UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

ANÁLISIS DE LAS CARACTERÍSTICAS ACÚSTICAS DE LA PERCUSIÓN HUMANA

TRABAJO DE FIN DE MÁSTER

Máster en Ingeniería Informática y de Telecomunicación
Programa Oficial de Posgrado en Ingeniería Informática y de Telecomunicación

Daniel Hernández López
Ingeniero de Telecomunicación, UAM
Madrid, Febrero de 2010

ANÁLISIS DE LAS CARACTERÍSTICA ACÚSTICAS DE LA PERCUSIÓN HUMANA

AUTOR: Daniel Hernández López
TUTOR: Doroteo Torre Toledano

Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Febrero 2010

TRABAJO FIN DE MÁSTER

Título: *Análisis de las características acústicas de la percusión humana*

Autor: D. Daniel Hernández López
Ingeniero de Telecomunicación (UAM)

Tutor: D. Doroteo Torre Toledano
Doctor Ingeniero de Telecomunicación (UPM)

Tribunal: Doroteo Torre Toledano
Universidad Autónoma de Madrid

Daniel Ramos Castro
Universidad Autónoma de Madrid

Joaquín González Rodríguez
Universidad Autónoma de Madrid

Fecha de lectura:

Calificación:

Resumen:

En este Trabajo Fin de Máster se detalla el proceso seguido para comenzar con una nueva línea de investigación prácticamente desde cero, el análisis y tratamiento de señales de audio de percusión humana. Para ello, en primer lugar se presenta el concepto de percusión humana, sus tipos y las posibilidades que ofrece. En segundo lugar se analiza el estado del arte en tecnologías similares a la que se pretende desarrollar y se buscan puntos en común que puedan ser de utilidad.

Para poder desarrollar un trabajo de investigación en una línea prácticamente inexistente, el siguiente paso ha sido generar una base de datos del tipo de datos que se pretenden analizar. Por ello se explica el diseño y las características de la base de datos, así como los protocolos de grabación, revisión y una propuesta de protocolo de pruebas.

A continuación se realiza un análisis de las características de los datos recogidos, sus peculiaridades y detalles que puedan ayudar a establecer recomendaciones acerca del trabajo futuro en mecanismos de clasificación de los eventos de percusión humana analizados.

Palabras clave:

Percusión humana, análisis espectral, análisis acústico, percusión vocal onomatopéyica, beat box, imitación de percusión, consulta por percusión, transcripción de audio a MIDI.

Abstract:

In this final Master degree paper it´s detailed all the process of a new line of an investigation project from the beginning, the analysis and treatment of human percussion audio signals. To explain that, in the first place, the concept of human percussion will be defined, mentioning its types and all the possibilities than they can offer. In the second place, we will find a short study of the state of the art in similar technologies to be developed, and also we´ll try to find the things in common that could be useful to the investigation.

To be able to develop an investigation of these characteristics, following a line almost nonexistent, the next step was to generate a Data Base with the type of information that we are trying to analyze. That’s the reason why it’s explained the design and characteristics of the Data Base, as well as the recording protocols, the revision and also a proposal of some protocol proves.

Following to that, an analysis of the characteristics of all the information collected it´s been done, including its singularities and another details that can help to establish some recommendations about a future work of the classified mechanisms of the human percussion events analysed.

Key words:

Human drumming, spectral analysis, acoustic analysis, onomatopeic human drumming, beat box, percussion imitation, query-by-drumming, audio to MIDI transcription.

Agradecimientos

Quisiera agradecer en primer lugar y de manera muy especial a todos los que han contribuido de forma amable, paciente y desinteresada en la grabación de la base de datos drhuman y a todos los que aunque todavía no lo han hecho me han brindado su interés y predisposición a hacerlo. Ha sido y es la parte más dura del trabajo, tanto para ellos como para mí y por ello es por lo que les dirijo los primeros agradecimientos. También quiero dar las gracias a todos aquellos que habiendo participado o no en la grabación me han ayudado a conocer a más gente que grabar. No quiero olvidarme de Javier González Izaguirre y Luis Andújar Vegas, compañeros de Uncreated que me han ayudado en la grabación del material de video relacionado con la base de datos, Tadeo que también me ayudó con el video y Sergio Rodríguez del Val y Juan Pedro Moragues que me ofrecieron ayuda para la elaboración de los ritmos de la base de datos. Y especialmente agradecido a Samuel Ruiz (V.R.S.), quien además de haberme prestado material desinteresadamente, me ha escuchado, aconsejado y enseñado de audio más que muchos libros.

