Aplicación de una Red Neuronal Convolucional para el Reconocimiento de Personas a Través de la Voz

Alvaro Araujo*‡, Jesús Pérez†‡ y Wladimir Rodriguez‡
* *Centro de Estudios en Microelectrónica y Sistemas Distribuidos*
† *Laboratorio de Sistemas Discretos, Automatización e Integración*
‡ *Departamento de Sistemas Computacionales*
Universidad de Los Andes
Mérida, Venezuela
{aalvaro, jesuspangulo, wladimir}@ula.ve

***Resumen*—El reconocimiento de personas a través de la voz se ha popularizado recientemente en aplicaciones para teléfonos móviles, seguridad, robótica social, entre otras. Los métodos convencionales de reconocimiento a partir de la voz utilizan coeficientes para generar vectores de características presentes en el audio que alimentan modelos de aprendizaje automático. Aunque se obtienen excelentes resultados con estos métodos convencionales, se requiere una alta cantidad de muestras de voz por persona para entrenar los modelos de aprendizaje automático, lo cual es una tarea tediosa para los usuarios. En ese sentido, el objetivo de este trabajo consiste en utilizar un método no convencional en aras de disminuir la cantidad de muestras requeridas para el entrenamiento. El método propuesto consiste en generar imágenes con los espectrogramas de la voz, para entrenar una red neuronal convolucional que las clasifique. En este trabajo se comparan los resultados de entrenamiento y validación para distintas cantidades de muestras con el propósito de identificar la cantidad adecuada. Luego, con la cantidad identificada se comparan los resultados de entrenamiento y validación para distintos valores de un parámetro utilizado en la generación de los espectrogramas denominado ganancia, con el objetivo de optimizar los resultados. Los mejores resultados obtenidos indican que deben utilizarse 5 muestras con una ganancia de *0.9*, para obtener una exactitud de validación del *93,34%*.**

Palabras Clave*—*Espectrograma, Red Neuronal Convolucional, Reconocimiento de Personas, Red Neuronal Profunda.

I. Introducción

El reconocimiento de personas a través de la voz consiste en determinar la identidad de estas basándose en expresiones ya conocidas de un grupo de individuos [1]. Recientemente, este tipo de reconocimiento se ha popularizado en aplicaciones de teléfonos móviles [2], seguridad [3], robótica social [4], entre otras. En ese sentido, en los últimos años las investigaciones se han interesado en mejorar los métodos utilizados para realizar esta tarea [1][5][6][7][8]. La tendencia de las investigaciones está orientada hacia la utilización de modelos de aprendizaje automático, por lo tanto, para lograr el reconocimiento se deben tener muestras de las personas para entrenar estos modelos. Los métodos convencionales de reconocimiento a partir de la voz extraen coeficientes de las muestras de audio para generar vectores de características. Luego, estos vectores se utilizan para entrenar modelos de aprendizaje automático, tales como: Redes Neuronales Profundas [1][5][6], Redes Neuronales Convolucionales [6], Redes Neuronales Recurrentes de Memoria a Corto Plazo [1][5][7], entre otras. Aunque en trabajos recientes utilizan frases cortas como "Ok Google" [1], "Hey Siri" [5] o "Wei" [6], se aprecia una gran cantidad de muestras por frases pertenecientes a una misma persona [1][5][9][10]. La recolección de esa gran cantidad de muestras representa un desafío para los investigadores porque puede generar fatiga en los usuarios si este proceso es manual. En ese sentido, en esta investigación se propone la disminución de esa cantidad de muestras utilizando un método que no ha sido encontrado en otras publicaciones para realizar esta tarea. El método propuesto consiste en generar imágenes (espectrogramas) a partir de las señales que genera la voz, con el propósito de entrenar una Red Neuronal Convolucional que se encargue de clasificar las imágenes. Un espectrograma es una Representación de la Frecuencia en el Tiempo (TFR, por sus siglas en inglés) que permite una descripción precisa de las señales no estacionarias, la cual es calculada mediante la concatenación de espectros obtenidos a través de Transformadas consecutivas de Fourier de Tiempo Reducido (STFT, por sus siglas en inglés) [11]. Por otro lado, una Red Neuronal Convolucional (CNN, por sus siglas en inglés) es un tipo de red neuronal artificial profunda que obtiene este nombre de la operación matemática lineal entre matrices llamada convolución [12].

