Tecnologías para el análisis automático del contenido musical de grabaciones de audio

Martín Rocamora[(a)(b)]

(a) Estudio de Música Electroacústica, Escuela Universitaria de Música
(b) Instituto de Ingeniería Eléctrica, Facultad de Ingeniería
Universidad de la República, Uruguay.
rocamora@fing.edu.uy

Abstract

This paper gives a brief overview of a growing field in audio processing research that pursues the development of technologies for the automatic analysis of musically meaningful content information from an audio recording. This involves various problems and applications, from computer aided musicology, to automatic music transcription and recommendation. The type of content information that can be extracted and the processes that are typically involved are examined. A summary of the state of the art in two classic problems of the area, query-by-humming and automatic music transcription is presented. It also describes some of the work of the Audio Processing Group at the Universidad de la República related with these problems.

Resumen

Este artículo busca dar un breve panorama sobre un campo de investigación en crecimiento dentro del procesamiento de audio que consiste en el desarrollo de tecnologías para el análisis automático del contenido musical de grabaciones de audio. Esto involucra diversos problemas y aplicaciones que van desde la musicología asistida por computadora, hasta la transcripción y recomendación automática de música. Se examina el tipo de información de contenido que es posible extraer así como los procesos que típicamente están involucrados. Se presenta un resumen del estado del arte en dos problemas clásicos del área, la búsqueda de música por tarareo y la transcripción automática de música. Asimismo, se describe parte del trabajo del Grupo de Procesamiento de Audio de la Universidad de la República vinculado con estos problemas.

Palabras clave: procesamiento de audio, análisis de música, aprendizaje automático, music information retrieval.

[1] Introducción

El procesamiento de audio es un área clásica muy importante dentro del procesamiento de señales, que reviste relevancia debido a sus aplicaciones en telecomunicaciones, interacción persona-máquina y música, entre otras. Además de estas aplicaciones, existen actualmente nuevas tecnologías vinculadas al manejo del abundante material audiovisual disponible y de la creciente dificultad para su organización y búsqueda. Esto se ha convertido en los últimos años en un campo muy activo de investigación, en el que se busca desarrollar herramientas para extraer de forma automática información de contenido de archivos de audio.

Es así que gran parte de la investigación en procesamiento de señales de audio se ha enfocado durante las últimas décadas a la extracción de información musicalmente relevante a partir del análisis automático de una grabación de audio (Music Information Retrieval, MIR) (Downie 2003). Esto involucra diversos problemas y aplicaciones que van desde la musicología asistida por computadora (Leech-Wilkinson 2009), hasta la transcripción automática (Klapuri *et.al.* 2006) y la recomendación automática de música (Celma 2008). A modo de ejemplo, en los últimos años hemos presenciado el surgimiento y la popularización de sistemas automáticos de recomendación personalizada de música funcionando en aplicaciones comerciales (Last.fm, Pandora, Spotify, etc) (Song 2012). Parte del éxito de estas aplicaciones se debe a la incorporación de tecnologías de procesamiento de señales y reconocimiento de patrones que facilitan el análisis automático de grandes colecciones de música. En la Figura 1 se presenta un esquema de algunos tipos de aplicación del procesamiento de señales de audio vinculados a la música y los problemas involucrados. Se indica además los que corresponden a la extracción automática de contenido musical.

En la Universidad de la República, Uruguay, viene funcionando desde el año 2005 un grupo de investigación multidisciplinario (Grupo de Procesamiento de Audio, GPA[1]) integrado por docentes del Departamento de Procesamiento de Señales del Instituto de Ing. Eléctrica (IIE) de la Facultad de Ingeniería (FING) y del Estudio de Música Electroacústica (eMe) de la Escuela Universitaria de Música (EUM). Su cometido es la investigación de un amplio espectro de problemas que involucran el desarrollo de técnicas de procesamiento digital de señales y su aplicación a señales de audio con contenido musical. En particular técnicas de análisis de audio aplicadas a la representación tiempo-frecuencia y a la resíntesis de sonido, la separación de fuentes en señales de audio polifónico, y la extracción de contenido musical (melodía, ritmo, armonía, etc) de grabaciones de audio. También es objeto de estudio de este grupo la aplicación de estas técnicas en tareas como la búsqueda de música por contenido en bases de datos de grabaciones, y el análisis, transformación y resíntesis de audio en la composición electroacústica y artes audiovisuales.

El propósito de este artículo es presentar un panorama resumido de la disciplina conocida como MIR, es decir, la extracción automática de contenido musical a partir de grabaciones de audio, a través de la descripción del estado del arte en algunos de sus problemas clásicos. A su vez, se busca poner en contexto y describir brevemente el trabajo de investigación del GPA desarrollado en el área.

El resto del artículo se organiza de la siguiente forma. A continuación, la sección [2] discute en qué consiste la extracción de contenido musical y cuáles son sus principales dimensiones. Seguidamente, la sección [3] repasa brevemente el estado del arte en uno de los problemas clásicos de búsqueda de música por contenido y en la transcripción automática de música. La sección [4] presenta de

1 http://iie.fing.edu.uy/investigacion/grupos/gpa/

forma resumida algunos de los principales trabajos del GPA en relación a dichos problemas y la sección [5] cierra el artículo con consideraciones finales.