En el ámbito académico quisiera mostrar mi agradecimiento a Doroteo Torre Toledano, mi tutor. No sólo por su tutela y consejo académico, sino por su confianza en mí y mis intereses de investigación, la libertad que me ha dejado y el trato que he recibido de su parte. También quisiera agradecer su colaboración a Daniel Ramos tanto por el consejo académico como por el apoyo personal, así como a todo el equipo de los laboratorios del ATVS, que me siguen tratando estupendamente.

Por último, aunque no menos importante, también quería mostrar mi gratitud a los miembros de mi familia y amigos que han confiado en mí y no han cuestionado las decisiones que he tomado, así como a los que me han apoyado en todo momento pese a su desacuerdo. Muy especialmente agradecido a David Gracia y Víctor M. Martín-Tadeo (Tadeo) por estar siempre que les he necesitado.

INDICE DE CONTENIDOS

1 Introducción y objetivos

1.1 Concepto de percusión humana y tipos

Entendemos como percusión humana todos aquellos sonidos de percusión o que imitan a la percusión emitidos por personas sin el empleo de instrumentos musicales de percusión o dispositivos electrónicos diseñados al efecto. Cuando hablamos de imitación de instrumentos de percusión podemos distinguir entre aquellos de altura definida (que producen notas identificables) y de altura indefinida. En este estudio nos referimos de forma genérica a los de altura indefinida, concebidos para marcar el ritmo, no para crear melodías, aunque en particular, este estudio se ciñe a la imitación de batería, por lo que también se incluyen los timbales que se consideran de altura definida.

Si nos adentramos en el ambiente musical actual (pop, rock, música electrónica…) podemos darnos cuenta de que cada vez más gente sin estudios musicales se adentra en la composición musical. Hoy en día, gracias a la tecnología (o por culpa de esta) cualquiera que tenga un ordenador en su casa tiene la posibilidad de componer piezas musicales sin demasiado esfuerzo. Una tarjeta de sonido y cualquier software de edición son más que suficientes para empezar a componer temas con un resultado en cuanto a calidad de sonido bastante aceptable. Los que más han empleado esta tecnología son sin duda los que se dedican a la música electrónica, que llevan años empleando sintetizadores, cajas de ritmos y todo tipo de emuladores virtuales para generar piezas musicales. Con el paso del tiempo, la aparición del protocolo MIDI y la mejora de los sistemas informáticos, todos los aparatos que tenía una persona dedicada a la música electrónica han ido desapareciendo para quedarse únicamente en el PC (o más habitualmente Mac). Los únicos dispositivos hardware que se conservan son interfaces que permiten al usuario comunicarse con la aplicación software. De estas interfaces las más comunes son los teclados, debido que pueden ser tocados con mayor o menor intensidad, controlando ese parámetro en la aplicación de instrumento virtual, y en el caso concreto de la percusión, los pads.

Sin embargo en otros ambientes musicales, como el rock o el pop, también se han establecido de forma natural medios de comunicación musicales entre los diferentes componentes de una banda. Por ejemplo, si el baterista no sabe tocar la guitarra pero quiere hacer saber al guitarrista cómo quiere que interprete una parte de la canción, éste la tarareará y seguramente el guitarrista entienda qué es lo que quiere decir. El proceso inverso es también muy habitual, es decir, el vocalista, guitarrista o bajista quieren hacer saber al batería cómo les gustaría que fuera un ritmo para acompañar a una determinada melodía. Ante esta situación, aquellos que sepan tocar la batería pueden sentarse y tocar lo que quieren que el baterista interprete. Si tanto el baterista como el otro instrumentista saben leer y escribir el ritmo en una partitura se puede recurrir a este método para hacerse entender. Pero aquellos que no dominan ninguno de estos métodos suelen recurrir a otros mucho más instintivos como realizar el ritmo pegando golpes con las manos sobre una superficie o imitarlo con sonidos onomatopéyicos de la voz (/*tu ka tu tu ka*/). Estos últimos métodos es a lo que nos referimos cuando hablamos de percusión humana y es lo que estudiamos en este trabajo.