El documento se organiza de la siguiente manera: en la Sección II, se exponen trabajos previos relacionados con el problema que se intenta abordar; en la Sección III, se explica la construcción del conjunto de datos y se describe el modelo de aprendizaje automático utilizado; en la Sección IV, se exponen los resultados obtenidos; en la Sección V, se discuten los resultados previamente presentados; y finalmente, en la Sección VI se presentan las conclusiones generadas.

II. Antecedentes

Los antecedentes de esta investigación tienen tres características: primero, proponen métodos de reconocimiento de personas a través de la voz; segundo, utilizan modelos de aprendizaje automático; y tercero, han sido publicados en los últimos tres años.

En [1], presentan un enfoque integrado que mapea el enunciado a evaluar (la frase "Ok Google") con algunas expresiones de referencia. El método utiliza i-vector y d-vector para representar las muestras (aproximadamente 500 muestras por persona). El enfoque fue probado en dos modelos: primero, utilizando una Red Neuronal Profunda (DNN, por sus siglas en inglés), y segundo, utilizando una Red Neuronal Recurrente de Memoria a Corto Plazo (LSTM, por sus siglas en inglés). Los resultados obtenidos fueron satisfactorios para ambos casos debido a que lograron una tasa de error de 2.04% en el caso de la DNN y de 1.36% en el caso de la LSTM.

En [5] se propone una transformación discriminativa generalizada a través del Aprendizaje Curricular (CL, por sus siglas en inglés) donde se utiliza el principio general de aprender conceptos más simples antes de aprender gradualmente los más complejos. El conjunto de datos de entrenamiento está compuesto por más de 20 muestras por persona de la frase "Hey Siri". Para esta tarea se utilizó una DNN con el propósito de extraer información específica de la voz basada en la frase de entrada. El CL propuesto consta de tres pasos: primero, se aprende acerca de la frase establecida; luego, se aprende contenido del texto; y por último, se aprende acerca de las condiciones acústicas. Los resultados obtenidos indican que el CL es mejor que el Entrenamiento Multi-Estilo (MST, por sus siglas en inglés) en todas las condiciones acústicas, sin embargo, los resultados obtenidos en ambos casos son aceptables con una tasa de error que no supera el 4%.

En [6], proponen un enfoque basado en la combinación de un componente Convolucional, un componente de Retardo de Tiempo y una Red Neuronal Profunda (CT-DNN, por sus siglas en inglés). Se hacen comparaciones entre i-vector y d-vector, para un conjunto de datos que está compuesto por aproximadamente 24 muestras por persona, de tres eventos triviales: tos, risa y "Wei" (saludo corto en China); cuya duración varía entre 0.2 y 1.0 segundos. Los resultados obtenidos oscilan entre 10% y 14% de error para los tres eventos y los mejores resultados están asociados al d-vector con el saludo "Wei".

En [7] presentan una interpretación del Modelo de Fondo Universal (UBM, por sus siglas en inglés) y la consideran como una función de mapeo que transforma las emisiones verbales en un vector de características. Las muestras por persona tienen una duración de 70 segundos. Con esta interpretación, probaron tres vectores de características (i-vector, Coeficientes Cepstrales en las Frecuencias de Mel (MFCC, por sus siglas en inglés) y la combinación de ambos) y dos modelos de redes neuronales (LSTM y Red Profunda de Creencias (DBN, por sus siglas en inglés)). Los mejores resultados obtenidos están asociados a la combinación de i-vector y MFCC para la construcción del vector de características y el modelo LSTM, con una exactitud del 99.5%.

