Revista RITI
Investigación en Tecnologías de la Información

Revista de Investigación en Tecnologías de la Información
ISSN: 2387-0893
editorial@riti.es
Universitat Politècnica de Catalunya
España

Robles Contreras, Carmen Victoria; Carrillo Ruiz, Maya;
Hernández Ameca, José Luis; Robles Mendoza, Francisco Javier

Identificación del acento en hablantes de español mediante el análisis de atributos MFCC y aprendizaje supervisado

Revista de Investigación en Tecnologías de la Información,
vol. 12, núm. 26, 2024, julio-diciembre, pp. 19-27
Universitat Politècnica de Catalunya
España

DOI: https://doi.org/10.36825/RITI.12.26.002

Identificación del acento en hablantes de español mediante el análisis de atributos MFCC y aprendizaje supervisado

Accent identification in spanish speakers through MFCC attribute analysis and supervised learning

Carmen Victoria Robles Contreras
Benemérita Universidad Autónoma de Puebla
carmen.roblesco@alumno.buap.mx
ORCID: 0000-0001-7964-7883

Maya Carrillo Ruiz
Benemérita Universidad Autónoma de Puebla
maya.carrillo@correo.buap.mx
ORCID: 0000-0001-6152-456X

José Luis Hernández Ameca
Benemérita Universidad Autónoma de Puebla
joseluis.hdxameca@correo.buap.mx
ORCID: 0000-0002-7672-5409

Francisco Javier Robles Mendoza
Benemérita Universidad Autónoma de Puebla
francisco.roblesm@correo.buap.mx
ORCID: 0009-0007-3176-5005

doi: https://doi.org/10.36825/RITI.12.26.002

Recibido: Julio 19, 2023
Aceptado: Agosto 19, 2024

Resumen: El reconocimiento del hablante tiene múltiples aplicaciones en la vida real. El propósito de este estudio es determinar la viabilidad de clasificar muestras de habla humana, específicamente hablantes de español, a partir de su acento distintivo. En este trabajo se utilizaron los Coeficientes Cepstrales en las Frecuencias de Mel combinados con algoritmos de aprendizaje automático, tales como: *Random Forest*, KNN, SVM, SGD y Redes Neuronales, para identificar la nacionalidad de personas hispanohablantes por medio de grabaciones de voz obtenidas del corpus *Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech*. Se realizó un preprocesamiento de los datos, extrayendo 50 MFCC de cada grabación, con estos se construyó el conjunto de datos para la experimentación. Se realizaron experimentos con diferentes subconjuntos. Los mejores resultados se obtuvieron con individuos pertenecientes a cuatro países de Latinoamérica, incluyendo individuos del sexo masculino y femenino. Para la etapa de clasificación se utilizaron redes neuronales. La precisión obtenida fue de 99.84%.

Palabras clave: *Reconocimiento del Acento, MFCC, Algoritmos de Aprendizaje Automático, Aprendizaje Supervisado.*

Abstract: Speaker recognition has multiple real-life applications. The purpose of this study is to determine the feasibility of classifying samples of human speech, specifically Spanish speakers, based on their distinctive accents. In this work, Mel-Frequency Cepstral Coefficients (MFCC) combined with machine learning techniques were used to identify the nationality of Spanish-speaking individuals through voice recordings obtained from the Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech corpus. Data preprocessing was performed by extracting 50 MFCC from each recording, which formed the dataset for experimentation. Experiments were conducted with different subsets, and the best results were obtained with individuals from four Latin American countries, including both males and females. Neural networks were employed for the classification stage, achieving an accuracy of 99.84%.

Keywords: *Accent Recognition, MFCC, Machine Learning Algorithms, Supervised Learning.*

1. Introducción

El reconocimiento del hablante tiene múltiples aplicaciones en la vida real. Desde implementaciones en seguridad, como en la personalización de servicios interactivos. Para tal caso, los Coeficientes Cepstrales en las Frecuencias de Mel (MFCC, por sus siglas en inglés) combinados con técnicas de aprendizaje automático se han utilizado últimamente en diferentes trabajos de reconocimiento del hablante. Estos coeficientes también se han utilizado para el reconocimiento del acento entre hablantes del mismo idioma, por ejemplo, el reconocimiento de la nacionalidad de migrantes en países anglosajones [1].