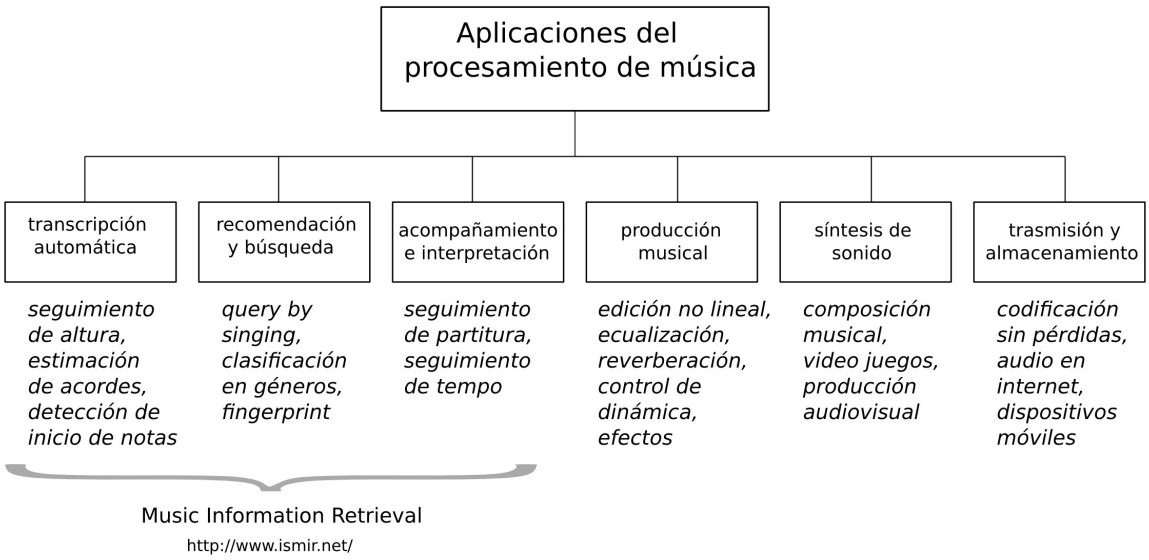


Figura 1. Algunos tipos de aplicación del procesamiento de música y problemas involucrados. Se indica los que corresponden a Music Information Retrieval.

[2] Descripción de contenido musical

Cabe preguntarnos a qué nos referimos con contenido musical en este contexto. Tal como se señala en (Herrera *et.al.* 2011), un descriptor de contenido musical puede entenderse de forma general como todo dato que puede constituir un predicado del contenido del archivo sonoro con el fin de caracterizarlo. De esta forma, 120 pulsos por minuto, tener forma de sonata, involucrar sonidos de clarinete, estar en tonalidad de do mayor, etc. son todos posibles descriptores de un supuesto archivo de audio. La descripción puede tener en cuenta diferentes aspectos musicales, tales como ritmo, armonía, melodía, estructura o instrumentación. Si bien es posible, con ciertas limitaciones, derivar algunos descriptores de forma automática a partir de un archivo de audio, existen otros para los cuales necesariamente debe intervenir un oyente experto.

A primera vista nuestro objetivo parece ser el de automatizar la transcripción a notación musical a partir de un archivo de audio, ya que una partitura tiene explícita o implícitamente mucha de la información mencionada. Si bien la transcripción automática de música es uno de los problemas clásicos de MIR, la tecnología actual sólo es capaz de transcribir de forma exitosa música monódica o polifonías simples y bajo ciertas restricciones. A su vez, debemos reconocer que una partitura en notación musical convencional deja de lado varios aspectos, como por ejemplo cuestiones de interpretación (p.ej. variaciones micro-temporales o de entonación), y no es aplicable a todas las tradiciones musicales existentes algunas de las cuales ni si quiera utilizan algún tipo de notación.

Es importante señalar que existe una gran variedad de descriptores, que pueden corresponder a diferentes niveles de abstracción, escalas temporales e incluso depender del contexto. La Figura 2 presenta un esquema de posibles niveles de abstracción en análisis automático de música a partir de una señal de audio. Los descriptores que podemos denominar de bajo nivel tienen que ver con aspectos

como la energía instantánea de la señal, la ubicación y duración de eventos, los contornos de altura (o más precisamente de frecuencia fundamental), o con representaciones espectrales de tiempo corto que pueden asociarse al timbre. Para obtenerlos se parte típicamente de algún tipo de representación de la señal (temporal, espectral o tiempo-frecuencia), y se aplica una transformación que busca capturar algún aspecto particular. Estos descriptores de bajo nivel pueden combinarse y agregarse temporalmente para conformar descriptores de más alto nivel y a diferentes escalas temporales, asociados a entidades musicales como notas o acordes. A su vez, los descriptores obtenidos pueden usarse nuevamente para generar descriptores de más alto nivel, por ejemplo combinar las notas para obtener melodías. Ocurre que a medida que se crece en abstracción se pasa de información que podríamos considerar más objetiva a información cada vez más subjetiva, y de la simulación de procesos cognitivos innatos a la recreación de habilidades aprendidas.