Pues bien, lo que se pretende con este trabajo es, a continuación del mismo, a largo plazo, estudiar la posibilidad y las formas de transcribir los sonidos de percusión humana

recogidos con un micrófono a una representación musical más manejable informáticamente como puede ser MIDI.

Se pueden observar innumerables tipos de percusión humana, cada persona puede tener una forma que le resulte más fácil para imitar sonidos de percusión. En el diseño de la base de datos se ha intentado ser lo más flexible posible para ajustarnos a la realidad de la situación, pero para poder realizar un trabajo humanamente abarcable se han distinguido los siguientes tipos de percusión humana:

- *Imitación onomatopéyica*. Consiste en imitar de forma onomatopéyica los sonidos de la batería. Por ejemplo “tum”, “plas”, “ding”, “ta” serían algunos ejemplos para imitar un tom, un crash, un ride o una caja respectivamente.
- *Beat Box*. Este método también consiste en imitar la batería con la boca, pero emitiendo sonidos que no tienen porqué ser silábicos.
- *Imitación por golpes*. Consiste en imitar una secuencia, por ejemplo de bombo y caja, dando golpes con las manos sobre una mesa, las piernas o cualquier superficie.
- *Imitación con pies y manos*. Es un caso similar al anterior, consiste en imitar que estamos tocando la batería dando golpes sobre las rodillas, el abdomen o una mesa para imitar la caja o los toms y pisar sobre el suelo para imitar los golpes de bombo. Para los que saben tocar la batería es el método más natural puesto que los movimientos son muy similares.
- *Estilo libre*. Consiste en construir una batería a partir de elementos cotidianos como cacerolas, cojines, libros, cubos o cualquier cosa que al ser golpeada pueda imitar diferentes sonidos de percusión y luego tocarlos como si se estuviera tocando una batería.

1.2 Introducción a MIDI

MIDI son las siglas de *Musical Instruments Digital Interface* o interfaz digital de instrumentos musicales. Se trata de un protocolo de comunicación que apareció en el año 1982, fecha en la que distintos fabricantes de instrumentos musicales electrónicos se pusieron de acuerdo en su implementación. Aunque originalmente se concibió como un medio para poder interconectar distintos sintetizadores, el protocolo MIDI se utiliza actualmente en una gran variedad de aplicaciones: grabación musical, cine, TV, ordenadores domésticos, presentaciones multimedia, etc.

Dado que este protocolo es bastante eficiente en cuanto a enviar cantidades de datos relativamente grandes a una velocidad respetable, se ha convertido en un elemento de gran utilidad para compositores, educadores, programadores y aficionados intentando crear música con varios instrumentos. Con la ayuda de un ordenador o un secuenciador hardware, se pueden crear arreglos multipistas, líneas o partes instrumentales, etc. Además la edición es muy fácil y permite alterar la velocidad de reproducción y la altura tonal independientemente.

Generar sonido a partir de un sintetizador MIDI en vez de hacerlo partiendo de un sampler tiene muchas ventajas. La primera de ellas es que se necesita una gran cantidad de espacio de almacenamiento para guardar el audio muestreado (por ejemplo en forma de archivos WAV o AIFF) que necesita un sampler. Se necesitan unos 10 Mb de espacio en disco para almacenar 1 minuto de audio estéreo muestreado en calidad CD (16 bits y 44,1kHz). En

comparación, los archivos de datos MIDI tienen un tamaño insignificante. Una secuencia MIDI típica utiliza sólo unos 10 Kb por minuto.

El archivo MIDI no contiene datos de audio muestreado, sino más bien una serie de instrucciones que el sintetizador u otro generador de sonido utiliza para reproducir el sonido en tiempo real. Estas instrucciones son mensajes MIDI que indican al instrumento qué sonidos hay que utilizar, qué notas hay que tocar, el volumen de cada una de ellas, la duración, etc.

