Ciencia y Mar
2025, XXIX (86): 77–85
DOI: 10.59673/cym.v29i86.9

Acceso Abierto

Información Científica y Tecnológica

Extracción de características en señales de audio digital usando espectrogramas para la clasificación de géneros musicales

Jorge Ochoa-Somuano[1] Ricky Didier Peralta-Reyes[1]
José Francisco Delgado-Orta*[1] & María Alejandra Menéndez-Ortiz[1]

Resumen

La música ha sido una herramienta fundamental para la comunicación humana, al fungir como un medio para transmitir sentimientos y emociones. *Su forma y características varían según el contexto histórico y cultural en el que se desarrolla, lo que ha dado lugar a la aparición de diversos géneros musicales a lo largo de la historia en distintas culturas. Con el avance de la tecnología en las últimas décadas y el incremento en la accesibilidad a las grabaciones, el volumen de pistas musicales ha crecido exponencialmente. Lo anterior ha llevado a un incremento significativo en el tiempo y esfuerzo requeridos para su clasificación manual, lo que ha evidenciado la necesidad de automatizar dicho proceso. En el mismo contexto, este trabajo propone una metodología para la clasificación automática de géneros musicales, basada en el uso de espectrogramas como herramienta para la extracción de características. Específicamente, se emplean métricas estadísticas como la media y la desviación estándar para la representación de dichas características y, clasificadores como K-vecinos más cercanos y Máquinas de Vectores de Soporte para la validación del modelo. Los resultados de la clasificación muestran una exactitud de hasta 81% para el género clásico, 63% para el jazz y 71% para el rock.*

Palabras clave: clasificación, espectrogramas, Mel y MFCC, procesamiento de señales digitales.

Abstract

Music has been a fundamental tool for human communication, serving as a medium for conveying feelings and emotions. *Its form and characteristics vary depending on the historical and cultural context in which it develops, leading to the emergence of diverse musical genres throughout history across various cultures. With technological advancements over recent decades and the increased accessibility of recordings, the volume of musical tracks has grown exponentially. This proliferation has significantly increased the time and effort required for manual classification, underscoring the necessity of automating this process. In this context, we propose a methodology for the automatic classification of musical genres, leveraging spectrograms as a tool for feature extraction. Specifically, statistical metrics such as the mean and standard deviation are employed to represent these features, while classifiers like K-Nearest Neighbors and Support Vector Machines are utilized for model validation. The classification process achieved accuracy levels of up to 81% for the classical genre, 63% for jazz, and 71% for rock.*

Key words: classification, spectrograms, Mel y MFCC, digital signal processing.

Recibido: 26 de febrero de 2025. **Aceptado:** 28 de abril de 2025.

[1] *Instituto de Industrias, Universidad del Mar campus Puerto Escondido. Kilómetro 1.5 Carretrea Puerto Escondido - Sola de Vega Carretera Puerto Escondido 71980, San Pedro Mixtepec, Oaxaca, México.*

* **Autor de correspondencia:** *fdelgado@zicatela.umar.mx* (JFDO)

Introducción

La música es una forma de expresión artística, de comunicación y de transmisión de emociones que ha empleado el ser humano desde sus orígenes con diversos propósitos, razón por la cual cada pieza musical es considerada como una composición única que está sujeta a derechos de autor. La música se estructura con base en esquemas de representación simbólicos denominados pentagramas, que son utilizados para presentar gráficamente las notas musicales, los ritmos, las intensidades y los factores de tiempo en que debe sostenerse el sonido asociado; su manifestación física se percibe como una presión sonora (acústica) en un fenómeno que suele ser objeto de estudio de la teoría musical.

Estos esquemas han sufrido algunas transformaciones o adaptaciones necesarias para el manejo de la música (como señales) en los diversos medios analógicos y digitales de producción y reproducción de audio. Huber & Runstein (2017) y Miyara (2015) describen una señal de audio como una señal eléctrica analógica de tensión o voltaje, que posee una forma de onda similar a la que presenta una señal sonora. La señal de audio está acotada en el rango de frecuencias de audio que son percibidas por el ser humano, las cuales se sitúan aproximadamente entre los 20 Hz y los 20,000 Hz, rango al que se le denomina espectro audible (Albariño & Balut 2019).