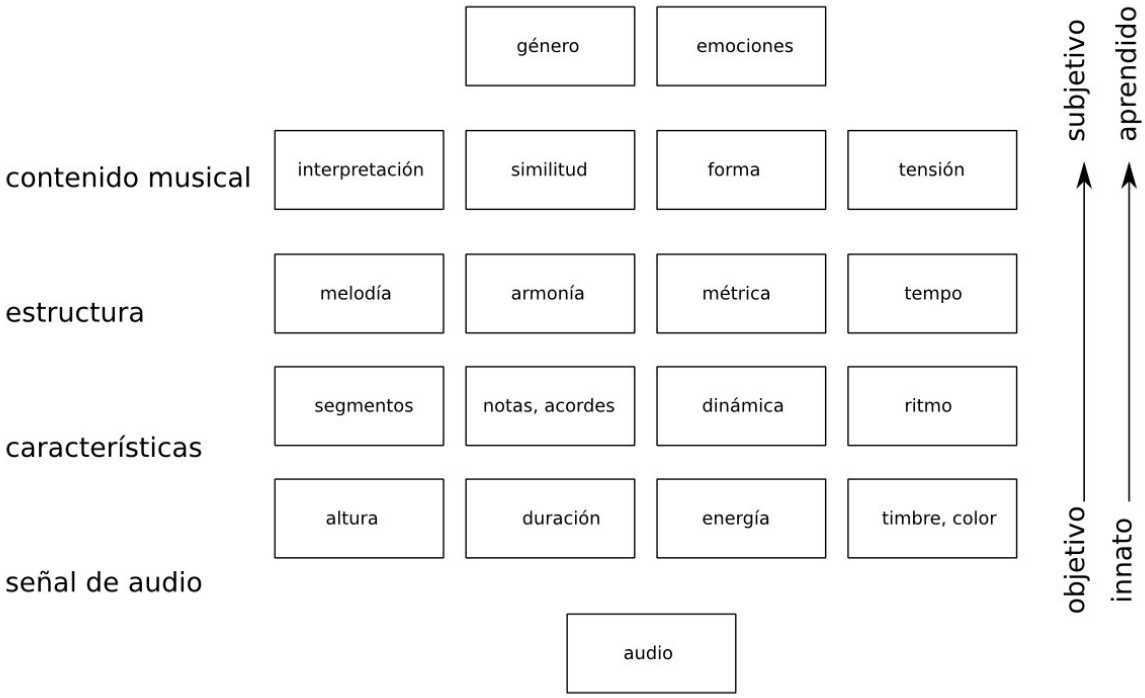


Figura 2. Esquema de posibles niveles de abstracción involucrados en la descripción de contenido musical a partir de señales de audio.[2]

Para alcanzar los niveles de abstracción más altos se suelen incorporar modelos semánticos que buscan capturar similitudes y diferencias entre conceptos. Estos modelos pueden ser formulados a partir de conocimiento musical o aplicando herramientas de aprendizaje automático. Esto último requiere de bases de datos de aprendizaje anotadas manualmente que permitan entrenar a los algoritmos en la tarea que se desea realizar.

La Figura 3 presenta un ejemplo de cálculo de descriptores de bajo nivel y de como pueden ser combinados para obtener entidades musicales, en este caso notas. Se parte de una señal de audio en la que una voz interpreta una cierta melodía. Los descriptores de bajo nivel calculados son el contorno de frecuencia fundamental y los inicios de notas. Luego esta información se combina para

2 Basado en una presentación de Petri Toivanen realizada en el Simposio Brasilero de Computación Musical, SBCM 2005, Belo Horizonte, Brasil.

determinar inicio, duración y altura de las notas en una escala igualmente temperada. Si se quisiera determinar el grado de similitud entre esta secuencia de notas y otra, para establecer si dos melodías se asemejan, la representación usada para la comparación debería ser invariante a trasposiciones de altura (mas grave o más aguda) y a las diferencias de tempo (más rápido o más lento). En la Figura 4 se muestra cómo podría codificarse la secuencia de notas en contornos de altura y duración para implementar este modelo semántico de similitud.

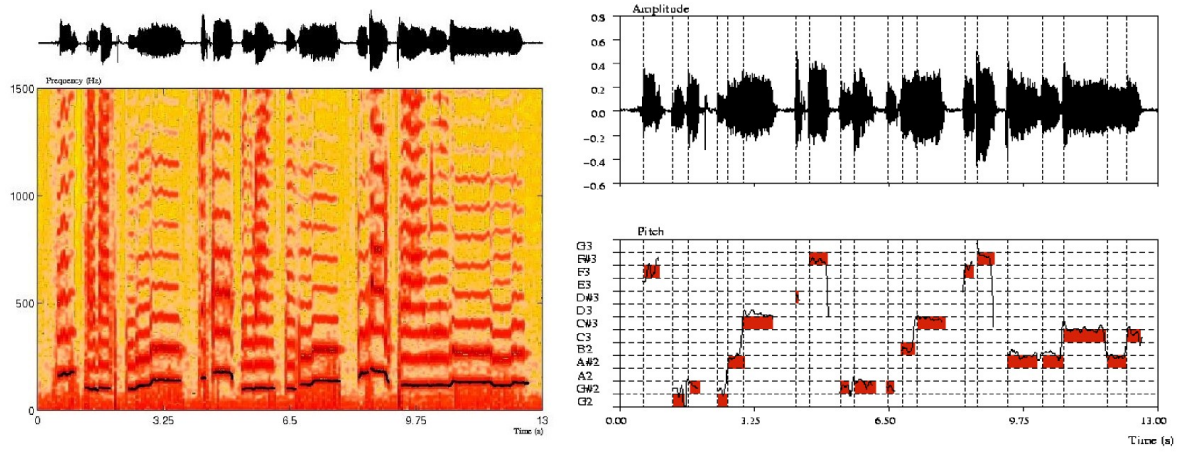
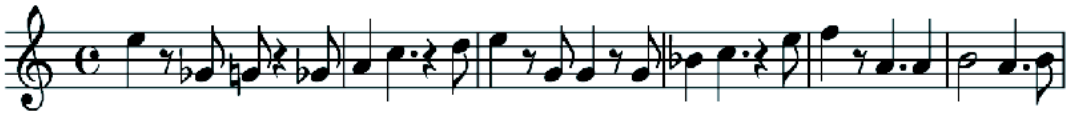


