400, 900 y 1600 i-vectores. De igual manera que en el primer enfoque, el vector que contiene los i-vectores se concatena y se transforma en un vector bidimensional (de tamaño 20x20, 30x30 y 40x40, respectivamente) que es la entrada del Back-End. Cabe señalar que todo el proceso realizado en este enfoque se lo hizo utilizando la librería de Sidekit ("Welcome to SIDEKIT 1.3.1 documentation! — SIDEKIT documentation," n.d.). Finalmente, en el tercer enfoque de los audio se extrae el espectrograma de mel en escala logarítmica dado que se ha demostrado que el mismo obtiene una muy buena representación de una señal de audio (Choi et al., 2018). El espectrograma se obtuvo utilizando 67 filtros de mel generando así un vector bidimensional de dimensiones 67x67, el cual, es la entrada del Back-End. En la figura 2 se puede observar el diagrama general de funcionamiento del sistema propuesto.

Por otra parte, para la identificación del locutor en el Back-End se utiliza una red neuronal siamesa que está basada en la arquitectura y metodología propuesta por Kotch et al. (Koch et al., n.d.). El modelo usado está compuesto por dos redes neuronales convolucionales (ConvNet) las cuales no son diferentes, sino que comparten la misma estructura y valores de parámetros (luego de cada ConvNet se realizó normalización del lote o batch). De esta forma, a la entrada de estas, se

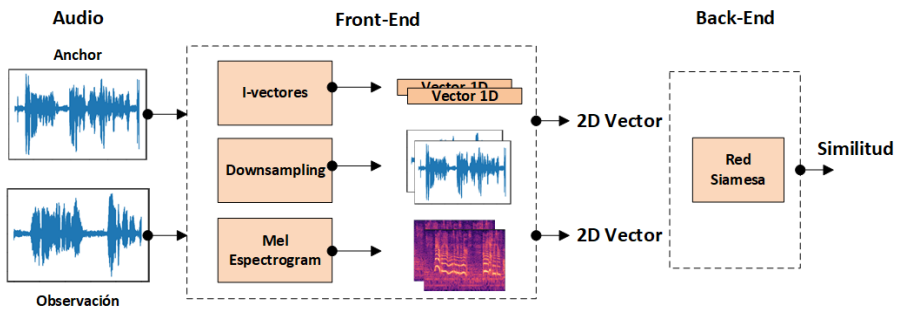

Fig. 2. Diagrama general de funcionamiento del Sistema Front-End Back-End

tendrán dos vectores (uno de referencia o anchor y una observación) los cuales pasan a través de la ConvNet y dan como resultado un vector de características de longitud fija por cada entrada. Ahora, si comparamos estos dos vectores pueden existir dos casos; si los dos audios pertenecen al mismo locutor sus vectores de características deben ser similares y si no son del mismo locutor deberían ser diferentes. Por tanto, de la salida de la red siamesa se calcula la diferencia absoluta (L1 distance) de los dos vectores de características que se obtiene de la ConvNet y este vector resultante es la entrada de una última capa densa de neuronas (fully-conected) que tiene una función de activación de tipo sigmoide, quien a su salida entregara un valor (1) en el caso de que las entradas sean del mismo locutor y otro valor (0) en el caso de que sean de un locutor diferente. De esta forma, durante el entrenamiento lo que se busca es que el modelo adapte sus parámetros de tal forma que pueda ser capaz de mejorar su exactitud al identificar si un par de observaciones pertenecen o no al mismo locutor. En la figura 3 se muestra la arquitectura de la red siamesa utilizada.

3.2 Métrica de Evaluación

En la tarea de identificación de locutor la idea principal es determinar a qué locutor pertenece la voz de entre varios locutores, por lo que este tipo de problemas podría ser catalogado como una clasificación multiclase. Sin embargo, en one-shot learning y con la red neuronal siamesa lo que se busca no es que el modelo aprenda a diferenciar entre clases, sino que aprenda a extraer características y que determine una medida de similitud entre dos observaciones. Así, la clasificación multiclase se transforma en una tarea de clasificación binaria donde una clase representa una similitud total y otra clase que no existe similitud. En este contexto, para evaluar el rendimiento de cada enfoque se tomó una estrategia denominada N-way One-shot clasification [9], donde N representa la cantidad de pares de observaciones a evaluar.