Los formatos analógicos de audio, primeros en surgir en el panorama de la historia de la música, emplearon representaciones continuas (funciones matemáticas) que permiten representar el espectro auditivo como la relación entre la amplitud y el periodo de una señal auditiva en términos de una función del tiempo, conocidos también como espectrogramas (Smith 2011). Esta representación fue adoptada también por los formatos digitales, debido a la estrecha relación que guardan ambos tipos de señales.

Con el auge de la era digital, detonada en las dos décadas recientes a causa de la producción masiva de reproductores multimedia digitales, la producción de música en los diversos formatos existentes se ha incrementado de manera significativa, lo que también se ha reflejado en la cantidad de información que es gestionada por los sistemas de reproducción, motivo por el que surge la necesidad de organizar los audios en términos de las características musicales que pueden ser identificadas en las representaciones de la señal como el ritmo, la tonalidad y la estructura melódica que define a su género, siendo este el elemento usual empleado en la música para organizar las obras musicales.

De acuerdo con Solano-Hernández (2019), esta clasificación obedece a diversos criterios como los medios sonoros y la función o los contenidos, conociéndose como género musical a la categoría que reúne composiciones musicales que comparten diferentes criterios, instrumentos y contextos sociales. De esta forma, el género representa una alternativa común para acceder a las piezas musicales, así como para reproducirlas de manera eficiente, razón que suma importancia a la extracción de las características auditivas en los archivos de música digitales.

La música en el procesamiento digital de señales

La transición de la música hacia los escenarios digitales ha sido soportada por el procesamiento digital de señales, donde cada documento o archivo digital de audio representa una señal digital de una dimensión, la cual es analizada mediante espectrogramas, los cuales, permiten

analizar con una mayor facilidad la estructura armónica de una señal de audio.

Un espectrograma se define en Smith (2011), para efectos del procesamiento digital de señales, como una representación gráfica de las intensidades de las magnitudes obtenidas a través de la STFT (Transformada de *Fourier* de Tiempo Corto , por sus siglas en inglés *Short-Time Fourier Transform*). La gráfica de las intensidades relacionadas suele desplegarse en una escala logarítmica, generalmente expresada en decibeles. En otras palabras, puede decirse que el espectrograma de una señal de audio es una representación visual de los componentes frecuenciales existentes en cada uno de los segmentos del audio analizado. Matemáticamente, la STFT puede expresarse en términos de la ecuación (1).

Donde x(n) es la señal de entrada en el tiempo n, w(n) es la función de ventana de longitud M, X_m (ω) es la DTFT (Transformada de Fourier de Tiempo Discreto, por sus siglas en inglés *Discrete-Time Fourier Transform*) de la señal centrada en el tiempo mR, y R es el tamaño del salto entre ventanas.

Existen numerosos trabajos en la literatura, tales como Costa *et al*. (2011), Kour (2015), Wyse (2017), Nirmal (2020), Aguilar-Sánchez (2023) y Zhang (2023) en los que se opta por utilizar una representación visual de la señal que permita conseguir una mejor extracción de las características del audio. Asimismo, otra técnica que es comúnmente empleada para el análisis de los componentes de las señales consiste en el escalamiento de las frecuencias del espectrograma a la escala de Mel (Zhang 2023). Esta transformación se logra mediante la aplicación del banco de filtros de Mel (ec. 2), los cuales se encargan de eliminar el ruido de la señal y los espacios en blanco (vacíos o interrupciones de la señal) que presentan los espectrogramas, ajustando la representación visual de la señal a la forma más parecida a la percepción de las frecuencias en el oído humano.

Ec. 2 $$\underline{X_{mel,i}} = M * |\underline{X_{stft,i}}|^2$$

Donde M se refiere a la matriz que representa el banco de filtros Mel, y el término cuadrático $|X_{stft,i}|^2$ equivale al espectro de potencia, y cada fila en M pertenece a un filtro Mel único. Del mismo modo, otra técnica empleada para mejorar la representación visual de los componentes frecuenciales es la de los Coeficientes Cepstrales de Frecuencia de Mel (MFCC por sus siglas en inglés *Mel Frequency Cepstral Coefficients*), los cuales son aplicados en trabajos como Kour (2015) y Aguilar-Sánchez (2023), mejorando el desempeño de los algoritmos de clasificación. El algoritmo de la figura 1 muestra el procesamiento de un audio que es filtrado mediante la FFT (Transformada Rápida de *Fourier*, por sus siglas en inglés *Fast Fourier Transform*), generando su espectrograma, y cuya salida es procesada por los filtros de Mel, transformando las frecuencias a una escala logarítmica, donde las señales son posteriormente deconvolucionadas mediante la transformada del coseno discreto, produciendo con ellos los MFCC resultantes.