En el caso de disponer de un ordenador, con la incorporación de una interfaz MIDI es posible conectar todo el sistema MIDI al ordenador y que éste actúe como componente central del estudio. Los programas secuenciadores actuales permiten manejar todo el estudio MIDI desde la pantalla del ordenador y grabar, editar y reproducir todo tipo de creaciones musicales. En los sistemas más sencillos, ni siquiera existen módulos de sonido externos, ya que la tarjeta de sonido del ordenador incorpora uno o varios sintetizadores y la propia interfaz MIDI. Con lo que sólo es necesario conectar un teclado controlador en la entrada *MIDI In* de la tarjeta de sonido y empezar a tocar (siempre y cuando esté todo bien configurado).

Por tanto según hemos visto, MIDI ofrece una ingente cantidad de posibilidades de trabajo relacionado con la música. Describe completamente la pieza que se está tocando y ello posibilita caracterizar la pieza para emplearla como disparador de iluminación, melodías programadas o video-arte en el escenario. Para información más detallada del protocolo MIDI se puede consultar el Anexo C.

1.3 Medios de interacción musical persona-computador

Básicamente en este apartado podemos diferenciar entre dos tipos de medios de interacción, el basado en MIDI y el basado en audio. Los ordenadores de hoy en día están preparados para interactuar con cualquiera de los dos tipos mediante la tarjeta de sonido. Dependiendo de la gama de la tarjeta de sonido nos encontramos con tarjetas con un número determinado de entradas de línea, en las que se pueden conectar señales que anteriormente hayan pasado por un amplificador de línea o más comúnmente denominado previo. Un canal de una mesa de mezclas, un módulo de efectos para guitarra o bajo, un canal de salida de audio de un sintetizador o un amplificador de señal microfónica entregan a la salida niveles de señal de línea. Algunas tarjetas de gama media incluso incorporan previos preparados para micrófono dinámico, de condensador (con alimentación fantasma) o para instrumento.

Mediante este tipo de entradas, todas a niveles similares, la tarjeta de sonido realiza una conversión AD (analógico a digital) con diferentes posibilidades de frecuencia de muestreo y resolución en bits. Una vez convertida en digital, la señal es encaminada hacia un programa de edición (Pro Tools, Cubase, Logic…). Este tipo de software es capaz de manipular un importante número de señales de audio de entrada simultáneamente y realizar tanto edición no lineal como manipulación de la señal (filtrado, ecualización, efectos…) mediante plugins (programas añadidos al editor básico). De esta forma podemos registrar y modificar composiciones de todo tipo siempre que tengamos micrófonos, un previo y la tarjeta. Se puede decir que es hoy por hoy el método más simple de interacción musical entre persona y ordenador orientada a la composición musical. Incluso si se trata de sintetizadores MIDI, en muchos casos, si no tenemos una librería con esos mismos

sonidos, deberemos grabarlos como audio. Ahora bien, últimamente lo más habitual ya no es conectar el sintetizador a la entrada de audio de la tarjeta. Ahora, como el sintetizador es virtual (software), únicamente necesitamos un controlador MIDI que maneje los parámetros de dicho sintetizador virtual.

Pero no todos los medios de interacción musical entre persona y ordenador tienen la composición musical como objetivo. A nivel de usuario, cualquiera tiene archivos de música en su ordenador o CDs y vinilos en la estantería. Pues bien, una forma de interacción entre usuario y ordenador referida a la música es lo que se conoce como MIR.

MIR son las siglas de *Music Information Retrieval* o recuperación de información musical, esto es realizar búsquedas, tesauros, clasificaciones, etc., de contenido musical. Este tipo de consultas basadas en contenido pueden tener como fuente distintas formas. Hoy en día la música se presenta en forma de ficheros de audio, ficheros MIDI y partituras. Esta variedad propicia que cada dimensión musical (timbre, tempo, estructura…) sea más fácil de encontrar en alguna de las diferentes formas, pero si únicamente disponemos de una de ellas será difícil encontrar todas las dimensiones. La conversión de MIDI a partitura, de MIDI a audio y de partitura a MIDI y de ahí a audio es una tarea que prácticamente está resuelta, por lo que parece ser que MIDI es el puente entre las diferentes formas de expresión musical. El paso que nos falta por resolver de forma satisfactoria es el de pasar de audio a MIDI, lo que nos lleva a los objetivos de este trabajo.

