Tecnologías para el análisis del contenido musical de archivos sonoros y para la generación de nuevos metadatos

Perfecto Herrera Boyer y Emilia Gómez Gutiérrez

Music Technology Group, Departamento de Tecnologías de la Información y las Comunicaciones, Universitat Pompeu Fabra y Departamento de Sonología, Escola Superior de Música de Catalunya

Introducción

Hace un par de décadas el conocer, organizar y explotar la información musical contenida en un archivo sonoro solo era posible a partir de un etiquetado manual que requería la intervención humana para anotar, de manera estructurada, dichos contenidos. La necesidad de acceder a dicha información proviene de diferentes colectivos de usuarios, con criterios, conceptos, expectativas e intereses muy variados: musicólogos, bibliotecarios, compositores, intérpretes, aficionados a la música, etc. Desde finales de los años 90 del siglo pasado se ha ido definiendo una comunidad de investigadores, desarrolladores de software y empresas cuyos objetivos incluyen el hacer posible la automatización de los procesos de descripción y explotación de tales contenidos.[1] La evolución de esta comunidad ha ido paralela a la tendencia a disponer de acceso a grandes colecciones de archivos sonoros musicales, que décadas atrás eran solamente patrimonio de instituciones culturales y educativas, o de empresas comerciales.

Contenido musical y metadatos

El contenido musical se puede definir como toda información referente a una obra musical que está representada en ella misma. Aquí distinguimos entre la señal sonora (definida como una serie de valores que representan las oscilaciones de amplitud de la onda sonora a lo largo del tiempo) y aquello que se puede decir al respecto de esa señal una vez escuchada por un oyente. El contenido es una representación de la obra, que se expresa mediante una serie de descriptores o características de la pieza. Incluye diferentes aspectos como la estructura, el ritmo, la melodía, la armonía o la instrumentación. También incluye conceptos que van más allá de aquellos propios de la teoría musical como por ejemplo la emoción que tiende a transmitir, la energía o el género musical.

Figura 1. Contenido extraíble y contenido manualmente etiquetable.

La figura 1 ilustra el proceso necesario para abstraer o inferir el contenido a partir de la señal, y muestra cómo dicho contenido es, en gran parte, etiquetable de forma manual, aunque una parte de ese etiquetado puede hacerse de manera automatizada. Los descriptores manuales constituyen la información que no está codificada en la pieza en sí: por ejemplo el año de composición, el lugar de grabación o el nombre del autor. Algunos de estos datos, a pesar de todo, se pueden intentar extraer con herramientas computacionales.

El objetivo de las tecnologías de procesado del contenido musical es el de describir automáticamente el contenido de señales de audio musicales y proporcionar medios para editarlo, visualizarlo y explotarlo con fines diversos (búsqueda, navegación, transformación, recomendación, etc.) Hay numerosas disciplinas involucradas en esta tarea: procesado de señal, musicología, psicoacústica y cognición musical, informática, ciencias de la información y de la computación, etc.

Descripción y transcripción

Las tecnologías actuales no permiten automatizar la transcripción a notación musical de un archivo sonoro musical polifónico, lo que nos permitiría

contar con una partitura digital de su contenido musical, aunque ello sí que es posible en ciertas circunstancias si el material es monofónico.[2] Por otra parte, convertir un archivo sonoro musical en partitura no garantizaría la extracción de todo el contenido albergado en él, dado que la notación musical solo caracteriza un subconjunto suyo. Si, además, consideramos que muchos documentos sonoros musicales pueden pertenecer a culturas musicales cuyas convenciones son diferentes de las occidentales, convendremos que, además de tecnologías de transcripción, necesitamos tecnologías de descripción, que buscan integrar metadatos o informaciones relacionadas con el contenido musical con la manera y el contexto en el que dicho contenido se puede recibir, categorizar o experimentar por parte de un oyente.

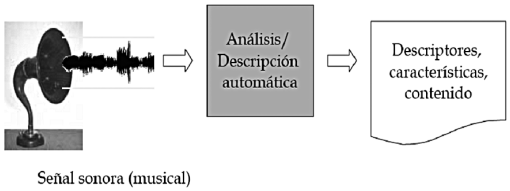

Figura 2. Elementos característicos de la descripción automática del contenido musical de un archivo sonoro.

Denominamos descriptores a todos aquellos datos que pueden constituir un predicado del contenido del archivo sonoro con el fin de caracterizarlo (por ejemplo, tener un tempo de 100 pulsaciones por minuto, una estructura en forma de sonata, contener sonidos de trompeta, haber sido compuesto en el siglo XIX, tener un centroide espectral promedio de 1000 Hercios, transmitir una sensación de tristeza, ser un ejemplo de *free jazz*, o presentar una tonalidad de sol mayor). Algunos de dichos descriptores pueden calcularse automáticamente a partir del análisis del archivo sonoro (y aceptando siempre una probabilidad de error nada despreciable)[3], mientras que otros requieren de la anotación manual realizada por un oyente experto (por ejemplo, difícilmente podríamos distinguir automáticamente si una grabación corresponde al año 1970 o 1975, si dicha información no aparece en los metadatos textuales de la grabación; solo un humano podrá, bajo ciertas circunstancias, realizar el trabajo requerido para su correcta datación).