Los filtros aplicados a la señal permiten extraer las características MFCC de los audios para su posterior

Ec. 1 $$X_m(\omega) \cdot = \sum_{n=-\infty}^{\infty} x(n)w(n-mR)e^{-j\omega n} = DTFT_{\omega}\left(x * SHIFT_{mR}(w)\right)$$

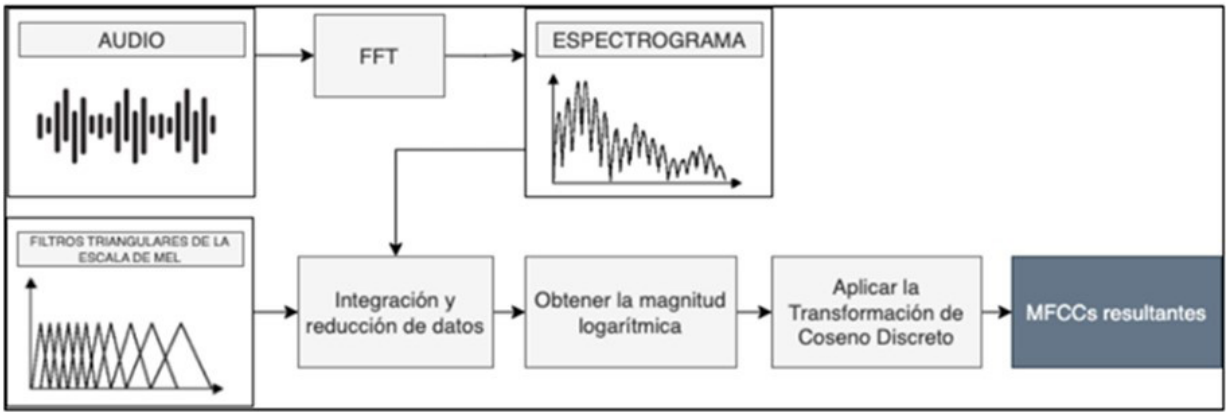

Figura 1. Etapas del algoritmo para la obtención de los MFCCs de un audio, figura adaptada de Aguilar-Sánchez (2023).

análisis en las tareas de aprendizaje automático (entrenamiento y prueba). Por ejemplo, para el entrenamiento de los clasificadores, propuestas como Wyse (2017) y Aguilar-Sánchez (2023) han utilizado métodos como las Redes Neuronales Convolucionales (CNN, por sus siglas en inglés *Convolutional Neural Networks*); en Zhang (2023) se emplean Redes Residuales (ResNet, acrónimo en inglés); en Tzanetakis & Cook (2002) se utiliza el algoritmo K-NN (K-Vecinos más Cercanos, por sus siglas en inglés *K-Nearest Neighbors*); y en Costa *et al*. (2011) y McFee *et al*. (2015) aplican SVM (Máquinas de Vectores de Soporte, por sus siglas en inglés *Support Vector Machines*). Al igual que en el entrenamiento, para la etapa de pruebas de clasificación de los géneros musicales se han utilizado clasificadores como K-NN (Tzanetakis & Cook 2002 y Aguilar-Sánchez 2023), SVM (Kour 2015, Nirmal & Mohan 2020, Tzanetakis & Cook 2002 y Aguilar-Sánchez 2023), y CNN (Wyse 2017, Aguilar-Sánchez 2023 y Zhang 2023). Lo anterior da lugar al planteamiento de una propuesta metodológica que permita la clasificación de audios mediante las técnicas del procesamiento digital de señales, las cuales permiten organizar una canción digitalizada en términos de su género musical.

