Tarea 8 - Reconocimiento de voz

Claudia Lissette Gutiérrez Díaz
Licenciada en Ciencias Computacionales, FCFM, UANL

1. Introducción

El reconocimiento de voz ha tomado mucha importancia en los últimos años, ya que las nuevas tecnologías, como la inteligencia artificial, el internet de las cosas, la ciberseguridad, entre otros, reciben sonidos, los interpretan y dan una respuesta a ello. Es por eso que es importante el poder reconocer el usuario de donde proviene la voz.

En este reporte haremos la comparación de dos metodologías que nos pueden permitir el reconocer al dueño de la voz de múltiples audios, esto con el uso de redes neuronales.

2. Descripción de los datos

El conjunto de datos que utilizaremos para este reporte es el encontrado en el artículo *Speaker Recognition* (Evans, 2021), el cual contiene discursos de cinco líderes destacados, Benjamín Netanyahu, Jens Soltenberg, Julia Gillard, Margaret Tacher y Nelson Mandela.

Originalmente, el discurso de cada orador era un audio extenso, se dividieron en un segundo cada uno para facilitar el funcionamiento del entrenamiento y prueba.

3. Metodología y resultados

Utilizando 6,375 fuentes de audio de un segundo cada uno, realizaremos un experimento donde encontraremos el modelo de clasificación más eficiente para poder identificar al líder dueño del audio.

En el artículo original utiliza como variables dependientes la matrices generadas por el MFCC (*Mel Frequency Cepstral Coeffcients*), divididas en 13 características de 32 componentes y se usa la red neuronal recurrente LSTM (por sus siglas en inglés, *Long short-term memory*) para encontrar la clasificación.

Por nuestra cuenta, usaremos las ondículas para generar matrices de 11 niveles de diferente tamaño aplicadas al audio que generan hasta 8000 registros, usaremos la red neuronal recurrente LSTM, y además, vamos a realizar una tercera comparación al realizar el entrenamiento del modelo usando redes convolucionales, al simular que la matriz dependiente es una imagen, encontrando así los patrones necesarios para reconocer en la prueba.

Para la prueba 1, usando MFCC con LSTM, usaremos 3 capas, donde la capa de entrada es de 128 neuronas, la segunda de 64, usando función de activación `softmax` y la tercera de salida de 5 neuronas con función de activación `softmax`.

Para la prueba 2, usando ondículas con LSTM, usaremos 3 capas, donde la capa de entrada es de 128 neuronas, la segunda de 64, usando función de activación `softmax` y la tercera de salida de 5 neuronas con función de activación `softmax`.

Para la prueba 3, usando ondículas con CNN, usaremos una primera capa convolucional de 32 neuronas de 3x3 con función de activación `ReLU`, una primera capa de `MaxPooling` de 2x2, una segunda capa convolucional de 64 neuronas de 3x3 con función de activación `ReLU`, una segunda capa de `MaxPooling` de 2x2, una tercera capa convolucional de 128 neuronas de 3x3 con función de activación `ReLU`, una tercera capa de `MaxPooling` de 2x2, una capa de aplanamiento, luego una capa completamente conectada de 128 neuronas con función de activación `ReLU` y por último una capa de salida de 5 neuronas con función de activación `softmax`.

Para las 3 pruebas, se usa el optimizador `Adam`, la función de pérdida `categorical_crossentropy` y como métrica `accuracy`.

Prueba	Exactitud	F1-Score
MFCC LSTM	0.96	0.96
Ondícula LSTM	0.32	0.30
Ondícula CNN	0.92	0.92

Cuadro 1. Comparación de pruebas

En la tabla 1, se puede observar la comparativa entre las 3 pruebas con respecto a la métrica de Exactitud (*Accuracy*) y el F1-Score. Podemos observar que la técnica utilizada en el artículo original es la más precisa de las 3 pruebas realizadas, donde los valores dependientes es el coeficiente de MEL. En la figura 1 podemos observar la matriz de confusión utilizando la técnica anadora, donde vemos que en la gran mayoría del conjunto de datos de prueba cayó en la categoría correcta, debido al gran porcentaje de exactidud del modelo.