1.4 Objetivos

El objetivo principal de este Trabajo de Fin de Máster y de su continuación es encontrar, mediante el análisis de las características acústicas de los eventos de percusión humana, rasgos y características que orienten la elección de sistemas automáticos de reconocimiento y clasificación de eventos de percusión humana.

Para ello, en primer lugar se ha confeccionado el diseño de una base de datos con el objetivo de recoger una amplia variedad de muestras de distintos tipos de percusión humana. De este modo veremos las distintas formas que tienen las personas de interpretar este tipo de percusión y podremos observar, dentro de cada tipología los siguientes aspectos:

- Las formas más habituales de imitación y parámetros comunes.
- Variabilidad del usuario y entre diferentes usuarios y formas de abordarla.
- Qué casos se pueden considerar errores y qué casos no.
- Errores más comunes y cómo enfrentarnos a ellos.

El objetivo a largo plazo será hacer que los sistemas de reconocimiento automático sean capaces de reconocer eficientemente el tipo de sonido de percusión que el usuario pretende imitar para poder transcribirlo a un lenguaje musical. Por lo que se ha visto anteriormente, parece que actualmente el protocolo MIDI sería el lenguaje que más posibilidades ofrecería a los potenciales usuarios de este tipo de sistemas, tanto para MIR como para aplicaciones de creación y producción musical.

2 Estado del arte

2.1 Estado del arte en MIR

Al contrario que en otros tipo de contenidos, como en el caso de texto, donde la búsqueda puede ser realizada buscando coincidencias entre un texto dado y los textos presentes en bases de datos, o como en el caso de imágenes y video, donde la búsqueda puede ser realizada buscando descriptores y meta-datos del contenido multimedia, en música no es tan fácil. Podemos realizar una búsqueda por el título de la canción, pero no siempre lo recordaremos. También podríamos intentar buscarla indicando el género y las características de la canción, pero por poner un ejemplo sería imposible discriminar entre las cientos de baladas rock que se han compuesto. Sin embargo sí que recordaremos la melodía principal, el ritmo de ésta o puede que incluso la letra del estribillo. Es por esto que la búsqueda de información musical debe ser realizada basada en contenido.

Hay diversos tipos de búsquedas que se pueden hacer referentes a contenidos musicales, y estas búsquedas estarán principalmente motivadas por el papel del usuario. No buscará lo mismo un usuario casual que un DJ o que un musicólogo. Por ello primero debemos considerar las necesidades de cada tipo de usuario para después recuperar información musical adaptada a dichas necesidades. Según (Orio, 2006) se puede hacer la siguiente clasificación en función del tipo de consulta.

Un usuario casual o doméstico, que simplemente requiere de contenidos musicales para disfrutar de la música y conocer nuevos grupos acordes a sus preferencias realizará búsquedas según los siguientes planteamientos:

- *Encuentra una canción que suene como esto*: estaríamos en un caso de *Query-By-Example* (QBE), el usuario puede tararear o silbar un fragmento de un tema que ha escuchado y el sistema deberá encontrar un tema o una selección de temas que sean similares a la melodía emitida por el usuario.
- Me gustan estas canciones, *encuéntrame más canciones que suenen como esto*: en este caso el usuario proporciona al sistema un conjunto de temas que le gustan y el sistema deberá filtrar y recomendar al usuario una serie de temas que el usuario no tiene porqué conocer pero deberían estar acordes con sus gustos. Son sistemas de recomendación y filtrado, se basan en una canción y buscan otras similares, de estos se derivan sistemas de generación automática de listas de reproducción.
- *Organiza mi colección de música*: en este caso el sistema deberá clasificar la música que el usuario tenga almacenada, y establecer vínculos entre las distintas piezas para que el usuario pueda navegar intuitivamente entre los distintos temas cuando quiera buscar una canción en particular. Podemos encontrar sistemas para búsqueda rápida o *browsing* mediante estilos o mediante clasificación con grafos

Un usuario profesional como pueda ser un DJ, alguien encargado de buscar los temas para la banda sonora de una película o un crítico que necesita recopilar canciones para realizar una comparativa pueden plantearse las siguientes preguntas:

- *Estoy buscando una banda sonora para…*: una película, un anuncio televisivo o un hilo musical de un restaurante.