Materiales y métodos

En la figura 2 se muestra la propuesta de solución para la clasificación de audios a partir de sus géneros musicales. El procesamiento consta de cinco etapas: la selección de un repositorio de audios digitales, el preprocesamiento, la extracción de características, el aprendizaje automático y la asignación del género para los audios.

En primer lugar, se seleccionan los archivos de audio del repositorio que sirven como entradas para los filtros STFT y MFCC, a partir de los cuales se obtienen sus espectrogramas. En segundo lugar, se extraen las características a las frecuencias de los espectrogramas, éstas sirven como datos de entrada para la fase de aprendizaje automático, donde tanto las características como los mecanismos del aprendizaje implementados alimentan una base de conocimiento. En tercer lugar, el aprendizaje automático utiliza la información de la base de conocimiento para estimar el género musical del audio analizado, con lo cual, un algoritmo de clasificación asigna un género específico a cada archivo de audio. Los procesos de las cinco fases de la metodología propuesta se detallan en las siguientes secciones.

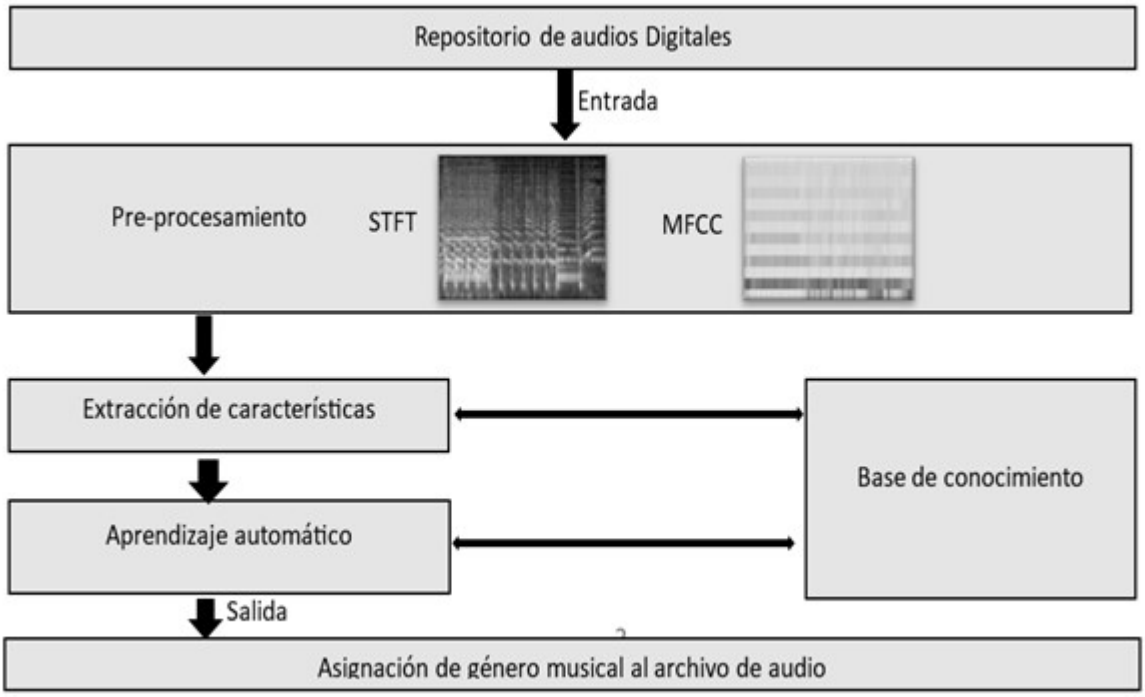

Figura 2. Metodología para la calificación de géneros musicales a partir de audios digitales.

Repositorio de audios digitales. Se utilizó la base de datos (GTZAN), que cuenta con más de 1,000 canciones clasificadas manualmente, cada una con una duración de 30 segundos. El repositorio asociado es utilizado por Tzanetakis & Cook (2002) en su estudio sobre la clasificación de géneros musicales. Esta base de datos ofrece una clasificación variada en términos de diez géneros musicales, lo que la convierte en un repositorio con las características deseables para llevar a cabo las pruebas de la metodología propuesta.