Distinguimos entre descriptores de bajo, medio y alto nivel para diferenciar entre aquellos cuyo grado de abstracción se halla lejano a los conceptos que un oyente o un experto musical podría utilizar (bajo nivel) o aquellos que corresponden a las representaciones mentales comúnmente utilizadas por los oyentes o expertos musicales (alto nivel). A los descriptores de alto nivel los denominamos también descriptores semánticos y son los más difíciles de extraer automáticamente y, por ello, los más susceptibles a errores. Además, podemos definir descriptores relativos a diferentes niveles temporales. Por ejemplo, algunos descriptores se denominan «instantáneos» al referirse a un punto concreto de la grabación. Por el contrario, los descriptores «globales» hacen referencia a la obra completa o a un fragmento amplio de ella.

Buena parte de los descriptores se calculan a partir de realizar representaciones de la forma de la onda sonora basadas en el teorema de Fourier. A partir de operaciones como la denominada transformada de Fourier a corto plazo, es posible describir a bajo nivel el contenido frecuencial de un segmento sonoro muy corto. Cuando a este contenido le aplicamos, además, un modelo de enmascaramiento (algo que incluyen los codificadores de MP3) la descripción resultante guarda ciertas similitudes con la que la cóclea humana transmite al cerebro por vía del nervio auditivo. A partir de dicha descripción, y a menudo combinando una secuencia de dichas micro-descripciones, podemos obtener descriptores de más alto nivel como por ejemplo notas musicales o segmentos estructuralmente relevantes. En algunos casos los sistemas incorporan conocimiento musical y sonoro mediante modelos semánticos que capturan las similitudes y diferencias entre conceptos (por ejemplo, qué tienen en común los sonidos de violín, o qué tienen en común las piezas de jazz y qué las hacen diferentes de las de folclore). Para elaborar dichos modelos se utilizan colecciones previamente anotadas por humanos, que se suministran como ejemplos para que un sistema de aprendizaje automático (*machine learning*) abstraiga los rasgos relevantes para caracterizar un concepto así como los rasgos que permiten discriminarlo frente a otros conceptos relacionados.

STRUCT		CONCEPT LEVEL		MUSICAL CONTENT FEATURES				
CONTEXTUAL	global beyond 3 sec	HIGH II	EXPRESSIVE	cognition | emotion | affect = *syntactic+semantic concepts*				
				melody	harmony	rhythm	source	dynamics
		HIGH I	FORMAL	key profile	tonality cadence	rhythmic patterns tempo	instrument voice	trajectory articulation
	global < 3 sec	MID	PERCEPTUAL	successive intervallic pattern	simultane intervallic pattern	beat IOI	spectral envelope	dynamic range sound level
				pitch		time	timbre	loudness
NON-CONTEXTUAL	local + spatial	LOW II	SENSORIAL	periodicity pitch pitch deviations fundamental frequency		note duration onset offset	roughness spectral flux spectral centroid	neural energy peak
	local + temporal	LOW I	PHYSICAL					
				frequency		duration	spectrum	intensity

Figura 3. Tabla resumen de niveles de descripción y facetas musicales (Lesaffre y otros, 2003).

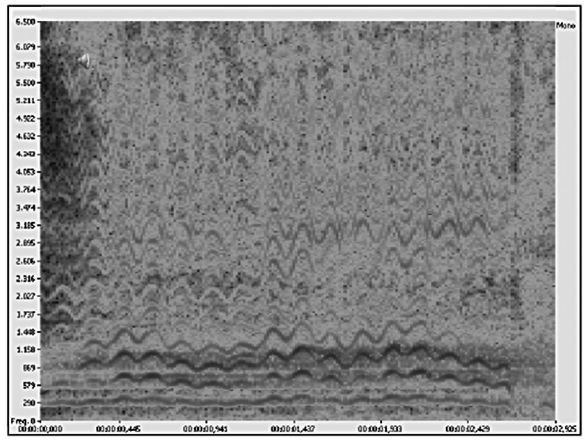

Figura 4. Espectrograma de un fragmento de canto. Horizontalmente representamos tiempo y verticalmente representamos las frecuencias del espectro. El tono oscuro de las líneas indica la cantidad de energía para cada armónico del espectro. La fundamental es la línea oscura más inferior, y los armónicos más prominentes aparecen por encima. Nótese también cómo estos no son líneas planas sino que presentan pequeñas oscilaciones correspondientes al vibrato impreso al cantar las notas.

Facetas susceptibles de descripción automatizada

Durante los últimos años se han hecho esfuerzos considerables para obtener automáticamente descriptores de las diferentes facetas musicales: timbre, ritmo, melodía, armonía/tonalidad y estructura. Como se ha comentado anteriormente, aquello que representan no siempre equivale a los conceptos más utilizados en teoría musical, y la precisión de las técnicas existentes es limitada, pero aún así ofrecen una ayuda computacional a la catalogación manual. Especialmente relevante es la posibilidad de procesar masivamente colecciones sonoras de miles o de millones de archivos.

Timbre

Las técnicas de descripción de timbre o instrumentación se basan en analizar la señal sonora y extraer su representación frecuencial o espectro, que representa la intensidad relativa de cada frecuencia audible. Dicha representación se calcula para un período de tiempo, y por lo tanto podemos estudiar la evolución temporal del espectro, denominada sonograma o espectrograma. El espectro nos indica por tanto qué componentes de frecuencias hay en un sonido en cada instante, que están relacionadas con su altura (o frecuencia fundamental) y su timbre.