- *Búscame canciones que tengan un ritmo como este*: esta es la típica consulta que realizaría un DJ para buscar el siguiente tema a pinchar en su sesión para que ambos temas casen y tengan un enlace natural.

Musicólogos, teóricos de la música y músicos tendrían otro tipo de necesidades de búsqueda. Por poner un ejemplo, un musicólogo realizando un estudio de la música renacentista puede necesitar encontrar patrones de tonalidad o tempo, o necesitar establecer relaciones estructurales entre las músicas de dos regiones geográficas próximas.

La forma más habitual, y además la más intuitiva, de realizar una consulta basada en contenido es la denominada *Query-By-Humming* (Paiva et al., 2005 y Ghias et al., 1995) consistente en tararear o silvar una melodía para que el sistema encuentre canciones con melodías similares. El método seguido comienza por extraer una serie de posibles candidatos a pitch de cada trama de audio, posteriormente se construyen las trayectorias de los pitch que más destaquen. Generalmente, como no se espera que el usuario afine perfectamente las notas, se realiza una primera estimación de las notas que se tratan de interpretar y después se emplea como descriptor de la melodía la diferencia tonal entre notas sucesivas, es decir como una *melodía diferencial*. Las notas muy cortas, y notas poco destacadas y relacionadas armónicamente con otras habitualmente son eliminadas en previsión de que sean fallos o interpretaciones de los arreglos de la canción, que se desvían de la melodía principal.

Otra forma de petición de búsqueda propuesta para DJs es *Query-By-Beat-Boxing* (Kapur et al., 2004), en la que el usuario imita la percusión del tema que está buscando mediante beat box y el sistema busca patrones de percusión similares. En este artículo únicamente se contemplan bombo, caja y charles, describen como mejor método para separar estas 3 clases la tasa de cruces por cero, después extraen el tempo y los golpes de cada elemento en dicho tempo y realizan comparaciones con esto. Lamentablemente, al no tener etiquetado el audio, simplemente informan de un buen comportamiento desde el punto de vista cualitativo. Es una buena referencia a tener en cuenta en el sentido de que han conseguido implementar un sistema como uno de los que se persigue tras el análisis realizado en este TFM.

Una vez tenemos la señal musical representada para poder realizar una búsqueda hay distintos tipos de sistemas para realizar dicha búsqueda. Pueden ser basados en índices mediante N-gramas (Downie y Nelson, 2000) o frases musicales (Melucci y Orio, 1999). También hay aproximaciones basadas en coincidencia de secuencias mediante modelos ocultos de Markov (Shifrin et al., 2002) o *Dynamic Time Warping* (Hu et al., 2002). Otros métodos de búsqueda están basados en métodos geométricos como pianoroll (Fig. 2-1), que es como nos solemos encontrar representados los archivos MIDI, o representaciones similares (Typke et al., 2004) donde las melodías polifónicas son representadas como puntos ponderados en un plano, en donde el eje de ordenadas representa la nota, el eje de abcisas el instante de ejecución de la misma y el peso es la duración de la nota.

