Detección de Plagio de Audio mediante Correlación Visual Multi-Escala de Espectrogramas y Cromagramas

Enrique Andres Castelan Rosas
Facultad de Ingeniería
Universidad Veracruzana
Veracruz, Veracruz
kisabakasumi@gmail.com

***Resumen*—La detección automática de *sampling* no autorizado en la música moderna presenta desafíos significativos debido a las técnicas de producción actuales. La compresión dinámica agresiva y la saturación de frecuencias bajas en géneros como el Trap o Reggaeton a menudo enmascaran las características acústicas de las grabaciones originales utilizadas como base. Este proyecto aborda la problemática desde una perspectiva de visión artificial aplicada al dominio de la frecuencia. Hemos desarrollado un sistema que convierte señales de audio en representaciones visuales (Espectrogramas Mel y Cromagramas) y aplica técnicas de filtrado selectivo para aislar la información melódica relevante. El núcleo del sistema utiliza un algoritmo de coincidencia de plantillas (*Template Matching*) basado en correlación cruzada normalizada, implementado con un enfoque multi-escala para detectar coincidencias independientemente de las variaciones de *tempo* (velocidad). Los resultados demuestran la capacidad del sistema para localizar fragmentos reutilizados en mezclas complejas donde los métodos tradicionales de huella acústica suelen fallar, proporcionando evidencia visual y auditiva alineada para el análisis forense.**

***Index Terms*—Ingeniería de audio, análisis forense, visión por computador, cromagrama, espectrograma, correlación cruzada, procesamiento de señales.**

I. Introducción

El uso de fragmentos de grabaciones existentes (*sampling*) es una piedra angular de la producción musical contemporánea. Sin embargo, la identificación de la fuente original dentro de una nueva obra derivada se ha vuelto técnicamente compleja. Las herramientas convencionales de identificación musical, diseñadas para reconocer canciones completas e idénticas, a menudo fracasan cuando el fragmento ha sido alterado en tono, tiempo o está enterrado bajo nuevas capas de instrumentación y procesamiento de audio moderno.

El desafío principal radica en la diferencia tímbrica y de mezcla entre grabaciones de distintas décadas. Una línea de guitarra de los años 90 suena drásticamente diferente cuando se incorpora sobre una base de percusión digital actual con graves saturados.

Este trabajo propone una solución ingenieril que se aleja del análisis puramente acústico y trata el problema como uno de reconocimiento de patrones visuales. Al transformar el audio en mapas de tiempo-frecuencia, podemos aplicar algoritmos maduros de visión por computador para encontrar ïmágenes"(patrones musicales) dentro de otras ïmágenes"(canciones completas), superando las barreras de la producción de audio.

II. Objetivos del Proyecto

II-A. Objetivo General

Diseñar e implementar una herramienta de software para el análisis forense de audio capaz de detectar y localizar fragmentos musicales reutilizados entre dos pistas, utilizando técnicas de procesamiento de imágenes sobre representaciones espectrales.

II-B. Objetivos Específicos

- Implementar una etapa de preprocesamiento de señal (DSP) que incluya filtrado pasa-banda para eliminar el ruido de frecuencias extremas (sub-graves y agudos excesivos) que no aportan información melódica.
- Generar representaciones visuales robustas del audio, específicamente Espectrogramas Mel para textura y Cromagramas para contenido armónico (notas musicales).
- Desarrollar un motor de búsqueda visual basado en áreas (Correlación Cruzada Normalizada) que sea capaz de manejar variaciones de escala temporal (cambios de BPM).
- Automatizar la generación de reportes de evidencia, incluyendo la localización temporal exacta del hallazgo y la extracción de clips de audio sincronizados para su verificación humana.

III. Estado del Arte

Las técnicas tradicionales de recuperación de información musical (MIR) se han basado predominantemente en el "fingerprinting"de audio. Estos sistemas generan hashes compactos basados en los picos de energía más prominentes de un espectrograma. Si bien son extremadamente eficientes para identificar grabaciones idénticas (uso típico de aplicaciones comerciales de reconocimiento musical), su robustez disminuye considerablemente cuando el audio sufre transformaciones lineales de tiempo (*time-stretching*) o adición significativa de ruido de fondo.

Investigaciones más recientes han explorado el uso de la información de croma (Chromagramas). Un cromagrama proyecta la energía espectral en las 12 notas de la escala cromática, independientemente de la octava. Esto resulta en una representación que captura la progresión armónica y melódica de una pieza, siendo altamente invariante a los cambios de instrumentación o timbre. Nuestro enfoque se basa en esta línea, utilizando los cromagramas no solo para análisis estadístico, sino como imágenes directas para algoritmos de coincidencia de patrones visuales.