La presente investigación utiliza los MFCC para identificar el origen de una persona hispanohablante, empleando grabaciones de individuos de diferentes países de Latinoamérica.

El objetivo fundamental es establecer si es posible identificar y distinguir acentos en el habla, lo que podría tener aplicaciones en campos como la lingüística forense, la tecnología del reconocimiento de voz y la investigación sociolingüística. Se busca demostrar que esta clasificación es viable y efectiva.

Este trabajo está organizado de la siguiente manera: En la sección 2 se presenta el estado del arte, en la sección 3 se explica conceptos básicos que incluyen el algoritmo de extracción de los MFCC y los algoritmos de clasificación utilizados, en la sección 4 se exponen las características del conjunto de datos utilizado, en la sección 5 se presenta la metodología, después en la sección 6, se discuten los resultados. Por último, se plantean las conclusiones y trabajo futuro, en la sección 7.

2. Estado del arte

Honnavalli D. y Shylaja S. S. en [2] plantean mejorar el reconocimiento y la respuesta de voz ya que hay un mercado creciente de tecnologías en todo el mundo que requiere de dichos métodos. La mayoría de los sistemas de reconocimiento de voz a gran escala no tienen exposición a una amplia variedad de acentos de hablantes del inglés, lo que desfavorece el reconocimiento preciso del hablante. Así que, la clasificación del acento es una característica importante que se puede utilizar para aumentar la precisión de los sistemas de reconocimiento de voz. Los autores, presentan un método para distinguir a los hablantes de inglés de la India y de los Estados Unidos a través de su acento. Para ello extraen los MFCC del corpus VCTK, se realiza sobremuestreo de los datos poco representados y se aplican técnicas de aprendizaje supervisado. La precisión más alta se alcanza con redes neuronales y es del 95%. En total se utilizan 5 algoritmos alcanzando un promedio de precisión del 76%. Los resultados obtenidos indican que la concatenación de características MFCC de manera secuencial y la aplicación de una técnica de aprendizaje supervisado adecuada en los datos proporcionan una buena solución al problema de detección y clasificación de acentos.

En [3] Mannepalli *et al*., hablan sobre la importancia de la detección de acentos, en idiomas con una variedad de estos. El telugu es un idioma hablado ampliamente en el sur de la India que tiene diferentes acentos, incluyendo Andhra Costera, Telangana y Rayalaseema. En este trabajo, se recolectaron muestras de hablantes nativos de diferentes acentos del idioma telugu tanto para el entrenamiento como para las pruebas. Se extrajeron características por medio de MFCC para cada muestra y se utilizó el modelo de mezcla gaussiana (GMM) para clasificar el habla en función del acento. La precisión general del sistema, propuesto para reconocer al hablante y su región de origen basado en el acento, fue del 91%.

Ma, Z. y Fokoué E. en [4] comparan diferentes clasificadores utilizando MFCC. Para cada señal de voz, se utiliza el vector de media de la matriz MFCC como entrada para el clasificador. Es decir, para cada señal, los MFCC en realidad forman una matriz *n x q* donde *n* es el número de marcos (*frames*) de ventana y *q* es el número de MFCC. Así se tomar los valores medios de cada uno de los *n* vectores de columna. Se analiza una muestra de 330 señales, que incluye 165 voces estadounidenses y 165 voces no estadounidenses. Los resultados muestran que, en comparación con otros clasificadores, el algoritmo de vecinos más cercanos *(k-nearest neighbors*) tiene una precisión promedio igual a 93.98 % en las pruebas, después de usar validación cruzada de tamaño 500.

3. Conceptos básicos

En esta sección se presentan brevemente la definición de los Coeficientes Cepstrales en las Frecuencias de Mel y los algoritmos de aprendizaje utilizados para la obtención y comparación de los resultados.