En [9], se presenta un enfoque múltiple de i-vector que utiliza varios extractores que forman múltiples matrices de diferentes dimensiones para mejorar el rendimiento, combinando la información contenida en cada i-vector. El conjunto de datos utilizado está formado por alrededor de 80 muestras por persona con una duración de aproximadamente 20 segundos. Los resultados obtenidos muestran que el rendimiento de una Red Neuronal Recurrente (RNN, por sus siglas en inglés) con el enfoque múltiple de i-vector logra mejores resultados en comparación con los enfoques de i-vector de generación única y el Modelo de Mezclas Gaussianas (GMM, por sus siglas en inglés), para errores del 7.31% y 11.08% respectivamente.

De los antecedentes presentados anteriormente, se observa que todos utilizan vectores de características para alimentar modelos de aprendizaje automático, y además, se nota un interés por mejorar los resultados mediante combinaciones de las características que forman los vectores y distintos modelos de aprendizaje automático. Particularmente, se observa en los tres primeros antecedentes una tendencia a utilizar frases cortas, sin embargo, utilizan una cantidad relativamente elevada de muestras por persona. En ese sentido, en aras de probar nuevas combinaciones, en este trabajo se alimentará con imágenes (espectrogramas) un modelo de aprendizaje automático, con la intención de estudiar la disminución de la cantidad de muestras por persona.

III. Método

Con el propósito de realizar un conjunto de pruebas que permitan generar conclusiones respecto al número mínimo de muestras por persona y la ganancia óptima utilizada para generar los espectrogramas pertenecientes a estas muestras, se realizaron las siguientes tareas: construcción del conjunto de datos, procesamiento de los datos, y entrenamiento-validación del modelo.

A. Construcción del Conjunto de Datos

Con la finalidad de construir el conjunto de datos que se utilizará para el entrenamiento y validación del sistema, se grabaron audios de 30 personas (10 Mujeres y 20 Hombres) con una edad comprendida entre 18 y 60 años. Por cada persona se tomó una muestra de 20 grabaciones, las cuales tienen una duración de 2 segundos cada una; estas grabaciones contienen la frase "Hola Robot", se encuentran en formato wav (wave), tienen una frecuencia de muestreo de 44100 Hercios y utilizan dos canales (sonido estéreo). En la Figura 1 se observa un gráfico de nivel perteneciente a una de las muestras, donde se distingue el segmento donde es pronunciada la frase.

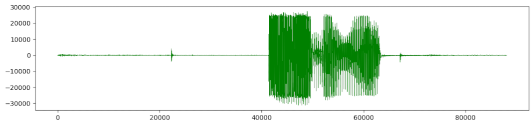

Figura 1. Gráfico de Nivel Perteneciente a una de las Muestras

B. Procesamiento de los Datos

El procesamiento propuesto tiene como objetivo, generar imágenes (espectrogramas) y preparar los vectores de

información que serán utilizados por el modelo de aprendizaje automático. Los cinco pasos que integran el procesamiento se describen a continuación:

- Truncamiento de Colas: en este proceso, se limita el audio al dominio del tiempo donde se realiza la pronunciación de la frase "Hola Robot". El criterio de truncamiento actúa sobre los segmentos de los extremos (colas) de la grabación y si el valor máximo en un segmento es menor al 16% del valor máximo registrado en la grabación, el mismo es suprimido de esta última.
- Normalización de Nivel: con la finalidad de regularizar el nivel de sonido en cada grabación, se decidió ajustar el valor máximo registrado en la misma al 40% del valor máximo soportado por el tipo de archivo de audio utilizado (wav), y luego, se ajustaron a esta proporción generada los demás fragmentos de la grabación (ver Figura 2).