Las técnicas de análisis computacional de timbre se basan, por tanto, en el análisis de forma de onda y espectro para describir las cualidades del sonido y estimar, en la medida de lo posible, qué instrumento(s) está(n) representado(s). Actualmente podemos llegar a etiquetar automáticamente fragmentos o archivos musicales polifónicos con los instrumentos predominantes, aunque también, sin necesidad de etiquetaje, podemos identificar fragmentos tímbricamente similares en diferentes archivos o dentro de uno mismo (véase la figura 5). El etiquetado automático requiere la elaboración, con ayuda de técnicas de aprendizaje automático, de «modelos tímbricos» correspondientes a los instrumentos que se desea identificar.

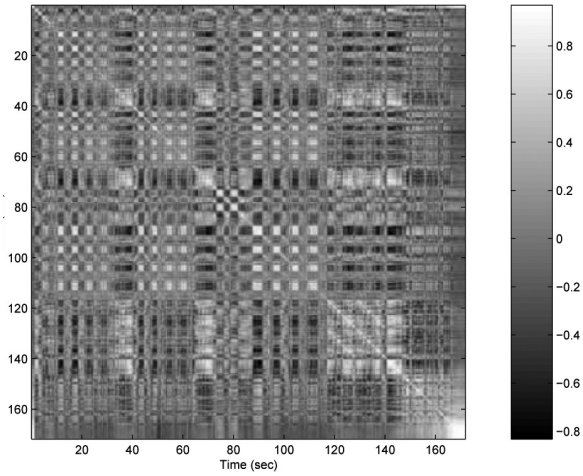

Figura 5. Matriz de similitud tímbrica de «Magical Mystery Tour». Los dos ejes representan tiempo y el brillo representa el grado de similitud de cada instante de análisis con el resto de instantes de la canción. La típica estructura repetitiva y seccional de una canción pop puede visualizarse claramente (Cooper y Foote, 2002).

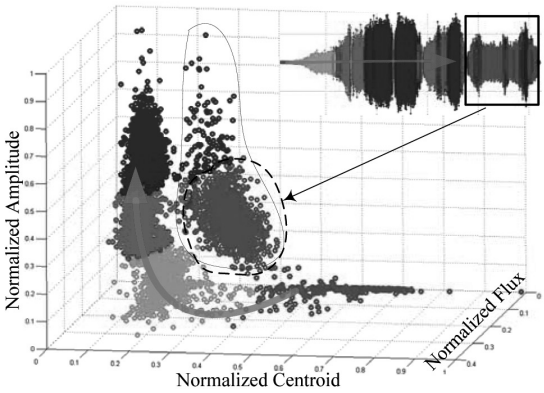

Figura 6. Ejemplo de visualización y segmentación basadas en información de timbre. La energía (eje y), el brillo o centroide espectral (eje x), y el flujo o variación espectral a corto plazo se utilizan para diferenciar secciones de la composición *Riverrun* de Barry Truax (Park y otros, 2009).

Melodía

La melodía es una faceta muy importante para caracterizar una pieza y, por ejemplo, reproducirla mediante el tarareo. La descripción melódica está íntimamente relacionada con la transcripción, aunque proporciona una descripción más amplia y relacionada con la interpretación o las características melódicas. Aunque los humanos tenemos una habilidad especial para descifrar la melodía predominante en un fragmento musical polifónico, los sistemas artificiales aún no pueden hacer lo propio con un grado de fiabilidad aceptable. Por lo tanto, el ámbito más adecuado para trabajar con descripciones melódicas es el de los instrumentos monofónicos o las polifonías simples. Los pasos para una transcripción melódica se representan en la figura 7.

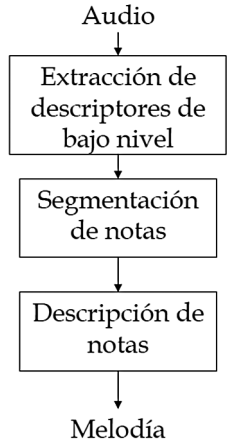

Figura 7. Pasos para una transcripción melódica.

El primer paso es el de extraer descriptores de bajo nivel relevantes, principalmente la frecuencia fundamental (relacionada con la altura), y energía (relacionada con la intensidad). Se han propuesto numerosos métodos para la estimación de la frecuencia fundamental de una señal sonora, que se relaciona con su altura. A pesar de ello, todavía no es un problema resuelto para todos los instrumentos y condiciones, incluso en sonidos monofónicos. Por lo que respecta a la energía, esta se extrae directamente de la señal de audio, sin error, y en algunos casos se calcula independientemente para diferentes bandas del espectro de frecuencias audibles. Ello permite, por ejemplo, detectar los ataques de los instrumentos en diferentes tesituras. La figura 8 muestra un ejemplo de extracción de curva de altura para un fragmento de saxofón, donde cada granja de la gráfica del centro, en la que aparece la frecuencia fundamental detectada, se corresponde con un semitono de la escala temperada.

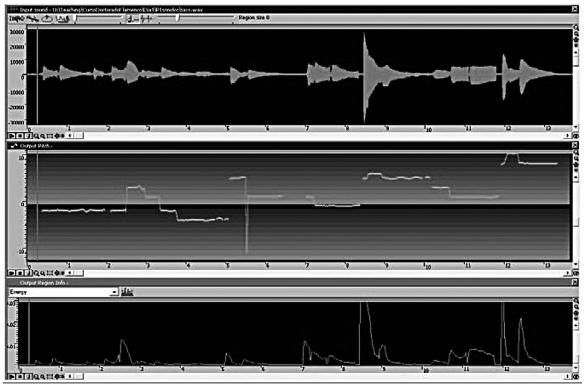

Figura 8. Señal de audio (arriba), frecuencia fundamental (centro) y energía (abajo) extraídas de un fragmento de saxofón.