3.1. MFCC

Los Coeficientes Cepstrales en las Frecuencias de Mel sirven para la representación del audio basada en la percepción auditiva humana. Un problema fundamental en el procesamiento de sonido, particularmente, del habla, consiste en obtener una codificación compacta de las características del archivo de audio. La técnica más usada para la extracción de estas características son los Coeficientes Cepstrales en las Frecuencias de Mel, como puede observarse en los trabajos presentados en [2] y [3]. En esencia, los MFCC se utilizan para extraer características de una señal de audio que sean útiles para una tarea, eliminando ruido de fondo y otras señales que la distorsionen. El proceso de extracción de los MFCC se compone por:

a. *Pre-enfatizado*: La señal primeramente se pre-enfatiza para potenciar los componentes de alta frecuencia.
b. *Muestreo*: La señal que ya fue pre-enfatizada se divide en muestras sobrepuestas cortas, típicamente de 20-30 milisegundos de duración.
c. *Función de ventana*: Una función de ventana se aplica a cada marco para reducir el manchado espectral proveniente de la discontinuidad entre las muestras recortadas.
d. *Transformada de Fourier*: Se aplica una transformación rápida de Fourier (FFT) a cada muestra para convertir el dominio de la señal de tiempo a frecuencia.
e. *Filtros de Mel:* La densidad espectral de potencia resultante se convierte a escala de frecuencia de Mel, que es una escala de frecuencias perceptualmente definidas que refleja la forma en que los humanos percibimos el sonido.
f. *Análisis Cepstral:* Una transformada de coseno discreta se aplica al logaritmo del espectro de frecuencias de Mel para obtener un conjunto de coeficientes cepstrales. Los primeros coeficientes son típicamente rechazados, puesto que suelen relacionarse al volumen de la señal, dejando un conjunto pequeño de coeficientes que se consideran con mayor relevancia. [1]

3.2. Algoritmos de clasificación

La clasificación supervisada es el proceso mediante el que un algoritmo analiza las características de un conjunto de datos y crea un modelo poblacional en base a una fracción de los datos de entrada, el cual se define como "conjunto de entrenamiento". Posteriormente, con la fracción restante de datos se realizan pruebas, dejando que el modelo clasifique por su cuenta los datos y luego comparando las clases asignadas por el modelo con las clases reales, a partir del número de aciertos se calculan medidas de evaluación del modelo.

a. ***Máquina de soporte vectorial (SVM).*** SVM es un algoritmo de clasificación supervisada que fue presentado por primera vez por Vladimir Vapnik y sus colegas en 1990. La idea básica detrás de la SVM es encontrar el hiperplano que separa dos clases de forma que se maximice el margen entre las clases. Ese margen está definido como la distancia entre el hiperplano y los nodos más cercanos a él de cada clase. Las SVMs pueden manejar la separación no lineal de datos transformando el espacio de ubicación en uno de mayor dimensión, donde se puede separar de forma lineal. Esto se hace a partir de una función *kernel* que computa el producto interno entre parejas de nodos en el espacio de dimensión mayor. Tienen ventajas con respecto a otros algoritmos de clasificación. Son efectivas en

espacios de muchas dimensiones, tienen un fuerte fundamento teórico, y pueden separar datos lineales y no lineales. Además de que son menos propensos a hacer sobreajuste que otros algoritmos [5].

b. ***Descenso de gradiente estocástico (SGD).*** El descenso de gradiente estocástico es un algoritmo de optimización usado para el entrenamiento de modelos de aprendizaje automático, particularmente redes neuronales profundas. Consiste en minimizar la función de costo al ajustar de forma iterativa los parámetros del modelo en dirección del gradiente de la función de costos. Estrictamente hablando, SGD no es un modelo de aprendizaje, sino más bien una técnica que se combina con otros modelos para mejorar su rendimiento [6].

c. ***K-vecinos más cercanos (KNN).*** El algoritmo de los K-vecinos es un algoritmo no paramétrico usado para la clasificación y regresión. Funciona asignando a la muestra una clase de acuerdo con la clase de sus "vecinos", es decir, aquellas muestras que se encuentran "cercanas" de acuerdo con una función de distancia que las ubica por sus características en un plano. En el caso de que existan varias muestras de clases diferentes a la misma distancia, ocurre un empate. Los empates se rompen de forma arbitraria. Tiene varias ventajas, incluyendo su simplicidad, flexibilidad y robustez [7].

d. ***Random Forest.*** El algoritmo *Random Forest* es un método de aprendizaje basado en ensamblaje que combina múltiples árboles de decisión para hacer una predicción final. En el caso de la clasificación, cada árbol predice independientemente la clase de una muestra y la clase con más votos es escogida como la predicción final. La idea clave de este algoritmo es introducir un porcentaje de aleatoriedad a la construcción de cada árbol, al seleccionar un subconjunto de características al azar para cada árbol [8].