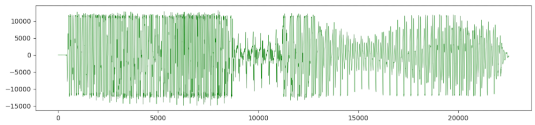

Figura 2. Una Muestra luego de ser Truncada y Normalizada

- Generación de Espectrogramas: a partir de cada grabación se genera una imagen que contiene un espectrograma. Esta tarea se lleva a cabo utilizando FFmpeg [13], la cual es una plataforma que permite manipular archivos multimedia. En este proceso de generación se utilizan los siguientes parámetros: como función de ventaneo se utiliza la función *Hann*, la escala utilizada para calcular la intensidad de color fue *logarithmic* (logarítmica), la manera como se desliza el espectro a lo largo de la ventana se establece como *replace* (las muestras comienzan de nuevo a la izquierda cuando llegan a la derecha), y por último, el valor utilizado para el parámetro *gain* (ganancia) se escogió como caso de estudio, por lo tanto, fue utilizado un conjunto de valores para el mismo. Este último parámetro se refiere a una ganancia en la escala con la cual se calcula la intensidad del color, que por defecto tiene un valor predeterminado de 1 (sin ganancia) y su escala está comprendida desde 0 a 128 incluyendo valores decimales. Se decidió estudiar el efecto que ejerce una variación en la ganancia, ya que gracias a la naturaleza de este parámetro, es posible intensificar diferentes detalles en las imágenes generadas a medida que este varía, lo cual probablemente permita la detección de patrones que normalmente serían difíciles de detectar con el valor de la ganancia predeterminado. Esto último posiblemente tenga incidencia en el rendimiento del modelo. En la Figura 3 se observa un conjunto de espectrogramas generados con diferentes valores de ganancia, donde se logra observar un cambio en los espectrogramas.
- Recorte y Redimensionado de las Imágenes: las imágenes generadas se limitan a un rango de frecuencias correspondientes al conjunto [0,5000] Hercios y se redimensionan a un tamaño de 400x133 píxeles. En la Figura 4 se observan las imágenes de un subconjunto de la información procesada.

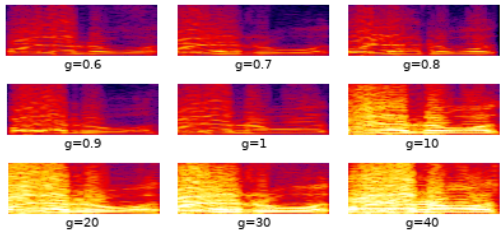

Figura 3. Espectrogramas Generados con Diferentes Valores de Ganancia

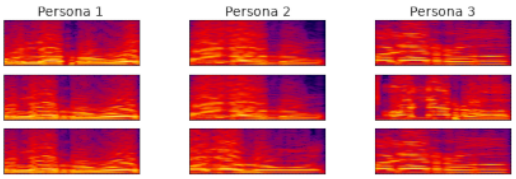

Figura 4. Espectrogramas Pertenecientes a Tres Personas

- Preparación de los datos: en este último paso del procesamiento, se verifica que los datos tratados estén completos y se procede a construir el vector de atributos y el vector de destino. El primero contendrá los valores de cada píxel perteneciente a cada imagen y el segundo contendrá la información correspondiente a cada clase. Dado que los datos utilizados para generar el vector de destino son categóricos (1 categoría por cada persona) y el modelo utilizado no trabaja con datos categóricos, se genera una matriz de clases binarias.

C. *Definición y Entrenamiento del Modelo*

Debido a que el reconocimiento se basará en imágenes, la propuesta aquí presentada se desarrolla mediante el uso de una Red Neuronal Convolucional (CNN, por su sigla en inglés); este tipo de Red Neuronal Profunda es la más eficiente y útil para este tipo de datos [12][14]. El modelo de la CNN con el cual se obtuvo el mejor resultado, consta de dos capas convolucionales, seguidas de una capa de reducción por valor máximo y un aplanamiento de la red para capas completamente conectadas que funcionan para hacer predicciones, y además, se utilizó el optimizador SGD (*Stocastic Gradient Descent*). La cantidad y el tamaño de los filtros en las capas convolucionales se obtuvieron a través de una validación para diferentes valores de los mismos, siendo los valores que se especifican más adelante los que generaron mejores resultados; adicionalmente, en cada capa convolucional se utilizó una función de activación ReLU (*Rectified Linear Unit*). A continuación, se especifican las capas del modelo:

- Capa de entrada convolucional, 8 filtros con un tamaño de 2×2, función de activación ReLU.
- Capa convolucional, 4 filtros con un tamaño de 2×2, función de activación ReLU.
- Capa de reducción por valor máximo con tamaño 2×2.
- Apagado aleatorio (*dropout*) establecido en 20%.
- Capa de aplanado.
- Capa completamente conectada con 128 unidades y una función de activación del ReLU.
- Apagado aleatorio (*dropout*) establecido en 20%.
- Capa de salida completamente conectada con 30 unidades (clases) y una función de activación softmax.