Figura 3. Ejemplo de cálculo de descriptores de bajo nivel a partir de una señal de voz cantada. A la izquierda se indica sobre el espectrograma el contorno de frecuencia fundamental detectado. A la derecha, se muestra el resultado de la detección de eventos sobre la forma de onda y como pueden combinarse ambos descriptores para generar una transcripción a notas.



MIDI Note	53	43	44	43	46	49	51	53	44	44	44	47	49	53	54	46	46	48	46	48
Pitch Interval	*	-10	1	-1	3	3	2	2	-9	0	0	3	2	4	1	-8	0	2	-2	2
Duration (♪)	3	1	3	1	2	5	1	3	1	3	1	2	5	1	3	3	2	4	3	1
Relative Duration	*	$\frac{1}{3}$	3	$\frac{1}{3}$	2	$\frac{5}{2}$	$\frac{1}{5}$	3	$\frac{1}{3}$	3	$\frac{1}{3}$	2	$\frac{5}{2}$	$\frac{1}{5}$	3	1	$\frac{2}{3}$	2	$\frac{3}{4}$	$\frac{1}{3}$

Figura 4. Ejemplo de codificación de la secuencia de notas en contornos de altura y duración para lograr invarianza a transposición de altura y tempo. Las alturas se representan en números MIDI y las duraciones corresponden al intervalo entre inicios de notas sucesivos normalizados al valor de corchea.

[3] Extracción automática de contenido en señales de audio

En esta sección se presenta un breve panorama del estado del arte en dos problemas clásicos de la extracción automática de contenido de señales de audio aplicada a la música. El primero está orientado a la aplicación de búsqueda de

música por contenido. El segundo, vinculado estrechamente con el anterior, corresponde al problema más general de la transcripción automática de música.

3.1 Búsqueda de música por contenido

La búsqueda de material audiovisual se ha convertido en un campo muy activo de investigación en los últimos años. Se busca desarrollar herramientas de procesamiento para extraer automáticamente una descripción de su contenido, lo que permite clasificarlo y organizarlo de manera práctica y eficiente para su identificación y búsqueda. La búsqueda de imágenes basada en contenido tiene ya cierto desarrollo, pero es más reciente su estudio en el dominio del audio, si bien ya existen aplicaciones comerciales que persiguen este objetivo. Por ejemplo, Google Audio Indexing (GAudi) es una herramienta que permite buscar a través de vídeos analizando el contenido de audio con tecnología de transcripción de voz a texto. La búsqueda y recomendación automática de música es un terreno en el que también se aplica el procesamiento de audio para extracción de contenido. Por ejemplo, Spotify es un servicio de recomendación de música por internet muy popular y premiado que utiliza características calculadas a partir de la señal de audio usando tecnología de http://the.echonest.com/

La búsqueda de música por contenido comprende la clasificación automática en géneros musicales, p.ej. rock, jazz (Scaringella *et.al.* 2006) y el reconocimiento de patrones rítmicos o melódicos específicos. Es frecuente que las personas identifiquen un tema musical tarareando un fragmento de su melodía. Esta es entonces una forma natural de ingresar la consulta, lo que recibe el nombre de búsqueda de música por tarareo (QBH, query by humming). Otros enfoques son la búsqueda de música únicamente a partir de patrones rítmicos (Hanna *et.al.* 2009) o a través de un ejemplo (QBE, query by example), es decir usando como consulta un fragmento grabado del audio original y en ocasiones afectado por ruido ambiente o distorsiones (Haitsma 2002).

Búsqueda de música por tarareo

Los sistemas de búsqueda de música por tarareo buscan identificar la melodía cantada por el usuario en una base de datos de melodías. Desde el sistema propuesto en (Ghias *et.al.* 1995), se han considerado distintas formas de enfrentar este problema (Hu *et.al.* 2002). El enfoque más habitual consiste en transcribir la señal de voz a una secuencia de notas y buscar instancias similares a ese patrón en una base de datos de melodías en notación simbólica. A pesar de que el problema ha recibido mucha atención de la comunidad científica por más de una década, la generación automática de la base de datos de melodías contra la que son comparadas las consultas permanece como un problema abierto. En todas las propuestas, salvo muy pocas excepciones, la base de datos se compone de música codificada en alguna notación simbólica, p.ej. MIDI. Esto se debe principalmente a que no existen formas automáticas lo suficientemente robustas de extraer la melodía de una grabación para compararla con la melodía cantada por el usuario. Sin embargo, el alcance de este enfoque está limitado por la necesidad de transcribir manualmente (i.e. audio a MIDI) cada nueva canción de la base de datos. Una forma de evitar este problema es construir una base de datos de consultas provistas por los propios usuarios y comparar las nuevas consultas con las grabadas anteriormente (Pardo *et.al.* 2008). Este enfoque simplifica drásticamente el problema y es aplicado en servicios de búsqueda de música como SoundHound o Midomi. Sin embargo, el proceso no es automático sino que depende de la contribución de los usuarios. Además, no es posible buscar una canción hasta que un usuario la grabe y etiquete por primera vez. Para extender los sistemas QBH a gran escala es necesario automatizar