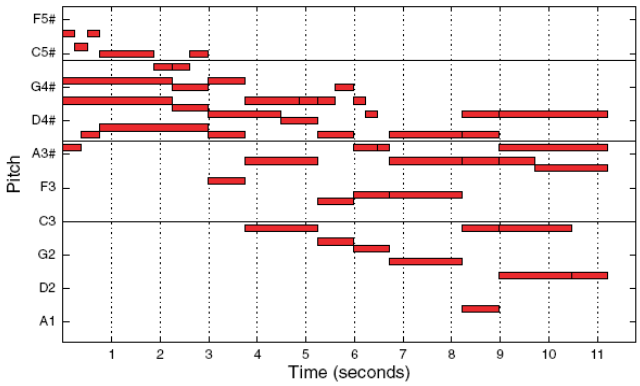


Figura 2-1: Representación de un Pianoroll (Orio, 2006)

Otros sistemas desarrollados en este ámbito son sistemas de recomendación y filtrado, que suelen estar basados en el timbre de una grabación y buscan timbres similares en otras canciones (Li et al., 2004), de estos se derivan sistemas de generación automática de listas de reproducción. También podemos encontrar sistemas para búsqueda rápida o browsing mediante estilos (Fig. 2-2) o mediante clasificación con grafos.

Genre	Subgenre
Classical	Choir
	Orchestra
	Piano
	String quartet
Country	
Disco	
HipHop	
Jazz	BigBand
	Cool
	Fusion
	Piano
	Quartet
	Swing
Rock	
Blues	
Reggae	
Pop	
Metal	

Figura 2-2: Distintos géneros y subgéneros (Li et al., 2004)

Para realizar este tipo de consultas, que deben estar basadas en contenido, debemos extraer de dicho contenido algún tipo de descriptor. Según (Downie y Nelson, 2000), los sistemas se basan en 3 características básicas del sonido musical:

- *Pitch*: o frecuencia fundamental es la frecuencia dominante en un instante dado de la canción, gracias a esto podremos identificar la nota que se está interpretando. El pitch puede ser más grave o más agudo.

- *Intensidad*: es la energía con la que la nota es ejecutada y está relacionada con la amplitud de la señal. Habrá sonidos de intensidad baja y alta.
- *Timbre*: es la característica espectral que diferencia un instrumento de otro o simplemente que permite percibir dos sonidos distintos para un mismo pitch.

A partir de la combinación de estas 3 características fundamentales podemos caracterizar las facetas que serán útiles para la búsqueda de contenidos musicales:

- *Timbre*: depende de la percepción de la calidad de los sonidos, la cual está relacionada con los instrumentos musicales empleados, posibles efectos de audio y diferentes técnicas de interpretación. Todos los gestos interpretativos contribuyen a la percepción final del timbre de la grabación.
- *Orquestación*: es debida a la elección de los compositores de los instrumentos, voces o elementos de percusión que van a interpretar una obra.
- *Acústica*: puede ser considerada como una característica especial del timbre que incluye las características acústicas de la sala donde se ha ejecutado la obra, el ruido de fondo, el post-procesado, la producción y la mezcla de la grabación. Es una dimensión que recoge los procesos realizados con la finalidad de satisfacer las expectativas de calidad, ambientación y estilo.
- *Ritmo*: está relacionado con la repetición periódica, con posibles pequeñas variaciones de un patrón temporal de sonidos. Esto sonidos no tienen porqué tener un pitch reconocible, ya que la percepción del ritmo está relacionada con los tiempos de ejecución de los sonidos. Por esto los sonidos percusivos y sin un pitch específico son usados como los mejores portadores de información rítmica.
- *Melodía*: está compuesta por una secuencia de sonidos con el mismo o similar timbre, con un pitch que varía dentro de ciertos rangos predefinidos de frecuencia (notas). La dimensión musical suele ser transportada por la voz cantada o por instrumentos monofónicos. Por si misma, la melodía es una característica multidimensional, ya que diferentes melodías pueden ser ejecutadas simultáneamente por diferentes instrumentos.
- *Armonía*: es la organización a lo largo del eje temporal de sonidos simultáneos con un pitch reconocible cada uno de ellos. Puede ser llevada a cabo por instrumentos polifónicos, por un grupo de instrumentos monofónicos o puede ir de forma implícita en la melodía.
- *Estructura*: es una dimensión horizontal con una estructura temporal diferente de las anteriores, estando relacionada con características macroscópicas como repeticiones, entrelazado de estrofas con estribillos, cambios en el ritmo o la tonalidad o eventos similares.