Para el entrenamiento del modelo se utilizó el 60% de los datos pertenecientes al conjunto de datos y el 40% restante se utilizó para la validación del modelo. Esta segmentación se llevo a cabo utilizando la función *train_test_split* disponible en la librería *sklearn* en el lenguaje Python, porque dicha función permite garantizar que la segmentación sea equilibrada en cuanto a la proporción por cada clase, y además escoge de manera aleatoria las muestras que formarán parte de cada conjunto. En el proceso de entrenamiento, se intentó determinar en primer lugar un valor relativamente pequeño para las muestras por persona que fueron utilizadas (n) sin que el rendimiento resultara comprometido. Luego de tener los valores de la exactitud para los diferentes valores de n, se escogió el valor más pequeño que no incidiera demasiado en el rendimiento y se empezó a explorar la exactitud para diferentes valores de la ganancia (g) con el fin de intentar encontrar una mejora en el rendimiento para un n relativamente pequeño.

IV. Resultados

Cada vez que se entrenó el modelo, se realizó un seguimiento de la exactitud, tal como se muestra en la Figura 5, la cual contiene una gráfica que muestra las exactitudes a medida que transcurren las épocas; esta gráfica corresponde a un entrenamiento del mejor caso obtenido, donde se logra apreciar que no existe un sobre ajuste en el modelo.

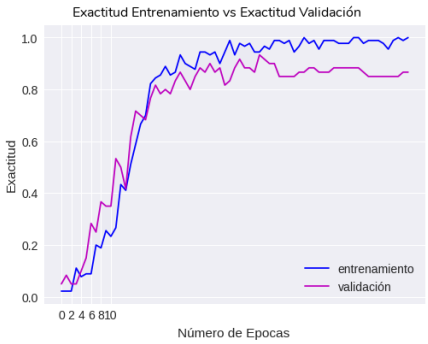

Figura 5. Exactitud en Entrenamiento con *n=5* y *g=0.9*

Los resultados obtenidos para la variación en el número de muestras por personas (n) se resumen en la Tabla I; de estos valores, se escogió un valor de *n=5* para experimentar con la ganancia. La elección de este valor en particular se debe a que representa un número de muestras manejable para los usuarios y la exactitud es cercana al 90%, lo cual es aceptable. A partir de este punto los resultados reportados fueron generados utilizando *n=5*.

Tabla I. Exactitud para Diferentes Valores de n

n	Exactitud de entrenamiento	Exactitud de validación
20	0.9783	0.9733
15	0.9741	0.9678
10	0.9711	0.9600
5	0.9667	0.8899

Con respecto a la ganancia, utilizando el valor por defecto (*g=1*, lo cual se traduce como ausencia de ganancia) se obtuvo un valor de exactitud en la validación del *88.99%*. Se observó una mejoría en la exactitud para valores de la ganancia comprendidos entre *0.7* y *30*, situándose los mejores resultados en valores cercanos a *0.9* y *20*. La Figura 6 contiene una gráfica con el reporte de los resultados obtenidos para diferentes valores de ganancia, donde el mejor rendimiento registrado se generó con una ganancia igual a *0.9* que registró una exactitud del *93.34%*.