completamente el proceso de generación de la base de datos. Existen solo unas pocas propuestas de un sistema de este tipo (Ryynänen *et.al.* 2008, Salamon *et.al.* 2013) y una de las más recientes proviene de nuestro grupo de investigación (Rocamora *et.al.* 2014). Si bien los resultados obtenidos son alentadores hay mucho margen de mejora posible para lograr alcanzar el desempeño de los sistemas tradicionales basados en bases de datos simbólicas. Para ello es imprescindible perfeccionar las herramientas existentes para la extracción automática del contenido melódico de grabaciones de música. Esto constituye una etapa esencial en el problema más general de transcripción automática de música.

3.2 Transcripción automática de música

La transcripción de música en el sentido de la representación puede considerarse como el proceso de transformar una señal acústica en una representación simbólica. Tradicionalmente, en la música escrita se utilizan notas para indicar la altura, inicio, y duración de cada sonido a interpretarse. El problema de automatizar a través de una máquina la transcripción de música es objeto de estudio desde hace décadas (Moorer 1977), pero aún permanece sin resolverse por completo (Klapuri *et.al.* 2006). Además de facilitar la búsqueda de música por contenido, estas herramientas tienen aplicación en la pautación de música para estudios musicológicos, análisis de interpretaciones, seguimiento de partituras y acompañamiento automático, entre otros.

Extracción de contenido melódico: F0 múltiple

La transcripción automática de música comprende la solución de diversos subproblemas. Uno de los principales es el reconocimiento de las alturas tonales o escalares de los eventos, lo que corresponde a la estimación de su frecuencia fundamental (estimación de F0). Existen ya diversas soluciones que resuelven este problema de manera bastante adecuada para el caso de texturas monódicas, es decir una sola altura a la vez (Cheveigné *et.al.* 2002). Típicamente, sin embargo, los casos reales implican texturas polifónicas, es decir con varias alturas simultáneas, lo que puede ser el resultado de la superposición de más de una línea melódica y/o de la presencia de instrumentos que producen varias notas simultáneas (acordes). Esto implica la estimación de F0 múltiple, para lo cual no existe aún una solución general y robusta.

Varias de los criterios más habituales de organización musical (coincidencia rítmica de diferentes voces, las relaciones armónicas entre alturas simultáneas) son precisamente los que hacen que el problema de la transcripción automática sea verdaderamente desafiante. En la composición y la orquestación suelen combinarse sonidos de distintos instrumentos de modo de conformar una entidad musical única desde el punto de vista perceptivo. En base a esto pueden identificarse los problemas esenciales de la estimación de F0 múltiple. El problema de agrupamiento consiste en identificar a qué fuente de sonido corresponde cada componente del espectro, considerando que éstas pueden incluso no ubicarse bajo una relación armónica perfecta. Otro de los problemas es el cálculo de la prominencia de cada fuente de sonido en función de la amplitud, la fase y la frecuencia de sus componentes. Además, en presencia de instrumentos percusivos la relación señal a ruido puede ser muy baja por momentos. Por último, uno de los principales problemas consiste en la coincidencia de parciales de distintos sonidos, lo que en la música occidental es una regla más que una excepción.

Existe una gran cantidad de propuestas para atacar el problema de estimación de F0 múltiple, véase (Cheveigné 2006) por una revisión. Dado que el problema

está fuertemente vinculado a la separación de audio en fuentes sonoras, una serie de propuestas se basan en los principios y mecanismos de la psicología del Análisis del Panorama Auditivo (ASA, Auditory Scene Analysis) (Bregman 1993). Los componentes sinusoidales se agrupan basándose en la relación armónica, el sincronismo en el cambio de sus parámetros, la proximidad en el tiempo y la frecuencia, y la proximidad espacial. Un tipo de abordaje clásico consiste en estimar la F0 del sonido más prominente y eliminarlo de la mezcla original, en un proceso iterativo (Klapuri 2003). Para el agrupamiento, además de la relación armónica, se utiliza el hecho de que la envolvente espectral de los sonidos varía suavemente en función de la frecuencia. Es habitual restringir el problema a la estimación de la melodía principal y en ocasiones considerando también la línea de bajo, en base a la hipótesis de que corresponden a las estructuras armónicas predominantes en las bandas de alta y baja frecuencia respectivamente y que tienen trayectorias continuas en el tiempo (Goto 2000). En la última década la extracción de la melodía principal ha sido un problema de investigación muy activo (Salamon *et.al.* 2014), que fue abordado por el grupo proponiendo enfoques novedosos y logrando buenos resultados.

Transcripción y análisis automático de música de percusión

Dentro de la investigación llevada adelante sobre transcripción automática de música, la percusión ha recibido típicamente menor atención. El problema de la transcripción automática de percusión consiste en obtener a partir de una señal de audio una representación simbólica que establezca el tipo de instrumento y la ubicación temporal de cada uno de los eventos. En muchos casos interesa además producir información sobre intensidad y modo de ejecución.