En segundo lugar la segmentación de notas se basa en detectar variaciones de los descriptores calculados anteriormente, teniendo en cuenta las variaciones que no son debidas a cambio de nota (vibrato, trémolo) y eliminarlas. La segmentación de notas es un problema complejo y aún en vías de investigación. Finalmente, cada nota se describe en términos de altura, duración y energía.

A partir de la transcripción melódica podemos calcular otros descriptores que nos servirán para caracterizarla, como la tesitura, la distribución de notas e intervalos y los patrones melódicos utilizados. La descripción melódica automatizada tiene diversas aplicaciones, como la búsqueda de melodías por tarareo (*query by humming*), el análisis expresivo, el análisis comparativo de estilos o la transcripción en sí misma.

Tonalidad

La descripción tonal aplica los conceptos de teoría musical a descriptores obtenidos a partir de la señal sonora. En este sentido utilizamos el término tonalidad para referirnos al sistema de relaciones entre una serie de notas, que forman armonías y melodías, y que tienen una tónica o nota central como elemento más importante. En este sentido, los métodos computacionales de descripción tonal caracterizan las relaciones entre las notas de una pieza musical. Además, la descripción tonal se relaciona con el estilo y el carácter de la obra y está en la base de su caracterización emocional. Por otra parte, posibilita encontrar relaciones y similitudes entre diferentes piezas, y es una información crucial para identificar versiones de una determinada composición.

Cuando trabajamos con grabaciones, el primer paso es extraer información sobre las notas presentes en un fragmento sonoro, para lo cual, y dada la complejidad de la transcripción automática que hemos mencionado anteriormente, se utilizan los descriptores de bajo nivel denominados chroma. Los descriptores de chroma, aunque no proporcionan una transcripción automática precisa, aproximan el valor de la intensidad relativa de cada semitono de la escala temperada. Podemos considerarlos como descriptores que aproximan los acordes de un fragmento musical. Dichos descriptores de chroma se comparan con los perfiles tonales mayores y menores, obtenidos a partir de experimentos con oyentes o derivados de la teoría musical. Así podremos determinar de qué acorde (descriptor instantáneo) o tonalidad (descriptor de un segmento sonoro mayor) se trata.

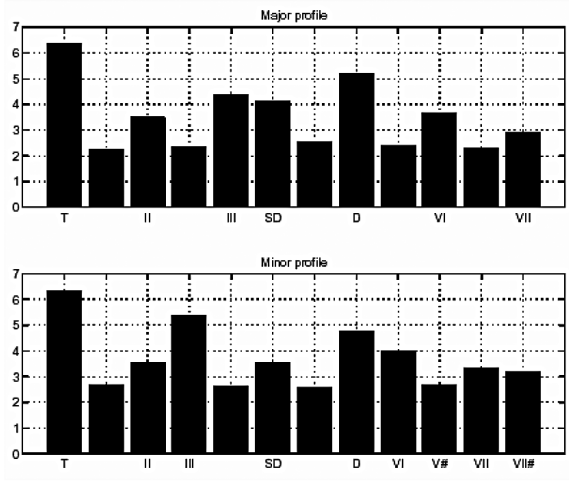

Figura 9. Perfiles de tonalidad mayor y menor que representan, estadísticamente, el peso relativo de cada uno de los grados de la escala según experimentos con oyentes habituados a música occidental. Estas plantillas normativas se utilizan para asignar la tonalidad a un fragmento musical según la similitud entre su chroma y dichos perfiles.

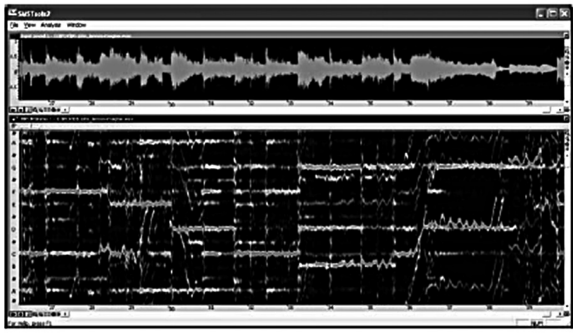

Figura 10. Ejemplo de visualización de chroma de un fragmento musical. Arriba: forma de onda. Abajo: el eje horizontal representa el tiempo y el vertical las notas presentes en el fragmento (líneas más claras) superpuestas en una sola octava.

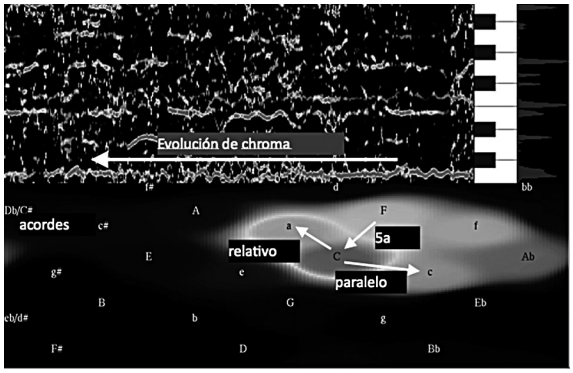

Figura 11. Sistema de visualización de chroma y acordes en el espacio denominado *harmonic network*. Los acordes mayores se representan en mayúscula y los menores en minúscula.