e. **Redes neuronales.** Las redes neuronales son un tipo de modelo de aprendizaje automático que imitan la forma en que funciona el cerebro humano. Estas redes están compuestas por capas de nodos interconectados que procesan y transmiten información a través de conexiones a las que se les asignan pesos. De forma sencilla puede ser visto como una colección de neuronas organizadas por capas, donde las neuronas de una capa solo se interconectan por un enlace dirigido ponderado con las neuronas de la capa posterior [9]. Buscan simular un proceso de aprendizaje aplicando diferentes herramientas de optimización, las cuales permiten, luego de una serie de recorridos a través de la estructura de la red neuronal, ajustar la función que facilita la predicción de la variable dependiente. Los errores encontrados en el proceso se emplean para retroalimentar el proceso de cálculo [10]. En términos de clasificación, las redes neuronales se utilizan para predecir la clase a la que pertenece un objeto o dato de entrada. Para hacerlo, se utiliza un conjunto de datos de entrenamiento que contiene ejemplos etiquetados con sus respectivas clases. El modelo de red neuronal ajusta gradualmente sus pesos con base en los datos de entrenamiento, de manera que la salida de la red se aproxime cada vez más a las etiquetas correctas. Los perceptrones multicapa (MLP, por sus siglas en inglés) son una clase de redes neuronales artificiales que consisten en una serie de capas de nodos o unidades interconectadas. Cada unidad está conectada a todas las unidades de la capa anterior y de la siguiente capa. En una red MLP, la información fluye desde la capa de entrada a través de las capas ocultas hasta la capa de salida. Cada capa oculta procesa la información recibida de la capa anterior y la transmite a la siguiente capa a través de conexiones ponderadas. La principal ventaja de los MLP es su capacidad para aprender funciones no lineales y complejas. Esto se debe a que las múltiples capas ocultas permiten que la red aprenda representaciones jerárquicas de los datos de entrada.

4. Metodología

La metodología utilizada está definida por tres pasos que se muestran en la Figura 1. En el preprocesamiento los audios representados como una relación entre tiempo y amplitud, se normalizan y se trasladan al dominio del tiempo y la frecuencia, para tener espectrogramas. Posteriormente, se extraen los coeficientes MFCC para cada audio. Dichos coeficientes permiten extraer características adecuadas de los audios para identificar contenido relevante. Obtenidos los coeficientes MFCC, a estos se le asigna una clase, en este caso la nacionalidad del hablante. Finalmente, los audios etiquetados se clasifican con diferentes algoritmos.

Figura 1. Pasos de la metodología.

5. Corpus utilizado

El corpus utilizado fue *Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech [*11] creado en el 2020, el cual está conformado por muestras de audio de alta calidad de frases grabadas por voluntarios latinoamericanos provenientes de: Argentina, Chile, Colombia, Perú, Puerto Rico y Venezuela, en idioma español. Estas muestras están en formato *.wav* y contienen su transcripción de manera adjunta en un archivo *.tsv*. Las muestras de audios también están separadas por género. En la Tabla 1 se observan las grabaciones en formato *.wav* para dos mujeres colombianas.

Tabla 1. Entradas contenidas en el corpus.

Nombre	Formato	Duración(seg.)
cof_00610_00008989777	.wav	5
cof_01523_01430571518	.wav	6

Fuente: Elaboración propia.

La cantidad de hablantes por país varía de 5 hasta 31, con un promedio de 17. En la Tabla 2 se presentan el número de muestras de audio de los países considerados para la experimentación.

Tabla 2. Número de muestras por país y género.

País	Género	Número de muestras
Chile	Masculino	2636
Chile	Femenino	1738
Venezuela	Masculino	1754
Venezuela	Femenino	1603
Perú	Masculino	2918
Perú	Femenino	2529
Colombia	Masculino	2534
Colombia	Femenino	2369

Fuente: Elaboración propia.

Todos los hablantes grabados son hablantes nativos del país y radicados ahí, excluyendo a Puerto Rico y Venezuela, donde los hablantes fueron grabados en Nueva York, San Francisco y Londres. La selección de frases fue diseñada para un sistema de conversación mexicano, por lo que se extrajeron todas las palabras exclusivamente usadas en México y se agregaron frases adicionales. Mientras que se usaron las frases estándar, se les permitió a los hablantes improvisar y usar la cadencia que sintieran natural para su dialecto. Treinta frases "canónicas" pertenecientes al corpus fueron grabadas por cada uno de los hablantes, de forma que pudieran estudiarse los contrastes fonéticos del dialecto.