Por ejemplo, en una estrategia 4-way one-shot clasification se tendría una observación base (audio a identificar a cuál locutor pertenece) y se compararía con respecto a otras 4 observaciones donde únicamente una es del mismo locutor, por lo que se espera que el modelo obtenga una métrica de similitud alta a esta observación y más baja para las otras observaciones que son de otros locutores. Así, si entre las dos observaciones que son del mismo locutor se obtiene la similitud máxima la clasificación es correcta, caso contrario, se la consideraría incorrecta. En la figura 4 se ilustra una clasificación correcta en un ejemplo de 4-way one-shot clasification. Si estas pruebas son realizadas m cantidad de veces, se puede obtener un resultado más general del rendimiento del enfoque, por lo que, la métrica final de evaluación de los enfoques está en función de:

$$S = \frac{mc}{m}$$

Donde mc representa cuantas clasificaciones se realizaron correctamente durante m cantidad de veces y S representa su exactitud.

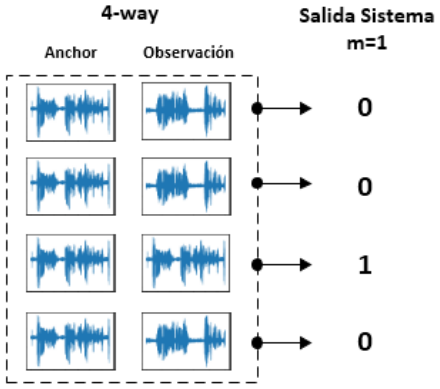

Fig. 4. Ilustración de una evaluación basada en N-way Clasification para N=4

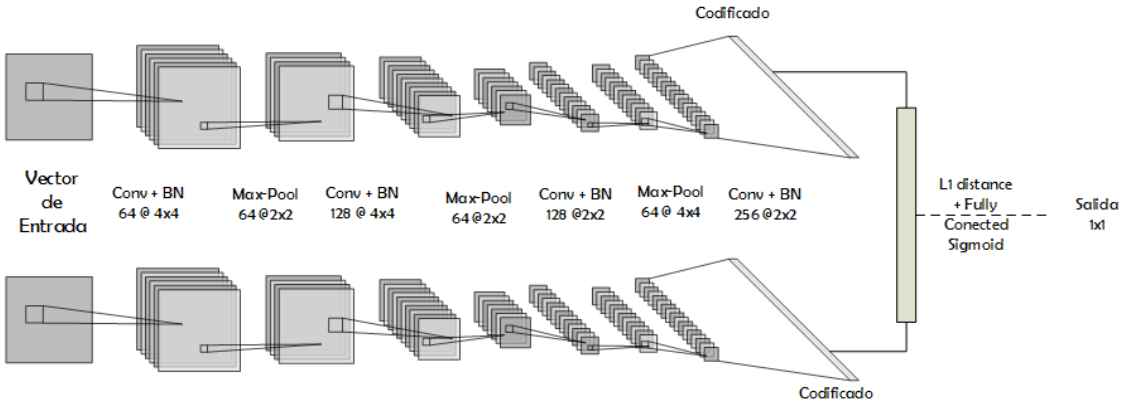

Fig. 3. Arquitectura de la red neuronal siamesa implementada en el Back-End

Si todo se clasificó correctamente, se obtendría un valor de S=1, cada vez que se cometa un error este valor ira disminuyendo hasta alcanzar 0 que sería el caso en que ninguna clasificación fue realizada correctamente.

3.3 Fase de Entrenamiento

En el entrenamiento del sistema se utilizaron 80 épocas (aproximadamente 20000 iteraciones), un tamaño de batch de 100, la función de perdida de entropía cruzada binaria (binary cross-entropy) y el algoritmo de Adam con una tasa de aprendizaje de 0.00001 (para tasas de mayor valor se volvía inestable). En las pruebas realizadas, el enfoque basado en i-vectores fue el que peor rendimiento obtuvo, como se puede observar en la figura 5 incluso para diferenciar entre 4 locutores (N=4) no se alcanza un buen resultado y al aumentar el número de locutores a diferenciar su exactitud fue disminuyendo aún más. Este comportamiento posiblemente se deba a que, la información que está presente en los i-vectores se vuelve dispersa con respecto al audio original y causa que en las capas convolucionales no sea posible extraer información relevante para medir la similitud.

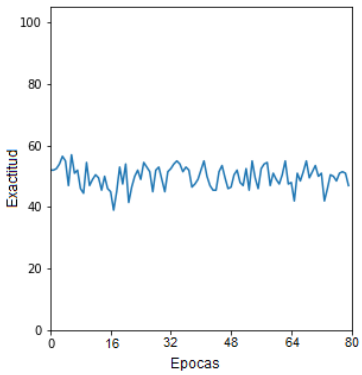

Fig. 5. Tendencia de la exactitud para tamaño del batch de 100 y N=4 en enfoque i-vectores

Por otra parte, en los otros enfoques ocurre un comportamiento opuesto al mostrado con i-vectores, en ambos casos se puede notar que si existe una tendencia de ascendente en la exactitud. En la figura 6, se muestra la tendencia de la exactitud en enfoque de downsampling durante el entrenamiento, para N=4. Mientras que en la figura 7, se muestra la tendencia del enfoque del espectrograma de mel, pero para N=7.