Ritmo

La descripción rítmica computacional se relaciona con aspectos temporales de la música, con la organización de los eventos musicales en el tiempo. A pequeña escala, el ritmo se describe a partir de las duraciones de las notas y de los silencios entre ellas, y a gran escala describimos el tempo, compás o patrón rítmico. Los métodos computacionales de descripción rítmica de sonido se derivan de métodos de descripción de partituras. Las áreas de trabajo fundamentales son la detección de pulsos (*beat*), la detección de tempo (*beats per minute),* la extracción de patrones rítmicos (métrica binaria o ternaria) y la descripción de características expresivas (desviación respecto a una partitura o notación inexpresiva).

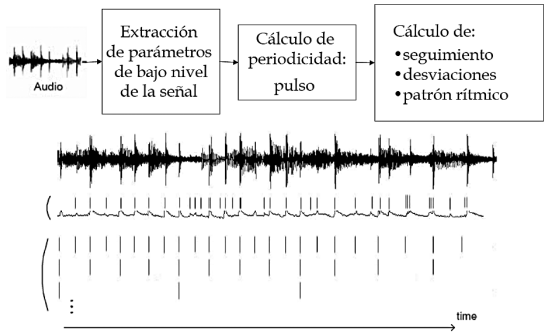

Figura 12. Esquema general de un sistema computacional de descripción rítmica.

El primer paso consiste en la extracción, a partir de la señal sonora, de parámetros relevantes, como pueden ser la evolución de la energía, tanto de manera global como distribuida en las diversas bandas del espectro de frecuencias. A continuación, se analiza la repetición o periodicidad de dicha energía para determinar el pulso o los pulsos existentes, y se estudia su variación y desviación a lo largo del tiempo y los patrones que forman los diferentes pulsos coexistentes (métrica). Las áreas de investigación se centran en algoritmos de detección de eventos (*onsets*) a partir de los parámetros de señal, algoritmos de cálculo de periodicidad y métodos de descripción métrica. Los métodos actuales funcionan relativamente bien con música muy rítmica y cuya métrica es estable en toda la pieza, pero en estilos con tempo o métricas variables aún no resultan suficientemente fiables.

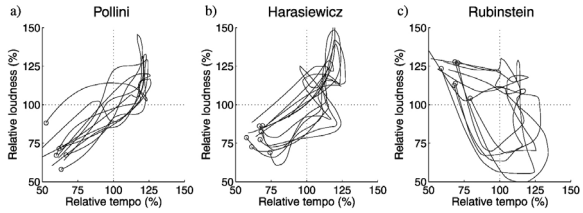

Figura 13. Comparación de interpretaciones de piezas de piano tocadas por diferentes intérpretes. Se representa el tempo respecto a la intensidad, extraídos automáticamente (Goebl y otros, 2004).

Estructura

La descripción estructural automatizada se centra en dos problemas fundamentales: la segmentación automática de grabaciones musicales y la búsqueda de fragmentos repetidos dentro de una pieza. Para ello se tienen en cuenta parámetros relacionados con la intensidad (energía), armonía (acordes), melodía (notas) y ritmo (eventos), y se buscan tanto sus cambios abruptos como sus pautas de recurrencia más o menos periódica. Una vez disponemos de delimitadores estructurales podemos, por ejemplo, realizar una navegación rápida a lo largo del archivo, o comparar segmentos de la misma clase dentro de una colección archivos.

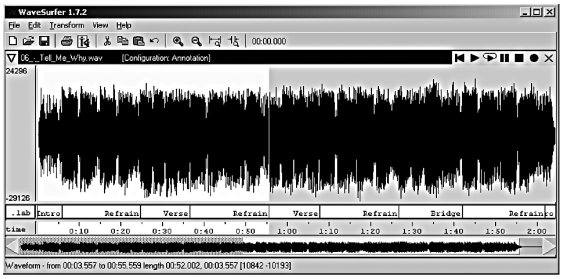

Figura 14. Segmentación de una canción en sus secciones más relevantes.

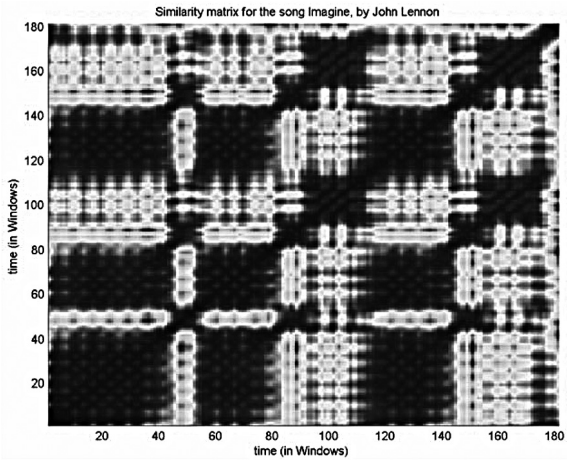

Figura 15. Matriz de similitud que representa repeticiones de acordes en la canción «Imagine» de John Lennon. La estructura de la canción puede determinarse a partir de esta información.