6. Experimentos y resultados

A continuación, se ilustran los pasos de la metodología con el corpus utilizado.

6.1. Preprocesamiento

El corpus original en promedio por nacionalidad tiene 4520 datos. Este fue balanceado en cuanto a país y género considerando únicamente 3206 grabaciones por nacionalidad. Así el corpus final utilizado está formado por 12824 grabaciones. En la Figura 2, puede verse graficada una grabación, el audio está representar como una relación entre tiempo y amplitud y puede observarse que el audio no está normalizado, por lo que presenta picos y hundimientos alejados del valor medio. Por lo que previamente a la extracción de los coeficientes MFCC se normalizó fijando la longitud de las muestras utilizadas a 93 milisegundos, con un muestreo de 22050 Hz, debido a que todos los audios se encuentran en estos o por arriba de estos valores.

El siguiente paso, de acuerdo con la metodología, fue trasladar los audios del dominio de la amplitud al de la frecuencia, como se visualiza en la Figura 3, donde puede observarse un espectrograma que es una representación visual del espectro de frecuencias del sonido conforme varia con el tiempo de la grabación de la Figura 2.

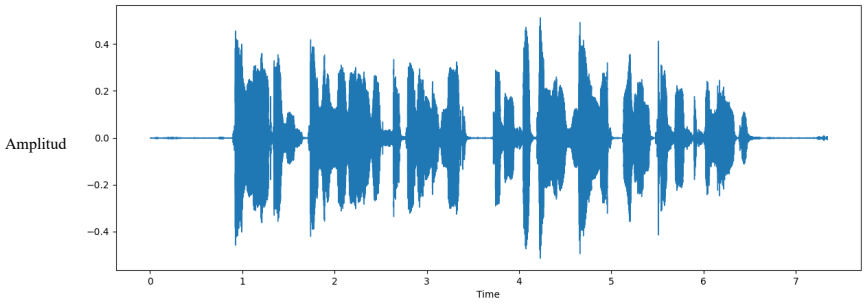

Figura 2. Relación entre amplitud y tiempo de una muestra de audio.

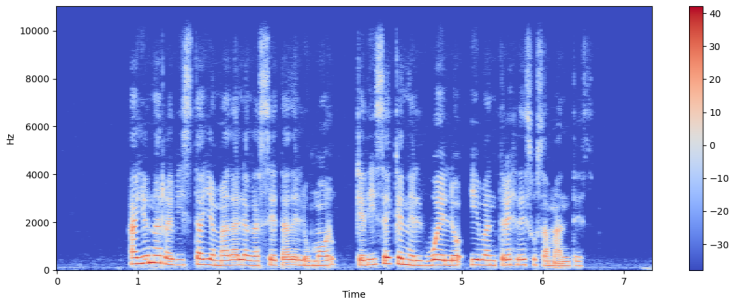

Figura 3. Espectrograma de una muestra de audio.

6.2. Extracción de MFCC

Concluido el preprocesamiento de las grabaciones, se caracterizó cada muestra por medio de sus MFCC. En la Figura 4, se puede observar la forma de los MFCC. Los parámetros de la representación son el tiempo y el valor del coeficiente. Podemos observar que, como es común en este tipo de representaciones, la mayoría de los valores calculados se encuentran dentro de un mismo rango, con pocos valores que se alejan de la media.

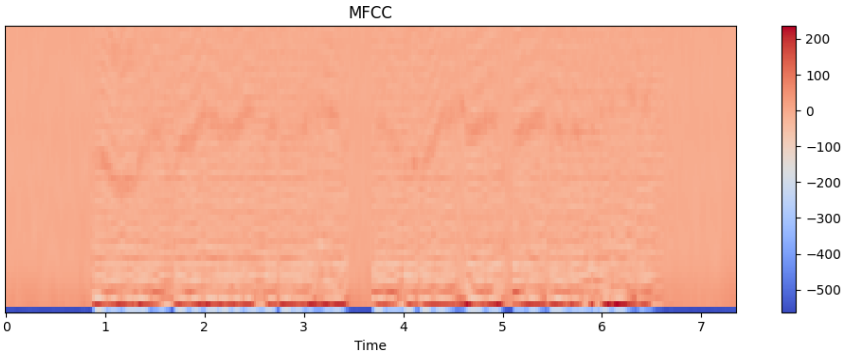

Figura 4. Forma de los MFCC's calculados de una muestra de audio.