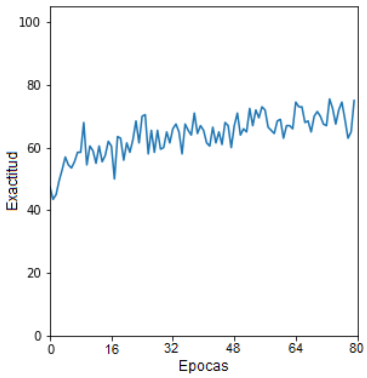

Figura 6. Tendencia de la exactitud para tamaño del batch de 100 y N=4 en enfoque Downsampling

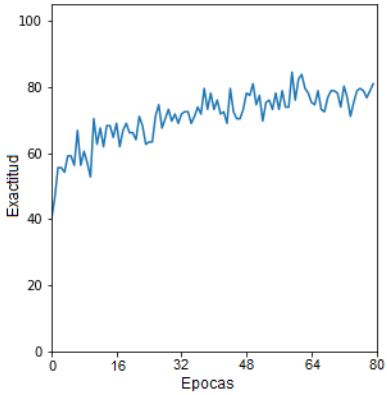

Figura. 7. Tendencia de la exactitud para tamaño del batch de 100 y N=7 en enfoque Espectro de Mel

Al comparar ambas figuras se puede notar la diferencia que existe en los dos métodos pues, aunque en el caso del espectrograma de mel se está utilizando una mayor cantidad de locutores (N mayor), supera al rendimiento de downsampling que utiliza menos locutores, se alcanzó una exactitud de 0.84 y 0.75, respectivamente.

4 Resultados

Los experimentos realizados tienen la finalidad de identificar el potencial que tienen estos tipos de sistemas para tareas de identificación del locutor, por lo que la métrica sobre la exactitud, se obtiene con respecto a un grupo de locutores diferentes a los locutores utilizados en el entrenamiento (grupo de testeo) y utilizando m=100. Es decir, la métrica mostrada es el promedio de 100 pruebas de testeo realizadas para esa observación. Las diferentes pruebas se realizaron modificando el valor de N, siendo el menor valor N=4 y el valor más alto utilizado N=40. En las pruebas realizadas para N=40, es decir para diferenciar a un locutor particular de entre otros 39 locutores, la exactitud disminuyó considerablemente con respecto a lo alcanzado para un menor valor de N. Por lo que, aunque se fijó el valor de 80 épocas para todos los experimentos, dado el bajo rendimiento y que el número de N es el mayor de todas las pruebas, se realizó una prueba adicional con el doble de épocas, es decir para 160 épocas en total.

Por otra parte, de los tres enfoques, el basado en i-vectores fue para el que peor rendimiento obtuvo en todas las pruebas, por lo que, no se muestran los resultados con respecto a i-vectores por su bajo rendimiento (para N=4 se alcanzó una exactitud de 54% y para valores mayores de N su exactitud fue menor) en comparación a los otros enfoques. En la tabla 1 se muestran los resultados de los enfoques basados en el espectrograma de mel y downsampling, se puede observar que para un valor de N=4 el rendimiento alcanza un S=0.9 para el enfoque del espectrograma de mel mientras que para downsampling se alcanza un S=0.75. Sin embargo, conforme se aumenta la cantidad de locutores a evaluar (N>4), la exactitud disminuye para ambos enfoques. En las pruebas adicionales realizadas para N=40 con el doble de épocas (160) se obtuvo una mejoría para ambos enfoques alcanzando un 66% de exactitud con el espectrograma de mel y 45% para downsampling, es decir una mejoría del 10% con respecto a las pruebas con 80 épocas.