Los primeros intentos de transcribir música de percusión fueron hechos a mediados de 1980 por Schloss (Schloss 1985), quien clasifica automáticamente distintos golpes de conga en interpretaciones naturales, usando la energía relativa entre diferentes porciones del espectro. Luego, Bilmes (Bilmes 1993) logra además diferenciar el sonido de diferentes congas usando un algoritmo de agrupamiento. Más adelante, se aborda la transcripción de música de percusión polifónica (Goto *et.al.* 1993), lo que agrega la dificultad de la mezcla de sonidos. La clasificación de sonidos aislados puede considerarse un problema suficientemente estudiado, pero el desempeño alcanzado cae notoriamente al clasificar sonidos simultáneos o transcribir interpretaciones reales.

La mayor parte del trabajo realizado en la última década se concentra en el reconocimiento y transcripción de sonidos de batería de pop/rock (FitzGerald 2004), donde se han logrado buenos resultados con un conjunto limitado de instrumentos (típicamente bombo, redoblante y hi-hat). La transcripción de sonidos de batería en música polifónica, en la que se admiten además sonidos de altura definida, es un problema abierto, abordado en trabajos más recientes (Gillet *et.al.* 2008). Muy pocos trabajos se orientan al reconocimiento de diferentes tipos de golpe en un mismo instrumento, lo que constituye un problema más desafiante ya que las diferencias de timbre tienden a ser más sutiles.

Los enfoques existentes para la transcripción de percusión pueden clasificarse básicamente en dos tipos (Fitzgerald *et.al.* 2006). La gran mayoría sigue una estrategia de reconocimiento de patrones aplicado a eventos sonoros que consiste básicamente en los siguientes pasos. En primer lugar se segmenta la señal de audio, ya sea por medio de la identificación de eventos o la construcción de una grilla regular de pulso (tatum). Luego se extraen características para cada segmento de audio, las cuales usualmente describen el contenido espectral y su evolución temporal, p.ej. centroide, MFCCs, ZCR (Fitzgerald *et.al.* 2006). Finalmente se procede a la clasificación usando los esquemas típicos de

reconocimiento de patrones, como SVM o K-NN. El otro enfoque habitual está basado en la separación de la señal de audio en flujos que contienen únicamente un tipo de sonido de percusión, usando diferentes técnicas como ISA, NMF o Sparse Coding (Gillet *et.al.* 2008). El siguiente paso es etiquetar cada flujo, indicando el tipo de instrumento. Por último, se obtiene la información temporal de los eventos aplicando un algoritmo de detección de ataques.

Por lo general, los métodos propuestos se concentran en procesamiento de señales de bajo nivel, sin aplicar análisis en niveles de abstracción más altos. Tal como ocurre en el reconocimiento automático de voz, en que los modelos lingüísticos han permitido un aumento considerable de desempeño, es razonable asumir que puede ser beneficioso incorporar ideas similares para modelar la dependencia estadística entre eventos mediante modelos musicológicos. Si bien ha habido relativamente poco trabajo en transcripción de percusión usando modelos musicológicos (Fitzgerald *et.al.* 2006), existen algunas propuestas que utilizan técnicas como N-gramas o cadenas de Markov [18]. Por otra parte, la combinación entre el reconocimiento de bajo nivel y el modelado de alto nivel requiere aún más desarrollo (Fitzgerald *et.al.* 2006).

[4] Algunos trabajos de investigación a nivel local

Desde hace varios años el Grupo de Procesamiento de Audio de la Universidad de la República trabaja en la extracción de contenido melódico de grabaciones de audio. En particular, una de sus principales contribuciones es el desarrollo de una representación tiempo-frecuencia no tradicional denominada Fan Chirp Transform (FChT), y su aplicación al análisis de música (Cancela *et.al.* 2010, Jure *et.al.* 2012). Dicha transformada permite obtener una representación precisa de los componentes de una señal armónica no estacionaria, como la voz cantada. Esto facilita la tarea de algoritmos de extracción de contenido melódico de grabaciones de música y, a su vez, permite su visualización con un grado de detalle y precisión que no es posible a través de técnicas de representación tiempo-frecuencia tradicionales. Un subproducto de esta investigación es la construcción de un sistema de extracción de la melodía principal, lo que constituye un primer paso en la transcripción automática de música. Esto a su vez tiene un fuerte impacto en aplicaciones de búsqueda de música por contenido, tema que ha sido abordado por el grupo en diferentes trabajos de investigación (López *et.al.* 2005, López *et.al.* 2006, Rocamora *et.al.* 2014). Parte de estas técnicas han sido incorporadas en una aplicación de análisis y visualización de audio existente, y puestas a disposición de sus potenciales usuarios (músicos, estudiantes de música, musicólogos, investigadores) durante un proyecto de investigación reciente (Jure *et.al.* 2012).

Por otra parte, la extracción automática de contenido rítmico es otra de las líneas de investigación abordada más recientemente y resulta complementaria de la anterior. En particular se busca el desarrollo de herramientas automáticas orientadas al análisis y estudio de músicas basadas en percusión, capaces de extraer contenido rítmico de señales de audio, analizar grandes colecciones de música e interactuar en tiempo real con intérpretes. Para ello se propone tomar como caso de estudio el Candombe afro-uruguayo, lo que plantea una serie de nuevos desafíos y oportunidades.