Pre-procesamiento. En esta etapa se realiza la conversión de la señal en el dominio de tiempo a espectrogramas mediante la STFT, para representar la señal en el dominio tiempo-frecuencia. Del mismo modo, para mejorar la representación visual, se aplican a los audios los filtros de Mel y los MFCCs, para obtener las características de la señal de audio como la textura tímbrica y la tonalidad de la melodía, así como otras variaciones que pueden ser observadas en las frecuencias existentes en la música. Para llevar a cabo las operaciones del procesamiento de los audios se utilizó la librería librosa (McFee *et al*. 2015) implementada en el lenguaje de programación *Python*.

La figura 3 muestra un ejemplo de los espectrogramas obtenidos a partir de las frecuencias de tiempo del filtro STFT sin ninguna alteración (Fig. 3a); la Figura 3b muestra la adaptación de dicho espectrograma a la escala de Mel, mediante la aplicación de los bancos de filtros de Mel; y finalmente, la Figura 3c muestra la representación frecuencial obtenida con los MFCC.

Extracción de características. Para llevar a cabo la extracción de características, así como los parámetros musicales, se calcula la media y la desviación estándar del conjunto de datos resultante de cada audio, como se realiza en Tepepa-Cantero *et al*. (2018). Para ello se utilizó NumPy (Harris *et al*. 2020), una librería de *Python* con la

cual se obtienen los vectores característicos para cada uno de los filtros (STFT, Mel y MFCC) los vectores característicos de los audios. Estos vectores se envían como entrada a la fase de aprendizaje automático, con la intención de identificar en ellos los patrones del género musical.

Aprendizaje automático. La fase de aprendizaje automático implementa los algoritmos de aprendizaje supervisado K-NN y SVM, dada su capacidad para manejar la complejidad inherente a las características extraídas. Para ello se emplean los algoritmos de la librería *Scikit-Learn* (Pedregosa *et al*. 2011), que contiene las rutinas y los algoritmos de aprendizaje automático implementados en el lenguaje de programación *Python*. Los algoritmos se aplican a una proporción de las representaciones de los espectrogramas de los archivos de audio (para entrenar los algoritmos), mientras que otra proporción es empleada con audios que no son utilizados en el entrenamiento, en la etapa de asignación del género musical.

Asignación de género musical del archivo de audio. Una vez que los algoritmos han sido entrenados, se pueden utilizar para asignar un género musical a un archivo de audio nuevo. Para esto, los modelos de clasificación que implementan los algoritmos evalúan las características extraídas del espectrograma y determinan el género musical al que mejor se ajusta, según los patrones aprendidos durante el entrenamiento. En esta etapa también se utilizó la librería *Scikit-Learn* (Pedregosa *et al*. 2011).