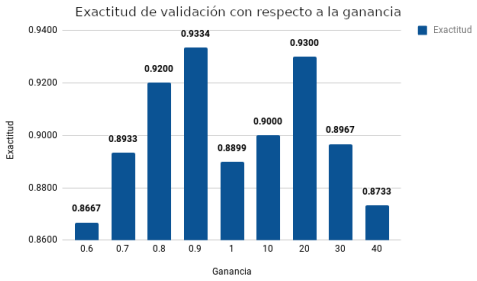

Figura 6. Exactitudes para Diferentes Valores de g con *n=5*

En la Figura 7 se aprecia la matriz de confusión perteneciente a la validación de un modelo generado con la combinación de *n=5* y *g=0.9*. En esta matriz se observa que, pese a la baja cantidad de muestras para la validación (*2* por persona), en un *13.33%* de los casos la persona fue identificada en una ocasión de manera exitosa y en un *86.67%* de los casos la persona fue identificada en ambas ocasiones de manera exitosa.

V. Discusión

Con respecto a los resultados obtenidos, se observa un buen porcentaje de exactitud de validación correspondiente a *97.33%* en el caso de utilizar *20* muestras, lo cual arroja un indicio que la solución basada en espectrogramas aquí propuesta funciona con un buen rendimiento; adicionalmente, se logró reducir el tamaño de las muestras a sólo *5* por persona (*3* para el entrenamiento, *2* para la validación) sin que el porcentaje de exactitud resultara afectado gravemente, obteniendo un porcentaje cercano al *90%*, el cual probablemente sea aceptable para aplicaciones que no sean de uso crítico; además, este porcentaje se logró aumentar gracias a una variación en la ganancia utilizada en la generación de los espectrogramas. En la Figura 6 se observa un comportamiento bimodal en la forma como se distribuye la exactitud, cuyos

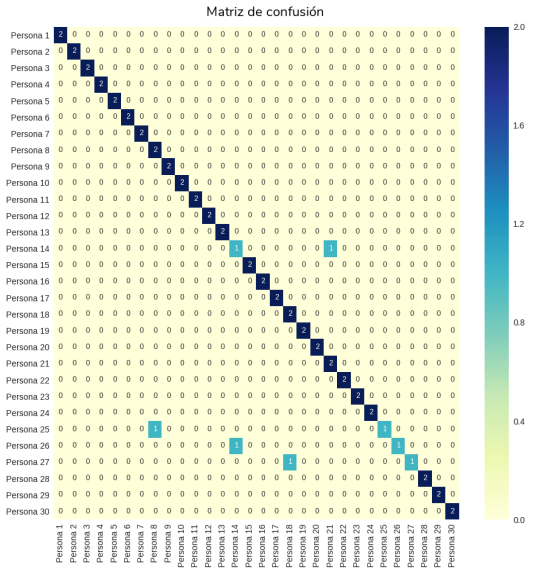

Figura 7. Matriz de Confusión de un Modelo con *n=5* y *g=0.9*

valores más altos fueron encontrados con una ganancia igual a *0.9* y *20*, encontrándose uno a la izquierda y otro a la derecha del valor predeterminado de la ganancia *g=1*. Una interpretación de estos resultados es, que un aumento o disminución en la ganancia puede traer consigo un mejor rendimiento. Cabe destacar que, en tanto el valor de la ganancia tienda a los extremos, la exactitud tenderá a bajar. Con respecto a la matriz de confusión presente en la Figura 7, se puede observar que en 4 ocasiones el modelo clasificó de manera incorrecta las grabaciones y, luego de consultar la información correspondiente a las grabaciones recolectadas, se pudo apreciar que los casos donde el modelo confundió a dos individuos, corresponden a personas del mismo sexo y aproximadamente de la misma edad, lo cual puede justificar las equivocaciones en el modelo.

VI. Conclusiones

En función de los resultados obtenidos, se puede decir que el sistema de reconocimiento de personas a través de la voz utilizando imágenes (espectrogramas) para entrenar una red neuronal convolucional, alcanzó un porcentaje de exactitud de validación aceptable, considerando que el enfoque propuesto busca minimizar la cantidad de muestras por persona utilizadas y presentar una alternativa a la resolución del problema de reconocimiento de personas. Aunque la tasa de error obtenida es mayor a las alcanzadas en [1][5][10], el modelo propuesto es notablemente más simple y el conjunto de datos utilizado es en gran medida de menor tamaño. El sistema propuesto demostró ser suficiente para ser implementado en aplicaciones que no requieran de un gran porcentaje de exactitud y en las cuales se produzcan interacciones con una cantidad reducida de personas en un ambiente estable.