En sintonía con parte del trabajo en curso en la comunidad científica de MIR, se espera que, como se señala en (Serra 2011), el estudio cuidadoso de una tradición musical particular fuera del paradigma de música comercial occidental pueda contribuir a la construcción de modelos más generales y ricos que los que actualmente dominan la investigación en tecnologías de la información aplicadas

a la música. Esto se basa en la observación de que si los esfuerzos orientados al desarrollo de las TIC aplicadas a la música están exclusivamente pautados por el mercado y no toman en cuenta una realidad multicultural, los avances alcanzados pueden en realidad profundizar el sesgo existente hacia la difusión, recomendación y acceso a un tipo muy reducido de música, restringiendo la diversidad de la oferta.

Por otra parte, el abordaje de problemas musicales en contextos culturales específicos plantea nuevos desafíos para las soluciones ya existentes y puede redundar en nuevas técnicas y metodologías aplicables a una amplia variedad de situaciones (Serra 2011). Por ejemplo, el problema de seguimiento automático de pulso y análisis rítmico computacional ha sido un campo activo de investigación desde hace varias décadas (Gouyon 2008). Esto tiene diversas aplicaciones que van desde análisis musicológico hasta software comercial para sincronismo de música para DJs o la industria audiovisual. Si bien los algoritmos existentes alcanzan un desempeño razonable para música occidental de tiempo marcado, tienen dificultades para manejar adecuadamente músicas sincopadas o polirítmicas, como ciertos tipos de música de origen Africano (Jehan 2005). Esto ha llevado al desarrollo de algoritmos de seguimiento de pulso para géneros musicales específicos (Hockman *et.al.* 2012).

En la investigación llevada adelante recientemente hemos confirmado que los algoritmos del estado del arte en seguimiento de pulso presentan un desempeño muy limitado al procesar interpretaciones de Candombe. Por otra parte, hemos propuesto un algoritmo que se basa en la detección de patrones rítmicos siguiendo un modelo Bayesiano basado en cadenas ocultas de Markov (HMM, Hidden Markov Models) como en (Whiteley *et.al.* 2006) con resultados muy alentadores. Creemos además que el enfoque de detección de patrones rítmicos para inferir el pulso y la estructura métrica es aplicable a otros tipos de música de percusión que, como en el caso del Candombe, presentan frases rítmicas repetitivas.

[5] Consideraciones finales

El presente artículo ha dado un breve panorama sobre un campo de investigación en crecimiento que consiste en el desarrollo de tecnologías para el análisis automático del contenido musical de grabaciones de audio. Se examinó el tipo de información de contenido que es posible extraer así como los procesos involucrados. Se repasó el estado del arte en dos problemas clásicos del área, la búsqueda de música por tarareo y la transcripción automática de música. Asimismo, se describió parte del trabajo del Grupo de Procesamiento de Audio de la Universidad de la República vinculado con estos problemas.

A modo de cierre podemos señalar que si bien las herramientas de análisis automático de contenido musical han experimentado avances significativos en los últimos años, lo que de algún modo se manifiesta en que han sido incorporadas en aplicaciones comerciales, su desempeño es en general todavía limitado comparado con el de un humano experto.[3] A pesar de ello, si el análisis automático es supervisado por un usuario puede resultar eficaz y facilitar el procesamiento de grandes cantidades de datos a una escala difícil de alcanzar con métodos tradicionales. Por estas razones, en el futuro próximo es esperable que estas herramientas se incorporen cada vez más a aplicaciones de búsqueda y recomendación de música, análisis musicológico, entrenamiento y aprendizaje musical, entre otras.

3 Una buena referencia del estado del arte en una amplia gama de problemas de MIR es la contienda anual de algoritmos MIREX accesible desde http://www.music-ir.org/

Referencias bibliográficas

J. S. Downie, *Music information retrieval*, Annual Review of Information Science and Technology, vol. 37, no. 1, pp. 295–340, 2003.

D. Leech-Wilkinson, *The Changing Sound of Music: Approaches to Studying Recorded Musical Performance*. London: Centre for the History and Analysis of Recorded Music (CHARM), 2009.

A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. Springer, 1st ed., 2006.

O. Celma, *Music Recommendation and Discovery in the Long Tail*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2008.

Y. Song, M. Pearce, and S. Dixon, *A survey of music recommendation systems and future perspectives.* In 9th Int. Symposium on Computer Music Modelling and Retrieval, pages 395-410, 2012.

P. Herrera and E. Gómez, Tecnologías para el análisis del contenido musical de archivos sonoros y para la generación de nuevos metadatos. Boletín de la Asocación Española de Documentación Musical, vol. 14, pages 28-38, 2011.

N. Scaringella, G. Zoia, and D. Mlynek, *Automatic genre classification of music content: a survey*. Signal Processing Magazine, IEEE , vol. 23, no.2, pp.133,141, 2006.

P. Hanna and M. Robine, *Query by tapping system based on alignment algorithm*. Acoustics, Speech and Signal Processing, ICASSP. IEEE Int. Conf. on, pp. 19-24, 2009.

J. Haitsma, *A Highly Robust Audio Fingerprinting System*, Proceedings of the 4th International Conference on Music Information Retrieval, pp. 107-115, 2002.