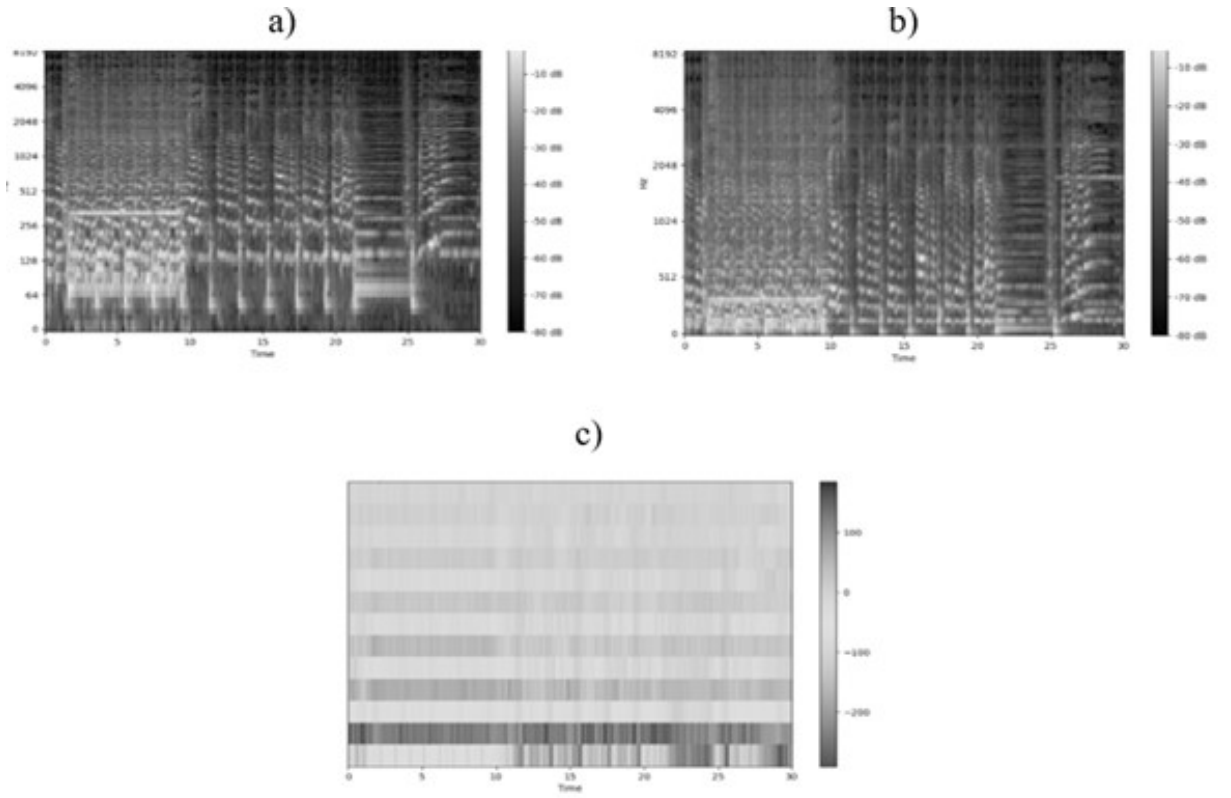

Figura 3. Espectrogramas generados a partir del filtro STFT en la etapa de pre-procesamiento. a) imagen resultante sin alteraciones, b) adaptación a la escala de Mel, c) representación de la frecuencia generada con los MFCC.

Resultados

Una muestra de 150 audios del repositorio como conjunto de entrenamiento fueron utilizados para probar la metodología. Estos se dividen en tres géneros musicales: rock, jazz y música clásica, con 50 audios para cada género. Del mismo modo, se utilizó un conjunto de pruebas compuesto por otros 150 audios, también distribuidos en los tres géneros con 50 audios para cada uno. En la tabla I se presenta un resumen de los resultados obtenidos en las pruebas unitarias de los clasificadores para cada uno de los filtros, de los cuales, se reporta el porcentaje de éxito promedio en el proceso de clasificación, lo anterior con base en las clases especificadas en los archivos de audio que se utilizaron.

Los resultados de la tabla I muestran variaciones significativas en el rendimiento de los algoritmos de clasificación, según la técnica de filtrado aplicada a la señal de audio, se obtuvieron los mejores resultados globales para el clasificador SVM con el filtro MFCC, con los cuales se alcanza un 82% de exactitud para el género clásico, 63% para jazz y 71% para rock.

Discusión

Los resultados de los clasificadores se sustentan en las distribuciones de las frecuencias de las características extraídas mediante el filtro MFCC. En ese sentido, la Figura 4 muestra los valores característicos para los audios cuyos resultados se presentaron en la Tabla I.

Las figuras 4a y 4b muestran las características extraídas con el filtro STFT, las figuras 4c y 4d muestran los vectores correspondientes para el filtro Mel y, por último, las figuras 4e y 4f visualizan los valores correspondientes para el filtro MFCC. La figura 4 muestra la correspondencia entre los resultados, destacando que las características de las figuras 4e y 4f correspondientes al filtro MFCC presentan una mejor distribución de las clases y un rango más amplio en el eje y. En contraste, los filtros de Mel representados en las Figuras 4c y 4d, exhiben un mayor solapamiento entre las clases de los distintos géneros y un rango más reducido. El filtro STFT (Figuras 4a y 4b) muestra un desempeño promedio para el entrenamiento, mientras que para los

Tabla I. Resultados de las pruebas de clasificación para los audios seleccionados del repositorio GTZAN.