Descriptores semánticos

Existen diferentes categorías conceptuales que podemos aplicar para describir o caracterizar un fragmento musical incluso sin ser expertos en materias musicales. Cualquier oyente es capaz de asignar descriptores de género o emocionales a un fragmento musical. Para realizar dicha caracterización los oyentes combinamos (a menudo automáticamente, sin ser conscientes de ello) información de timbre, tonal, melódica, rítmica y estructural que nuestro cerebro extrae al procesar estímulos musicales. La exposición a determinados géneros o contextos emocionales permite que, de manera automática, formemos asociaciones o inferencias entre características musicales y dichas categorías. Nuestros sistemas de descripción automática siguen ese mismo proceso. Así, disponemos de sistemas que pueden asignar etiquetas semánticas a un archivo sonoro con una alta fiabilidad y consistencia con los juicios que la mayoría de oyentes podría realizar. Nuevamente en este problema es esencial disponer de una buena colección de ejemplos para cada categoría que el sistema tiene que asignar, ya que un componente del sistema requiere de técnicas de aprendizaje automático que detectan relaciones entre los descriptores musicales y las diferentes categorías a aprender. Los sistemas actuales posibilitan la asignación de etiquetas emocionales tales como alegre, triste, agresiva o relajada, o de grandes géneros como pop, folclore, rock, tecno, clásica o jazz. Otras posibles etiquetas semánticas tienen que ver con la energía de la música (por ejemplo, etérea, tranquila o fuerte) o con el tipo de instrumentación (acústica, electrónica). En los casos en los que las etiquetas pueden tener una cierta variabilidad interpersonal es posible desarrollar sistemas personalizados que aprenden, para cada usuario, su manera específica de categorizar la música.

Presentación de la información

Frente a la tradicional búsqueda a partir de texto libre o de palabras clave, y a la presentación de listas de resultados, las tecnologías que hemos descrito en este artículo permiten plantear otras formas, menos centradas en la manipulación de texto, de buscar y de presentar información musical.

A la hora de buscar información se han desarrollado diferentes métodos que explotan la descripción del contenido musical:

- *Query by humming/singing*: el usuario tararea o canta una melodía con el fin de hallar música que contenga melodías similares;
- *Query by tapping:* el usuario pulsa rítmicamente una tecla o golpea una superficie sensible y esa secuencia se usa para hacer una búsqueda basada en el ritmo;
- *Query by playing:* el sistema proporciona algún tipo de interfaz musical (por ejemplo un teclado virtual) para que el usuario introduzca su petición de búsqueda;
- *Query by example:* se trata de proporcionar al sistema un archivo sonoro con un fragmento musical que hace de modelo para buscar en la colección otros fragmentos similares. La similitud puede estar predefinida o bien ajustarse a criterios que el propio usuario establece (por ejemplo, no considerar el tempo, o utilizar solo información tonal para la búsqueda).

Además de la búsqueda de información, la visualización de colecciones, subcolecciones o resultados de una búsqueda constituye otro tema importante para el que se han propuesto diversas innovaciones. Las colecciones se pueden visualizar como mapas geográficos en dos o tres dimensiones que pueden corresponder a características semánticas (por ejemplo, tempo, o género) o a rasgos de bajo nivel. Los denominados mapas auto-organizativos (*self-organizing maps* o SOM) han sido utilizados con frecuencia para organizar colecciones musicales y categorías aplicables a ellas en dos dimensiones. *Islands of Music* es el ejemplo más célebre de explotación de dicha técnica. En la figura 16 los géneros aparecen organizados de manera automática a partir del cálculo de similitudes entre canciones asignadas a cada uno de ellos. Vemos, por ejemplo, cómo géneros de tipo acústico (*acoustic, country, folk, blues, oldies*) aparecen arriba a la izquierda, mientras que estilos electrónicos (*ambient, chillout, electro, trance, techno*) aparecen en la zona opuesta, en la parte inferior del mapa aparecen próximos en la zona superior, muchos géneros donde la voz es predominante *(irish, blues, female vocalist, male vocalis, rnb, oldies, soul*) aparecen en la zona central; la zona derecha del gráfico está dominada por los géneros más oscuros y agresivos (*hard rock, metal, goth, psychodelic*), con conexiones directas entre ellos, mientras que entre ellos y otras «islas» próximas pero poco o nada relacionadas (*classical, new age*) existe un brazo de mar muy grande.

Las nubes de etiquetas (*tag clouds*) combinan texto con determinados efectos visuales transmisores de información. Por ejemplo, etiquetas con caracteres de mayor tamaño indican mayor presencia de archivos de esa categoría, o la disposición de etiquetas obedece a determinadas convenciones u organizaciones semánticas (véase la figura 18). Cuando el usuario pincha sobre una determinada etiqueta recupera aquellos archivos descritos con ella y, al estar dispuestas siguiendo algún tipo de ordenación lógico-semántica (por ejemplo, de más rápida a más lenta, de más alegre a más triste), la exploración permite la construcción de un mapa mental de la colección.

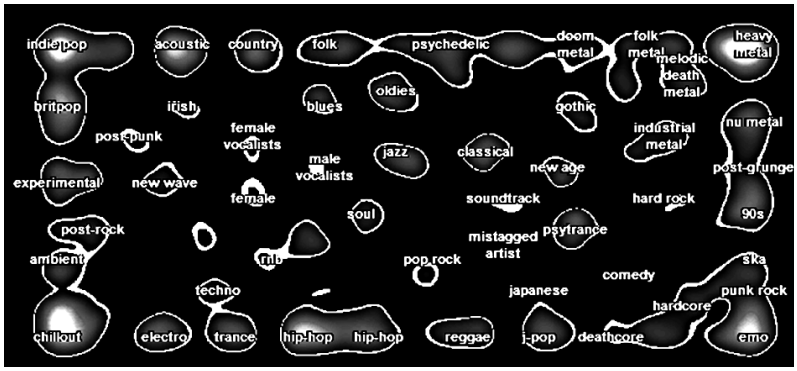

Figura 16. Mapa «Islands of Music» de géneros musicales escuchados en lastfm.com. Los géneros se ubican en base a similitudes en la música que comprenden (véase <http://www.lastfm.es/group/Playground/journal?action=display&c=1&entryid=120168692>).