Sin embargo, no desechamos el modelo diseñado por nuestra auditoría utilizando redes convolucionales, al arrojar un porcentaje de exactidud del 92 %, el cual, como observamos en la figura 2, al igual que en la anterior prueba, la gran mayoría de los elementos de prueba cayeron en su categoría correcta.

Por último, descartaríamos por completo la prueba 2 usando las ondículas con LSTM, ya que la exactitud fue muy baja al tratarse de un 32 %, el cual notamos en la figura 3 al verse las clasificaciones entre las predicciones y la realidad más disperso.

Considerando los tiempos de ejecución, la prueba del artículo original utilizando MFCC con LSTM, consumió menos recursos computacionales y tardó mucho menos tiempo que las otras dos pruebas, por lo que si se necesita realizar un nuevo modelo para reconocer voces, preferiría usar estos datos como datos dependientes.

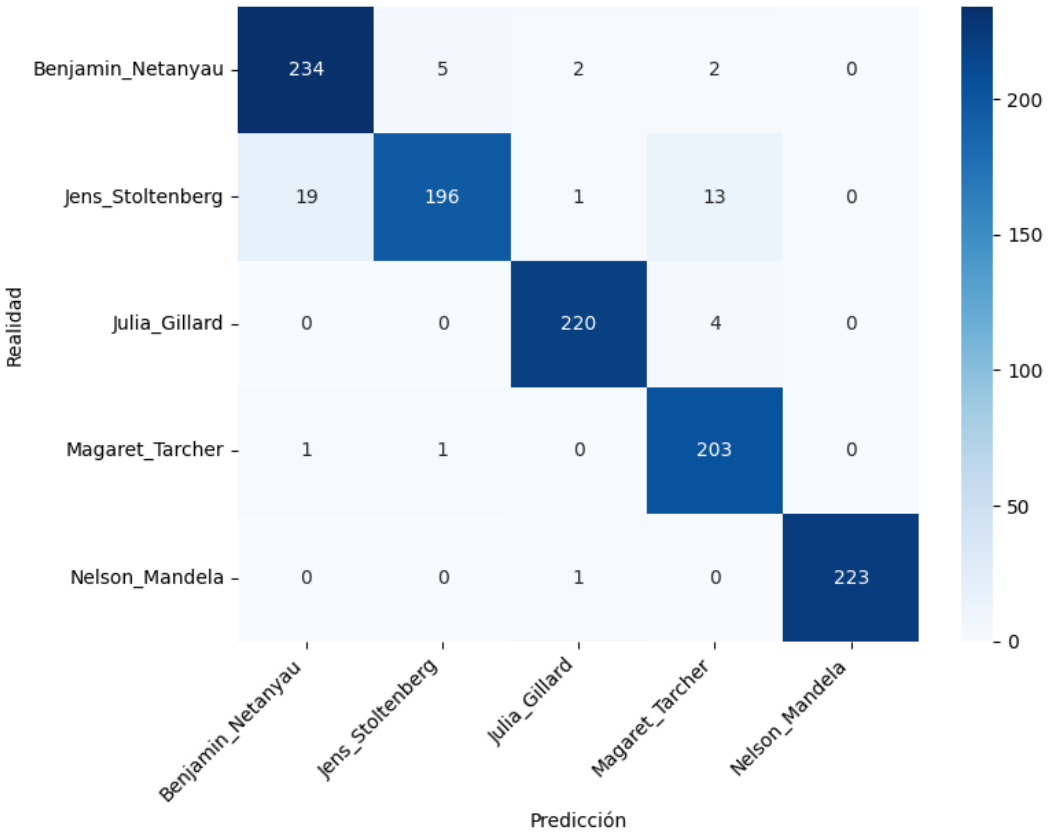

Figura 1. Matriz de confusión de MFCC por LSTM

Figura 2. Matriz de confusión de coeficientes de ondículas por redes convolucionales

Figura 3. Matriz de confusión de coeficientes de ondículas por LSTM

Referencias

Evans, K. (2021). Speaker Recognition Dataset [Accessed: 2024-07-07]. https://www.kaggle.com/datasets/kongaevans/speaker-recognition-dataset