Clasificador	Parámetros	Filtro	Exactitud %		
			Clásica	Jazz	Rock
K-NN	k = 40	STFT	72	40	36
		Mel	66	30	78
		MFCC	70	52	72
	k = 50	STFT	72	38	38
		Mel	66	30	78
		MFCC	70	48	46
SVM	kernel polinomial, C = 1, cef0 = 0	STFT	78	48	76
		Mel	58	36	80
		MFCC	77	51	67
	kernel polinomial, C = 1, cef0 = 6	STFT	65	53	73
		Mel	75	44	81
		MFCC	82	63	71

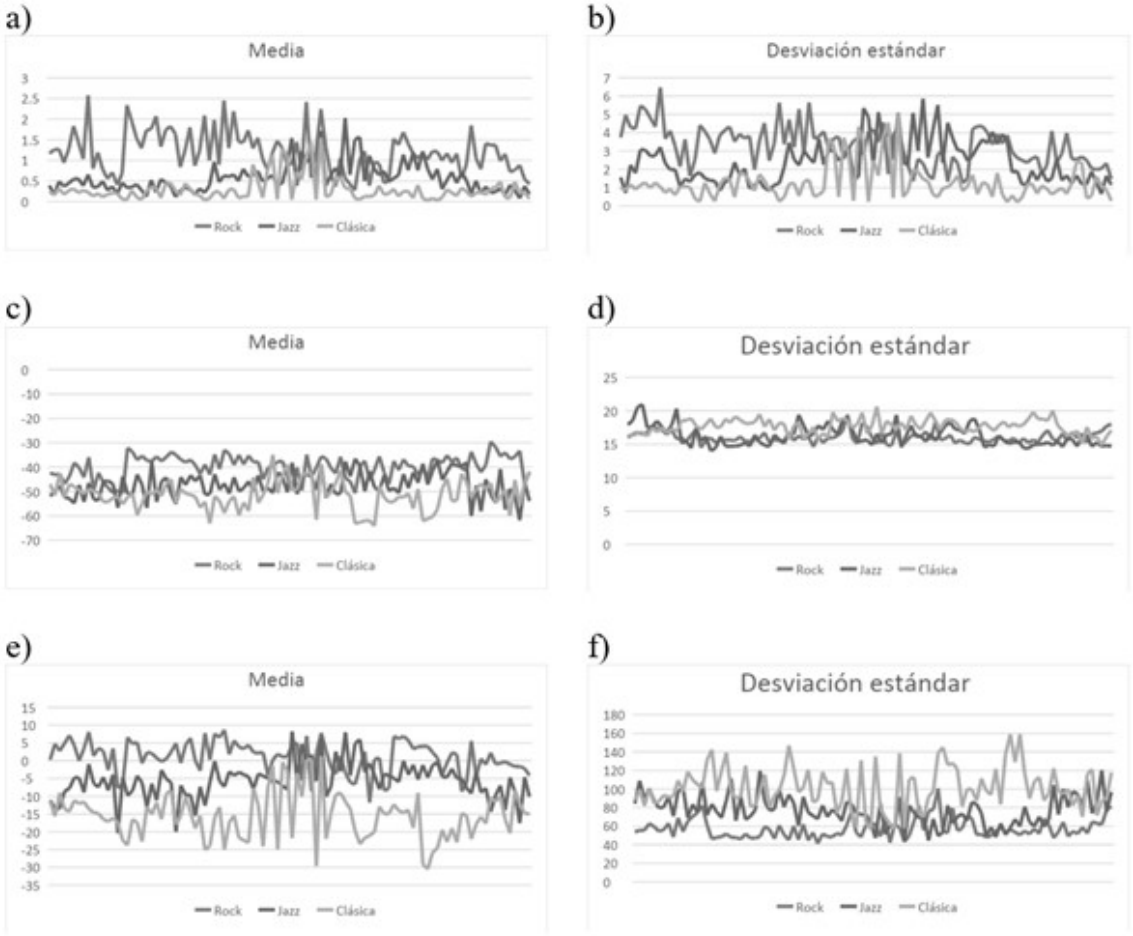

Figura 4. Valores característicos de las frecuencias de los audios seleccionados del repositorio GTZAN. a) media con STFT, b) desviación estándar con STFT, c) media con Mel, d) desviación estándar con Mel, e) media con MFCC, f) desviación estándar con MFCC.

ejemplos de prueba muestra que las clases se mantienen difusas. Del mismo modo, existe la posibilidad de que la omisión de otras características y de otros filtros con información relevante haya limitado la capacidad del modelo de aprendizaje de los clasificadores en la captura de los rasgos distintivos y la identificación de los patrones que permitan delimitar eficientemente las clases correspondientes a los géneros musicales.