Enfoque (Front-End)	N (Locutores)	Exactitud (m=100)
Downsampling	4	75 %
	7	61 %
	10	54 %
	20	47 %
	40	35 %
Espectrograma de Mel	4	90 %
	7	84 %
	10	76 %
	20	67 %
	40	56 %

Tabla 1: Resultados Obtenidos en la Fase de Experimentos para 160 Épocas

5 Conclusiones

En este trabajo se propone el uso de un sistema basado en one-shot learning para la identificación del locutor y se evalúa su funcionamiento utilizando audios que contienen tanto la voz de un locutor como ruido ambiental, para valorar su rendimiento en un ambiente real. En los experimentos realizados se pudo notar que, a medida que se aumentaba la cantidad de locutores a diferenciar, la exactitud disminuye hasta alcanzar un 66% al diferenciar de entre 40 locutores a la vez, para una cantidad mayor de locutores, considerando el comportamiento del sistema durante los experimentos el rendimiento podría ser aún menor. Se piensa que este comportamiento podría deberse a que, al aumentar la cantidad de locutores, es probable que varios de ellos tengan características similares en su voz causando que la información se solape y se complique la tarea de diferenciarlos. Por lo que en trabajos futuros se tomara en consideración el método usado en el Front-End, para evitar perder información que podría ser determinante a la hora de diferenciar dos audios con características muy similares en la voz. Sin embargo, cabe señalar que, incluso al utilizar Downsampling de una forma tan agresiva el sistema obtuvo un resultado aceptable, de esta manera se pudo constatar la capacidad que se tiene para identificar un locutor, aun cuando existe gran pérdida de información con respecto al audio original.

De entre todas las arquitecturas propuestas, la que mejor rendimiento obtuvo fue la que utiliza el espectrograma de mel en escala de la señal de audio original para determinar a quién pertenece la voz a través de una medida de similitud entre dos observaciones. Cabe destacar que, en este tipo de sistemas para la identificación se utiliza únicamente una muestra de audio y además existe la posibilidad de utilizar la voz de locutores diferentes a los presentados en el entrenamiento (algo que no es posible en un sistema de clasificación tradicional). Estas características, le brindan al sistema gran escalabilidad para ser utilizado en situaciones reales donde puede ser complicado adquirir gran cantidad de muestras de un locutor. Por lo que, según los resultados obtenidos y las ventajas que presentan estos sistemas durante el entrenamiento y el funcionamiento, los sistemas basados en one-shot learning tienen gran potencial para ser utilizados en tareas de identificación del locutor independiente del texto.

Bibliografía

Endres, W., W. Bambach, G. Flösser, 1971. Voice spectrograms as a function of age, voice disguise, and voice imitation. *Journal Acoustic Society of America*. Volumen:49, páginas 1842–1848.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on Acoustics, Speech and Signal Processing.*Volumen:29, páginas 254–272.

Donoso, R., 2014. Diseño e implementación de un sistema de reconocimiento de hablantes. *Universidad Carlos III de Madrid.*

Univaso, P., 2017. Forensic speaker identification: a tutorial. *IEEE Latin America Transaction.* Volumen:15, páginas 1754–1770.

Wan, V., W. Campbell, 2000. Support vector machines for speaker verification and identification. *Neural Networks for Signal Processing X.* Volume:2, páginas 775–784.

Kenny, P., 2006. Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. *Computer Science.*

Dehak, N., P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, 2011. Front-End Factor Analysis for Speaker Verification. *IEEE Trans. Audio Speech Lang. Proces.* Voumen:19, páginas788–798.

Fe-Fei, L., R. Fergus, P. Perona, 2003. A Bayesian approach to unsupervised one-shot learning of object categories. *Proceedings Ninth IEEE International Conference on Computer Vision.* Volumen:2,páginas 1134–1141, France.

Koch, G., R. Zemel, R. Salakhutdinov, 2015. Siamese neural networks for one-shot image recognition. *Proc. 32 International Conference on Machine Learning.* Volumen:37, France.

Santoro, A., S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, 2016. Meta-Learning with Memory-Augmented Neural Networks. *Proc. 33 International Conf. on Machine Learning*, USA.

Sung, F., Y. Yang, L. Zhang, T. Xiang, P. Torr, T. Hospedales, 2018. Learning to Compare: Relation Network for Few-Shot Learning. *IEEE Conference on Computer Vision and Pattern Recognition.* Páginas 1199–1208, USA

Vinyals, O., C. Blundell, T. Lillicrap, K. Kavukcuoglu, D. Wierstra, 2016. Matching Networks for One Shot Learning. *30th Conference on Neural Information Processing Systems.* páginas 3630–3638, España.

SIDEKIT 1.3.1 documentation, URL https://projets-lium.univ-lemans.fr/sidekit/ (accessed 9.5.19).

Choi, K., G. Fazekas, M. Sandler, K. Cho, 2018. A Comparison of Audio Signal Preprocessing Methods for Deep Neural Networks on Music Tagging. *26th Eur. Signal Process Conference,* páginas 1870–1874, Italy.