Por medio de la librería Librosa [12] en Python 3.10.9, que contiene métodos para el análisis de música y audio se calcularon los MFCC. Obtenidos los MFCC a estos se le asigno una clase. Recuérdese que el objetivo es identificar la nacionalidad de los hablantes a partir de su acento. Para tal efecto, se etiquetaron los datos de la siguiente manera: 0 – Chilenos, 1 - Venezolanos, 2 - Peruanos, 3 – Colombianos. A partir de los MFCC y el atributo clase se creó un conjunto de datos por país. Estos conjuntos de datos se combinaron para clasificarlos.

6.3. Clasificación

La implementación de los clasificadores se hizo por medio de la librería *sklearn* [13]. Se seleccionaron 50 MFCC por muestra. Esta cantidad fue determinada experimentalmente. Mannepalli *et al*. [3], describe que un número mayor mejora la clasificación. Así que se hicieron pruebas usando el clasificador K-Vecinos por su simplicidad, variando el número de MFCC de un rango de 40 hasta 55. Los primeros experimentos aumentaron la exactitud y precisión de forma proporcional al número de coeficientes, mientras que las pruebas con 55 MFCC no muestran una mejora significativa, por lo que se mantuvieron 50 MFCC por muestra de audio. Como puede observarse en la Tabla 3.

Tabla 3. Número de MFCC.

Número de MFCC	Exactitud	Precisión
40	93.10%	97.5%
45	94.56%	98.23%
50	95.63%	98.81%
55	95.13%	98.29%

Fuente: Elaboración propia.

Una vez determinada la cantidad de MFCC a extraer, los datos se dividieron en una proporción 80, 20. La primera para el entrenamiento, y la segunda para las pruebas. Al inicio se realiza un barajeo aleatorio de los datos y se utilizaron los algoritmos de clasificación explicados en la sección 3.

Se inició la experimentación tomando un corpus conformado por muestras de dos nacionalidades, solamente incluyendo mujeres. Se utilizó validación cruzada a 10 pliegues para sacar promedios de diferentes métricas. Los resultados pueden verse en la Tabla 4.

Tabla 4. Clasificación binaria: mujeres Chilenas y Venezolanas.

Herramienta	Exactitud	Precisión	Recuerdo
Random Forest	98.28%	98.75%	98.43%
KNN	95.63%	95.56%	94.91%
SVM	83.75%	85.86%	83.22%
SGD	91.25%	92.19%	91.25%

Fuente: Elaboración propia.

Con base en los resultados de la Tabla 4, donde puede observarse que la exactitud es más del 80% en todos los casos, se decidió ampliar el conjunto de datos, para incluir también hombres. Los resultados obtenidos se muestran en la Tablas 5.

Tabla 5. Clasificación binaria: Hombres y mujeres Chilenos y Venezolanos.

Herramienta	Exactitud	Precisión	Recuerdo
Random Forest	97.31%	99.33%	96.33%
KNN	97.43%	98.81%	97.82%
SVM	80.32%	99.30%	82.74%
SGD	96.72%	95.91%	91.98%

Fuente: Elaboración propia.

Al obtenerse, nuevamente resultados satisfactorios, el corpus fue ampliado, agregando dos nacionalidades, así como implementando un clasificador a base de redes neuronales. Los resultados obtenidos pueden mirarse en la Tabla 6, donde se observa que las redes neuronales, tienen los mejores resultados generales, siendo superadas en exactitud únicamente por el algoritmo *Random Forest*. En general, los resultados están arriba del 80% a excepción de la exactitud y precisión obtenida con SVM.

Tabla 6. Resultados comparando 4 países y mujeres y hombres.

Herramienta	Exactitud	Precisión	Recall
Random Forest	**97.12%**	99.68%	97.08%
KNN	94.50%	98.71%	96.52%
SVM	73.92%	61.60%	**99.25%**
SGD	80.47%	99.03%	81.13%
Redes Neuronales	96.41%	**99.84%**	98.74%

Fuente: Elaboración propia.