Una de las limitaciones encontradas se sitúa en el proceso de recolección de datos, porque aunque se construyó un buen conjunto, los datos correspondientes a cada persona fueron tomados en una sola sesión, por lo tanto, una dirección a tomar en trabajos posteriores sería recolectar los datos de cada persona en múltiples ocasiones (diferentes días en un horario variado), con el fin de considerar la pequeña variación a la cual está sujeta la voz en un intervalo razonable de tiempo. Otra limitación encontrada fue la dependencia del dispositivo de grabación, el cual tiene incidencia en la calidad del sonido y, por lo tanto, las grabaciones correspondientes a una persona particular se deben realizar con el mismo micrófono que se utilizó para capturar las muestras pertenecientes al conjunto de entrenamiento. Se recomienda para trabajos futuros probar métodos adicionales en la sección de procesamiento con la finalidad de estandarizar en mayor medida las imágenes generadas.

Referencias

[1] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end Text-dependent Speaker Verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5115–5119.

[2] C. Vazquez-Machado, P. Colon-Hernandez, and P. Torres-Carrasquillo, "I-vector Speaker and Language Recognition System on Android," in *2016 IEEE High Performance Extreme Computing Conference (HPEC)*, Sept 2016, pp. 1–6.

[3] N. Desai and N. Tahilramani, "Digital Speech Watermarking for Authenticity of Speaker in Speaker Recognition System," in *2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE)*, Sept 2016, pp. 105–109.

[4] Y. Zhan, H. Leung, K. C. Kwak, and H. Yoon, "Automated Speaker Recognition for Home Service Robots Using Genetic Algorithm and Dempster-Shafer Fusion Technique," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 9, pp. 3058–3068, Sept 2009.

[5] E. Marchi, S. Shum *et al.*, "Generalised Discriminative Transform via Curriculum Learning for Speaker Recognition," 2018. [Online]. Available: http://sigport.org/3144

[6] M. Zhang, Y. Chen, L. Li, and D. Wang, "Speaker Recognition with Cough, Laugh and "Wei"," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Dec 2017, pp. 497–501.

[7] T. Chenm and G. Liu, "Researches Based on Neural Network about Optimizing Combination of Speaker Verification," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, Dec 2017, pp. 1961–1965.

[8] A. Awais, S. Kun, Y. Yu, S. Hayat, A. Ahmed, and T. Tu, "Speaker Recognition using Mel Frequency Cepstral Coefficient and Locality Sensitive Hashing," in *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, May 2018, pp. 271–276.

[9] K. Cho, J. Roh, Y. Han, N. Kim, and J. Lee, "Real-time Speaker Recognition System Using Multi-stream i-vectors for AI Assistant," in *2018 IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2018, pp. 1–4.

[10] B. Makrem, J. Imen, and O. Kaïs, "Study of Speaker Recognition System Based on Feed Forward Deep Neural Networks Exploring Text-dependent Mode," in *2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, Dec 2016, pp. 355–360.

[11] F. Plante, G. Meyer, and W. A. Ainsworth, "Improvement of Speech Spectrogram Accuracy by the Method of Reassignment," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 282–287, May 1998.

[12] S. Albawi, T. Mohammed, and S. Al-Zawi, "Understanding of a Convolutional Neural Network," in *2017 International Conference on Engineering and Technology (ICET)*, Aug 2017, pp. 1–6.

[13] *FFmpeg*. [Online]. Available: https://ffmpeg.org/

[14] N. Jmour, S. Zayen, and A. Abdelkrim, "Convolutional Neural Networks for Image Classification," in *2018 International Conference on Advanced Systems and Electric Technologies*, March 2018, pp. 397–402.