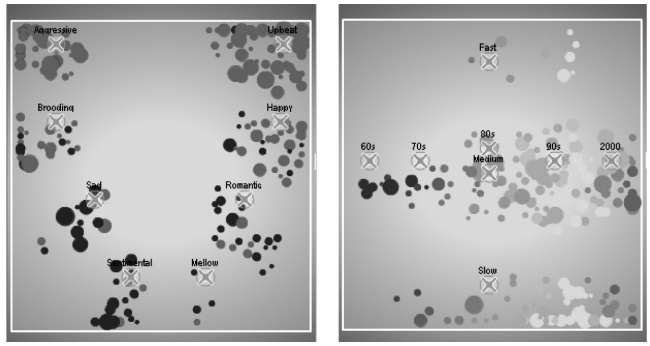

Figura 17. Interfaces de navegación basados en el contenido. A la izquierda, una colección organizada según categorías emocionales; a la derecha, organización basada en años (eje horizontal) y tempo (eje vertical) (van Gulick y otros, 2004).

Las etiquetas pueden asignarse de manera automática, pero también tienen mucho auge los enfoques colaborativos mediante los que una comunidad de oyentes asigna etiquetas según criterios personales y grupales, lo cual da lugar a las denominadas, respectivamente, «personomías» y «folk-sonomías», que son estructuras que representan una parte del conocimiento de la comunidad que los genera y utiliza activamente. Una vez etiquetados unos cuantos ejemplos, los algoritmos de clasificación automática pueden explotarse para sugerir etiquetas o directamente etiquetar nuevos archivos o aquellos que aún no han sido etiquetados por los usuarios.

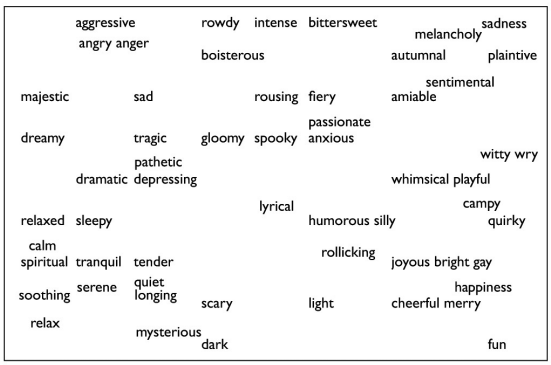

Figura 18. Nube de etiquetas emocionales asignadas a una colección musical y distribuidas a partir de la creación de un mapa auto-organizativo (Laurier y otros, 2009).

Un ejemplo que, aunque data de hace 9 años, no ha perdido vigencia ni interés es el explorador de la música del compositor Philip Glass.[4] Aunque la navegación en su obra se ha organizado a partir de descriptores semánticos manualmente anotados, muchos de ellos se podrían determinar automáticamente con las tecnologías aquí presentadas. Aparte de eso, la exploración de su obra se ve facilitada gracias a una interfaz de búsqueda multidimensional y muy directa.

Aplicaciones

La aplicación más desarrollada y en claro proceso de explotación comercial por diversas compañías consiste en la identificación de una determinada pieza musical registrada en diferentes soportes, medios y grados de calidad. Esta aplicación se basa en el cálculo de una «huella digital» (*fingerprint*) que identifica inequívocamente cada pieza musical existente en una colección. Las sociedades de gestión de derechos de autor utilizan dichas tecnologías para automatizar las estadísticas de reproducción pública de los temas musicales bajo su protección. También se han desarrollado detectores de versiones y *covers*, aunque todavía no conocemos ninguno en explotación comercial. Este problema difiere del anterior en que aquí se relaja el requerimiento de identidad: en lugar de buscarse una identidad física en el sonido lo que se busca es una similitud cuyo grado varía según los tipos de versiones (un *remix* a menudo solo preserva del original la frase del estribillo, mientras que un *cover* intenta parecerse lo más posible al original que está imitando). La organización de colecciones musicales personales a partir de criterios de similitud que explotan descriptores tímbricos, rítmicos, tonales y de género la utilizan, por ejemplo, algunos reproductores musicales de Sony, Philips, Bang & Olufsen o Yamaha (en este caso, el reproductor portátil *Bodibeat*[5] se combina con sensores de pulso, presión, etcétera para organizar automáticamente listas de reproducción acordes con la

intensidad del ejercicio o actividad que realizamos). Algunos distribuidores de música en internet empiezan a incorporar estas tecnologías, junto con la búsqueda basada en metadatos textuales asignados manualmente, para ayudar a explorar sus colecciones. En la web que Jamendo[6] utiliza para promocionar lo que se conoce como *library music* (música para bandas sonoras de videos, películas, espectáculos audiovisuales, etc.) podemos combinar descriptores de tempo, emocionales o instrumentales para filtrar los candidatos y aproximarnos a la sonoridad y musicalidad que necesitamos. También podemos experimentar con estas tecnologías en entornos de exploración musical de carácter generalista tales como Musipedia,[7] Midomi,[8] Audiobaba[9] o Mufin,[10] Desgraciadamente no conocemos aún aplicaciones específicas para bibliotecas o centros de documentación sonora y musical. Tal vez la especificidad y nivel de requerimientos de los usuarios de estas podría explicar este fenómeno. También, desgraciadamente, constatamos una desconexión entre las comunidades de profesionales que deberían estar implicadas en explotar estas tecnologías. Como decíamos al principio, las tecnologías de análisis del contenido musical de archivos sonoros no son un substituto de otras opciones de organización y búsqueda de información y su uso requiere tanto la adopción de nueva terminología como la aceptación de un cierto grado de imprecisión o la necesidad de corrección de datos por parte del usuario. No obstante, la reciente aparición de interfaces de programación de aplicaciones (API) ofrecidos por Canoris[11] y Echonest,[12] que nos facilitan el cálculo de descriptores via web, así como su integración dentro de un sistema que podemos diseñarnos a medida, podría empezar a cambiar esta situación ya que permiten la creación de entornos experimentales de descripción, acceso y organización de contenidos propios. Esperamos que nuestra introducción al tema sirva también para activar la curiosidad y los proyectos relacionados con estas tecnologías en ámbitos en los que aún no se han comenzado a introducir, ni siquiera de manera experimental.