Conclusiones

Con base en los resultados obtenidos, la técnica más efectiva fue el uso de coeficientes cepstrales de la frecuencia de Mel con el clasificador SVM, alcanzando una precisión del 81% para música clásica, 63% para jazz y 71% para rock; aunque cabe señalar que la utilización de segmentos de audio de 30 segundos pudo haber influido en la precisión de la clasificación. Es importante precisar que estos audios tienen ciertos segmentos en los que incluso el oído humano puede percibirlos como si se tratara de un género musical distinto al etiquetado. Por ello, el uso de segmentos más largos y de pistas con una mayor diferenciación entre géneros podría mejorar la precisión de la clasificación, aunque esto también aumentaría exponencialmente el tiempo de procesamiento requerido para el entrenamiento del modelo. Debido al desempeño favorable en las pruebas

realizadas, se recomienda el uso de MFCC con SVM. No obstante, se planea realizar pruebas adicionales con otros géneros para ampliar la validación en escenarios de mayor complejidad. Asimismo, se prevé explorar otras técnicas de extracción de características con el objetivo de evaluar y comparar su desempeño dentro de la metodología propuesta.

Agradecimientos

Agradecemos las observaciones y recomendaciones realizadas por los revisores para mejorar el contenido del presente documento.

Referencias

Aguilar-Sánchez, L.L. 2023. Análisis comparativo de las técnicas *deep learning perceptron*. Tesis de Licenciatura, Universidad de San Carlos de Guatemala, Guatemala. 98 pp.

Albariño, J. M., & Balut, P. 2019. La tecnología y el espectro audible en la producción musical. Pp. 45-53 *In*: Actas del III Congreso Internacional de Música y Cultura para la Educación, A Coruña, España.

Costa, Y., F. Gouyon & A. Koerich. 2011. Music genre recognition using spectrograms. Proceedings of the 11th International Conference on Intelligent Systems Design and Applications 11(1): 151-156.

Harris, C.R., K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk & M. Brett. 2020. Array programming with NumPy. Nature 585(7825): 357-362.

Huber, D.M., & R.E. Runstein. 2017. Modern Recording Techniques. 9a ed., Routledge, New York, USA. 616 pp.

Kour, G. 2015. Music genre classification using MFCC, SVM and BPNN. International Journal of Computer Applications 112(6): 12-14.

McFee, B., C. Raffel, D. Liang, D.P.W. Ellis, M. McVicar, E. Battenberg & O. Nieto. 2015. librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference 1(1): 18-25.

Miyara, F. 2015. Acústica y sistemas de sonido. 4a ed., UNR Editora, Rosario, Argentina. 255 pp.

Nirmal, M.R. & S.B. Mohan. 2020. Music genre classification using spectrograms. 2020 International Conference on Power, Instrumentation, Control and Computing (PICC) 1(1): 1-5.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Perrot & É. Duchesnay. 2011. Scikit-Learn: Machine learning in Python. Journal of Machine Learning Research 12(1): 2825-2830.

Smith III, J.O. 2011. Spectral audio signal processing. Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, USA. 352 pp.

Solano-Hernández, J. 2019. Identificación de géneros musicales. Tesis de Licenciatura, Universidad Nacional Autónoma de México, México. 124 pp.

Tepepa-Cantero, A., H.M. Pérez-Meana & M. Nakano-Miyatake. 2018. Algoritmos de aprendizaje supervisado para la clasificación de géneros musicales caracterizados mediante modelos estadísticos. Research in Computing Science 147(5): 119-128.

Tzanetakis, G. & P. Cook. 2002. Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing 10(5): 293-302.

Wyse, L. 2017. Audio spectrogram representations for processing with convolutional neural networks. Proceedings of the First International Workshop on Deep Learning and Music 1(1): 37-41.

Zhang, J. 2023. Music genre classification with ResNet and Bi-GRU using visual spectrograms. arXiv 1(1): 1-17.