A. Ghias, J. Logan, D. Chamberlin and B.C. Smith. *Query by humming: musical information retrieval in an audio database*. In: Proceedings of the third ACM International Conference on Multimedia, MULTIMEDIA ’95. ACM, New York, NY, USA, pp. 231–236, 1995.

N. Hu and R.B. Dannenberg, *A comparison of melodic database retrieval techniques using sung queries*. In: Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries, JCDL’02. ACM, New York, NY, USA, pp. 301–307, 2002.

B. Pardo, D. Little, R. Jian, H. Livini, J. Han, *The vocalsearch music search engine*. In: Proc. of the 8th ACM/IEEE-CS Joint Conf. on Digital libraries, USA, p. 430, 2008.

M. Ryynänen, A. Klapuri,*Query by humming of MIDI and audio using locality sensitive hashing.* In: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 2249–2252. 2008.

J. Salamon, J. Serrà, E. Gómez, *Tonal representations for music retrieval: from version identification to query-by-humming*. Int. Journal of Multimedia Information Retrieval, vol. 2 (1), 45–58, 2013.

M. Rocamora, P. Cancela, A. Pardo. *Query by humming: Automatically building the database from music recordings*. Pattern Recognition Letters, vol 36, pp. 272-280, 2014.

J. A. Moorer, *On the Transcription of Musical Sound by Computer*. Computer Music Journal, vol. 1, No. 4, pp. 32-38, November 1977.

A. de Cheveigné, H. Kawahara, *YIN, a fundamental frequency estimator for speech and music*. The Journal of the Acoustical Society of America 111 (4), 1917–1930. 2002.

A. de Cheveigné, *Multiple F0 estimation*, in Computational Auditory Scene Analysis: Principles, Algorithms and Applications, D. Wang and G. Brown, Eds., pp. 45–79. IEEE / Wiley, 2006.

A. S. Bregman, *Auditory Scene Analysis*. MIT Press, Cambridge, Massachusetts, 1993.

A. Klapuri, *Multiple fundamental frequency estimation based on harmonicity and spectral smoothness*, Speech and Audio Processing, IEEE Trans. on, 11(6):804-816, 2003.

M. A. Goto, *A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings*, Acoustics, Speech, and Signal Processing, 2000.

J. Salamon, E. Gómez, D. P. W. Ellis and G. Richard, *Melody Extraction from Polyphonic Music Signals: Approaches, Applications and Challenges*, IEEE Signal Processing Magazine vol. 31(2):118-134, 2014.

W. A. Schloss, *On the automatic transcription of percussive music from acoustic signal to high-level analysis*, Master's thesis, Stanford University, Stanford, CA, 1985.

J. Bilmes, *Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm*, Master's thesis, MIT, Cambridge, 1993.

M. Goto, M. Tabuchi, and Y. Muraoka, *An automatic transcription system for percussion instruments*, in Procs. of the 46th Annual Convention IPS Japan, 7Q-2, 1993.

D. FitzGerald, *Automatic Drum Transcription and Source Separation.* PhD thesis, Dublin Institute of Technology, Dublin, Ireland, 2004.

O. Gillet, G. Richard, *Transcription and separation of drum signals from polyphonic music*, Audio, Speech, and Language Proc., IEEE Transs. on, vol. 16, pp. 529 -540, 2008.

D. Fitzgerald, J. Paulus, *Unpitched Percussion Transcription*, in Signal Proc. Methods for Music Transcription (A. Klapuri and M. Davy, eds.), pp. 131–162, Springer, 2006.

P. Cancela, E. López, M. Rocamora. *Fan chirp transform for music representation*, Int. Conference on Digital Audio Effects, 13th. DAFx-10. Graz, Austria - 6-10 Sep. 2010

L. Jure, E. López, M. Rocamora, P. Cancela, H. Spontón, I. Irigaray, *Pitch content visualization tools for music performance analysis*, Int. Society for Music Information Retrieval Conf., 13th, Procs. ISMIR 2012. Porto, Portugal, page 493--498, 2012

E. López, M. Rocamora. *Tararira: Sistema de búsqueda de música por melodía cantada*. Brazilian Symposium on Computer Music. Belo Horizonte, Brazil - Oct. 2005.

E. López, M. Rocamora, Tararira: Query by singing system. In: The Second Annual Music Information Retrieval Evaluation eXchange (MIREX 2006), pp. 80–83, 2006.

X. Serra, *A multicultural approach in music information research*, in Int. Soc. for Music Information Retrieval Conf. (ISMIR), (Miami, Florida, USA), 2011.

F. Gouyon, *Computational Rhythm Description*, VDM Verlag, 2008.

T. Jehan, *Downbeat prediction by listening and learning*, Applications of Signal Processing to Audio and Acoustics. IEEE Workshop on, pp. 267-270, 2005.

J. Hockman, M.E.P. Davies, I. Fujinaga, *One in the Jungle: Downbeat Detection in Hardcore, Jungle, and Drum and Bass*, In Proc. of the 13th Int. Society for Music Information Retrieval Conf., Porto, Portugal, October 8-12, 2012, pp. 169-174, 2012.

N. Whiteley, A. Cemgil, and S. Godsill. *Bayesian modelling of temporal structure in musical audio*. In Proc. of the 7th Int. Conf. on Music Information Retrieval, pp. 29-34, Victoria, Canada, 2006.