Limitaciones y perspectivas de futuro

En este artículo hemos presentado lo que son y pueden aportar las tecnologías de análisis y descripción del contenido sonoro a la recuperación de la información musical almacenada en colecciones sonoras. Hemos examinado cómo dichas tecnologías permiten la extracción automática de rasgos descriptivos sonoros y musicales a diferentes niveles de abstracción y relacionados con diferentes facetas de la música. También hemos presentado diferentes maneras de recuperar la información indexada en base a esos descriptores, que en cualquier caso deben de considerarse como complementarios a los metadatos editoriales tradicionales así como a la información que se pueda extraer del procesado de partituras u otras representaciones simbólicas de la música. Aunque los avances experimentados en la última década son importantes y las promesas de cara al futuro resultan muy seductoras, los sistemas actuales presentan limitaciones importantes en cuanto a efectividad y eficiencia si se ponen demasiadas expectativas en la automatización. No obstante, creemos que el futuro a corto plazo asumible por instituciones y empresas que gestionan el acceso a grandes colecciones de archivos sonoros musicales consiste en incorporarlos dentro de un flujo de trabajo en el que sus resultados son tomados como sugerencias razonables por parte de un supervisor humano. En este sentido preferimos resaltar, como concepto final de nuestro artículo, el de *descripción asistida por ordenador*.

NOTAS

1 Pueden consultarse las comunicaciones presentadas en el principal congreso de dicha comunidad en <http://www.ismir.net/>

2 Probablemente la aplicación comercial más avanzada a este respecto es Melodyne, http://www.celemony.com/cms/>.

3 La precisión de los mejores algoritmos oscila entre 60% y 80%, dependiendo de los descriptores que se calculen.

4 http://www.philipglass.com/glassengine/#

5 http://www.yamaha.com/bodibeat

6 http://pro.jamendo.com/es search#pn=1&tag=&theme=&bpm=40-210&sound=&mood1=&search=&playlist=&profavorites=&voicegender=&country=&order=&lang=&inst1=&inst2=

7 http://www.musipedia.org/

8 http://www.midomi.com/

9 http://www.audiobaba.com/

10 http://www.mufin.com/

11 http://www.canoris.com/

12 http://developer.echonest.com/docs/v4/

ALGUNAS REFERENCIAS DE CARÁCTER INTRODUCTORIO

Orio, N. Music Retrieval: A Tutorial and Review, *Foundations and Trends in Information Retrieval*, 2006: 1-90.

http://dx.doi.org/10.1561/1500000002

Gouyon, F., Herrera, P., Gómez, E., Cano, P., Bonada, J., Loscos, A., Amatriain, X., Serra, X. (2008): *Content Processing of Music Audio Signals. Sound to Sense, Sense to Sound: A State of the Art in Sound and Music Computing*. 83-160.

http://smcnetwork.org/files/S2S2BOOK1.pdf

Casey, M.A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M.: *Content-Based Music Information Retrieval: Current Directions and Future Challenges*. Proceedings of the IEEE, 96(4): 668 – 696.

http://ieeexplore.ieee.org/xpl/freeabs_all jsp?arnumber=4472077

Materiales del Seminario *Intelligent Audio Systems*: *Foundations and Applications of Music Information Retrieval*, por Jay LeBoeuf y Rebecca Fiebrink.

https://ccrma.stanford.edu/wiki/MIR_workshop_2010

REFERENCIAS DEL MATERIAL GRÁFICO UTILIZADO

Goebl, W., Pampalk, E., and Widmer, G.: «Exploring expressive performance trajectories: Six famous pianists play six Chopin pieces». En *Proceedings of the 8th International Conference on Music Perception and Cognition, Evanston (ICMPC8),* Adelaida: Causal Productions, pp. 505–509, 2004.

Lesaffre, M., Leman, M., Tanghe, K., De Baets, B., De Meyer, H., Martens, J. P.: «User dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology». En *Proceedings of the Stockholm Music Acoustics Conference*, Estocolmo, 2003.

Cooper, M. and Foote, J.: «Automatic Music Summarization via Similarity Analysis». En *Proc. Third International Symposium on Musical Information Retrieval (ISMIR'02),* París, 2002.

van Gulik, R., Vignoli, F, y van de Wetering, H.: «Mapping music in the palm of your hand, explore and discover your collection». En *Proceedings of 5th International Conference on Music Information Retrieval (ISMIR'04)*, Barcelona, 2004.

Park, T.H., Li, Z., Wu, W.: «EASY Does It: The Electro-Acoustic Music Analysis Toolbox». En *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR'09)*, Kobe, 2009.

Laurier, C., Meyers, O., Serra, J., Blech, M., Herrera, P.: «Music Mood Representation from Social Tags». En *Proceedings of the 10th International Society for Music Information Conference (ISMIR'09)*, Kobe, 2009.