IV. Diseño e Implementación

El sistema se construyó utilizando Python, aprovechando las bibliotecas `librosa` para el procesamiento de audio y `OpenCV` para las tareas de visión artificial. La arquitectura se divide en tres etapas principales.

IV-A. Etapa 1: Preprocesamiento de Señal y Generación de Imágenes

La calidad de la detección visual depende directamente de la limpieza de la señal de entrada.

1. **Carga y Normalización:** El audio se carga a una tasa de muestreo de 22.050 Hz, suficiente para capturar la información musical relevante optimizando el rendimiento computacional.
2. **Filtrado Espectral:** Se aplica un filtro pasa-banda digital, descartando frecuencias por debajo de 300 Hz y por encima de 8000 Hz. Esta decisión de diseño es crítica para eliminar la interferencia de líneas de bajo sintetizadas modernas que saturan el espectro y ocultan las melodías de rango medio de muestras más antiguas.
3. **Transformación Visual:** Se generan dos mapas:
   - *Espectrograma Mel:* Captura la textura tímbrica general en una escala perceptual.
   - *Cromagrama (CQT):* Captura la estructura de notas musicales, esencial para identificar melodías transportadas a otros instrumentos.
4. **Mejora de Contraste:** Las matrices resultantes se normalizan a imágenes de 8 bits y se les aplica Ecualización de Histograma Adaptativo (CLAHE) para resaltar los patrones armónicos débiles.

IV-B. Etapa 2: Motor de Búsqueda Visual Multi-Escala

El núcleo del sistema no busca puntos aislados, sino que compara regiones completas de la imagen. Utilizamos la Correlación Cruzada Normalizada (`TM_CCOEFF_NORMED`) como métrica de similitud.

El algoritmo toma un segmento representativo de la imagen de la canción original (el "patrón") y lo desliza sobre la imagen de la canción sospechosa. Para abordar el problema de los cambios de tempo, implementamos un enfoque de fuerza bruta controlado: el patrón original se redimensiona iterativamente en el eje temporal (ancho de la imagen) en un rango de factores de escala de 0.8x a 1.2x. Se calcula el mapa de correlación para cada escala y se selecciona el pico máximo global como la mejor coincidencia candidata.

IV-C. Etapa 3: Validación y Localización

Una vez identificado el punto de máxima correlación visual, el sistema verifica la coherencia de la textura en esa región utilizando Histogramas de Gradientes Orientados (HOG) como métrica secundaria. Finalmente, la coordenada del píxel coincidente se mapea de nuevo al dominio del tiempo (segundos) para determinar el punto de inicio exacto del sample en la pista sospechosa.

V. Resultados

Las pruebas se realizaron con pares de canciones conocidas por contener samples con modificaciones significativas de producción y tempo. El sistema demostró ser capaz de localizar correctamente los fragmentos reutilizados.

El filtrado pasa-banda demostró ser un componente crucial; sin él, la energía de los instrumentos de percusión modernos dominaba la correlación visual, resultando en falsos negativos. Con el filtrado activo y el escaneo multi-escala, el sistema logró identificar progresiones de notas idénticas a pesar de las diferencias tímbricas, devolviendo puntuaciones de confianza (correlación) consistentes y calculando con precisión el factor de modificación de tempo aplicado por el productor.

La herramienta genera con éxito mapas de calor visuales que resaltan la región del plagio y extrae clips de audio sincronizados, facilitando la verificación manual por parte de un perito humano.

VI. Conclusiones y Trabajo Futuro

VI-A. Conclusiones

El enfoque implementado, basado en el procesamiento de imágenes de área (correlación de plantillas) sobre cromagramas filtrados, ofrece una solución viable y robusta para la detección de plagio en escenarios de producción musical moderna. La abstracción visual del audio permite superar las limitaciones de los comparadores acústicos tradicionales cuando se enfrentan a mezclas densas o cambios de instrumentación. La implementación de una búsqueda multi-escala es indispensable para manejar las variaciones de tempo inherentes al uso creativo de samples.

VI-B. Trabajo Futuro

El desarrollo futuro se centrará en integrar técnicas de separación de fuentes de audio (como U-Nets para demixing) como paso previo al análisis visual. Esto permitiría aislar, por ejemplo, una pista de guitarra de una mezcla completa antes de compararla, aumentando teóricamente la precisión en casos de solapamiento extremo de instrumentos. Además, se explorará la optimización del código mediante computación paralela en GPU para reducir los tiempos de análisis en archivos de larga duración.

Referencias

[1] M. Müller, "Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications,"Springer, 2015.

[2] J. P. Bello et al., .A Tutorial on Onset Detection in Music Signals,ÏEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 1035-1047, Sept. 2005.

[3] R. Szeliski, Çomputer Vision: Algorithms and Applications,”2nd ed., Springer, 2022.