7. Conclusiones

Después de realizar los experimentos descritos, se concluye que los MFCC son útiles para caracterizar el acento de los hispanohablantes, además de que son fáciles de utilizar, no requieren mucho espacio de almacenamiento ni demandan gran capacidad de cómputo. Los algoritmos de clasificación se comportan de la forma esperada, a excepción de la SVM que obtiene los resultados más bajos. Destacando especialmente, las Redes Neuronales se erigieron como el método de clasificación más efectivo, logrando una precisión del 99.84% y una exactitud del 96.41%. Estos resultados no solo validan la utilidad de los MFCC, sino también equiparan el rendimiento de las redes neuronales al estado del arte en el campo del reconocimiento de acentos, consolidando su relevancia en aplicaciones prácticas, desde la seguridad en la autenticación de voz hasta la adaptación personalizada de servicios de voz.

10. Referencias

[1] Wei, H., Cheong-Fat, C., Chiu-Sing, C., Kong-Pang, P. (2006). *An efficient MFCC extraction method in speech recognition*. IEEE International Symposium on Circuits and Systems (ISCAS), Kos, Greece. https://doi.org/10.1109/ISCAS.2006.1692543

[2] Honnavalli, D., Shylaja, S. S. (2019). *Supervised Machine Learning Model for Accent Recognition in English Speech Using Sequential MFCC Features Advances in Artificial Intelligence and Data Engineering.* International Conference on Artificial Intelligence and Data Engineering, Udupi, India. https://doi.org/10.1007/978-981-15-3514-7

[3] Mannepalli, K., Narahari Sastry, P., Suman, M. (2016). MFCC-GMM based accent recognition system for Telugu speech signals. *International Journal of Speech Technology, 9* (19), 87-93. https://doi.org/10.1007/s10772-015-9328-y

[4] Ma, Z., Fokoué, E. (2014). A Comparison of Classifiers in Performing Speaker Accent Recognition Using MFCCs. *Open Journal of Statistics, 4* (4)*,* 258-266*.* http://dx.doi.org/10.4236/ojs.2014.44025

[5] Chervonenkis, A. Y. (2013). Early History of Support Vector Machines*.* En B. Schölkopf, Z. Luo, V. Vovk, (Eds.) *Empirical Inference* (pp. 13-20). Springer. https://doi.org/10.1007/978-3-642-41136-6_3

[6] Sigtia, S., Dixon, S. (2014). *Improved Music Feature Learning with Deep Neural Networks*. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Florence, Italy. http://dx.doi.org/10.1109/ICASSP.2014.6854949

[7] Viswanath, P., Hitendra Sarma, T. (2011). *An improvement to k-nearest neighbor classifier*. IEEE Recent Advances in Intelligent Computational Systems (RAICS), Trivandrum, India. http://dx.doi.org/10.1109/RAICS.2011.6069307

[8] Taha Jijo, B., Mohsin Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal Of Applied Science And Technology Trends, 2* (01), 20-28. http://dx.doi.org/10.38094/jastt20165

[9] Urquiza Aguiar, L., Campos Yucailla, P., Hidalgo Lascano, P., Becerra Camacho, F. (2020). Detección de Nodos en Zonas Ocultas en redes LAA a través de Aprendizaje Automático Supervisado. *Revista De Investigación en Tecnologías de la Información (RITI)*, *8* (15), 114–127. https://doi.org/10.36825/RITI.08.15.011

[10] del Castillo Collazo, N. (2020). Predicción en el diagnóstico de tumores de cáncer de mama empleando métodos de clasificación. *Revista De Investigación En Tecnologías De La Información (RITI), 8* (15), 96–104. https://doi.org/10.36825/RITI.08.15.009

[11] Guevara-Rukoz, A., Demirsahin, I., He, F., Chu, S. C., Sarin, S., Pipatsrisawat, K., Gutkin, A., Butryna, A., Kjartansson, O. (2020). *Crowdsourcing Latin American Spanish for Low-Resource Text-to-Speech*. Proceedings of the 12th Conference on Language Resources and Evaluation, Marseille, France. https://aclanthology.org/2020.lrec-1.801

[12] McFee, B., McVicar, M., Faronbi, D., *et al.* (2023). *librosa/librosa: 0.10.1*. https://doi.org/10.5281/zenodo.8252662

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12* (85), 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html