Ahora que conocemos las características que pueden servirnos de descriptores, podemos ver a continuación algunos artículos donde describen la metodología empleada para la extracción de los mismos. Por ejemplo en (Goto y Muraoka, 1998) el sistema propuesto fue uno de los primeros destinado a la extracción de patrones rítmicos a partir de archivos de audio de música popular (pop, rock, funky…) en tiempo real. Se basa en el reconocimiento de una estructura rítmica jerárquica en la que se reconoce por una parte el compás, por otra parte medios compases y por último cuartos de compás. Se asume que la estructura del compás de la canción es de 4/4 y que el tempo está entre 61 y 185 bpm. Para reconocer puntos clave en la definición del patrón rítmico se tienen en cuenta los instantes de ejecución de las notas, los cambios de acordes en el acompañamiento y los patrones de batería.

El tipo de sistemas como el descrito anteriormente también se emplea como primera fase de otros sistemas de reconocimiento, por ejemplo de acordes (Bartsch y Wakefield, 2005), de esta manera, una vez extraído el tempo parametrizamos cada una de las unidades de métrica musical (en el caso de 4/4 se dividiría cada compás en 4 unidades y se parametrizaría cada una de ellas), así el enventanado de la señal es dependiente del ritmo y el tempo. Luego si se quieren reconocer patrones de melodía se pueden emplear ventanas de menor tamaño, pero relacionadas con el ritmo como unidades de parametrización. Para el reconocimiento de acordes es habitual emplear unidades completas, cada una de las cuales es representada mediante un vector de 12 dimensiones (Fujishima, 1999). Cada uno de estos vectores se extrae de los coeficientes de energía de cada una de las bandas de un banco de filtros adaptado a las octavas de cada una de las 12 notas (*Croma*). De modo que para la nota Do se empleará un banco de filtros donde exista un filtro en cada octava centrado en la frecuencia de la nota para cada octava. La representación en el tiempo de estos coeficientes es lo que se conoce como cromagrama (Fig. 2-3 derecha). Una peculiaridad del cromagrama es que la distancia entre notas no solo depende de la frecuencia sino de la nota y la octava en la que se encuentre. Así en la gráfica de la izquierda de la figura 2-3 (Hélice de Shepard) podemos observar que para la nota C (Do) las componentes registradas serán las marcadas con un círculo. Por otra parte el ancho de banda de cada uno de los filtros de cada banco de filtros de cada nota del croma será variable, siendo más estrechos los filtros en bajas frecuencias que los filtros en altas frecuencias.

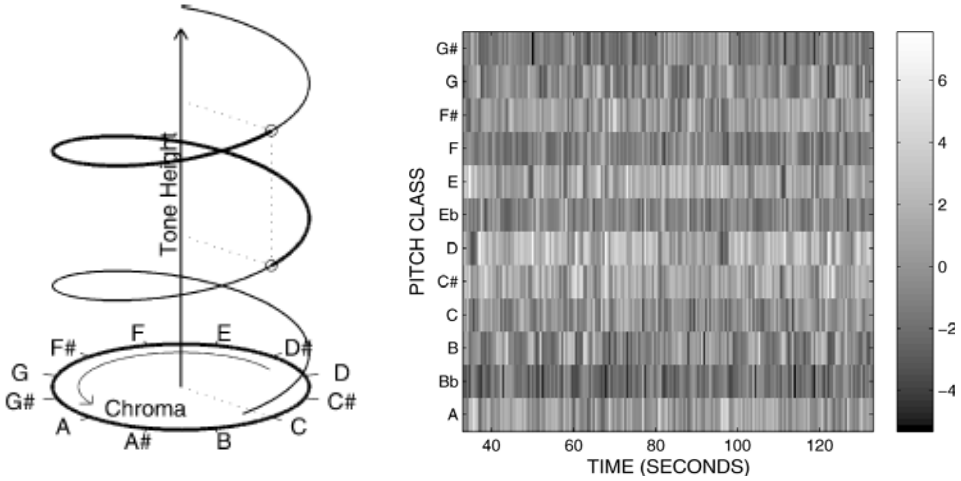


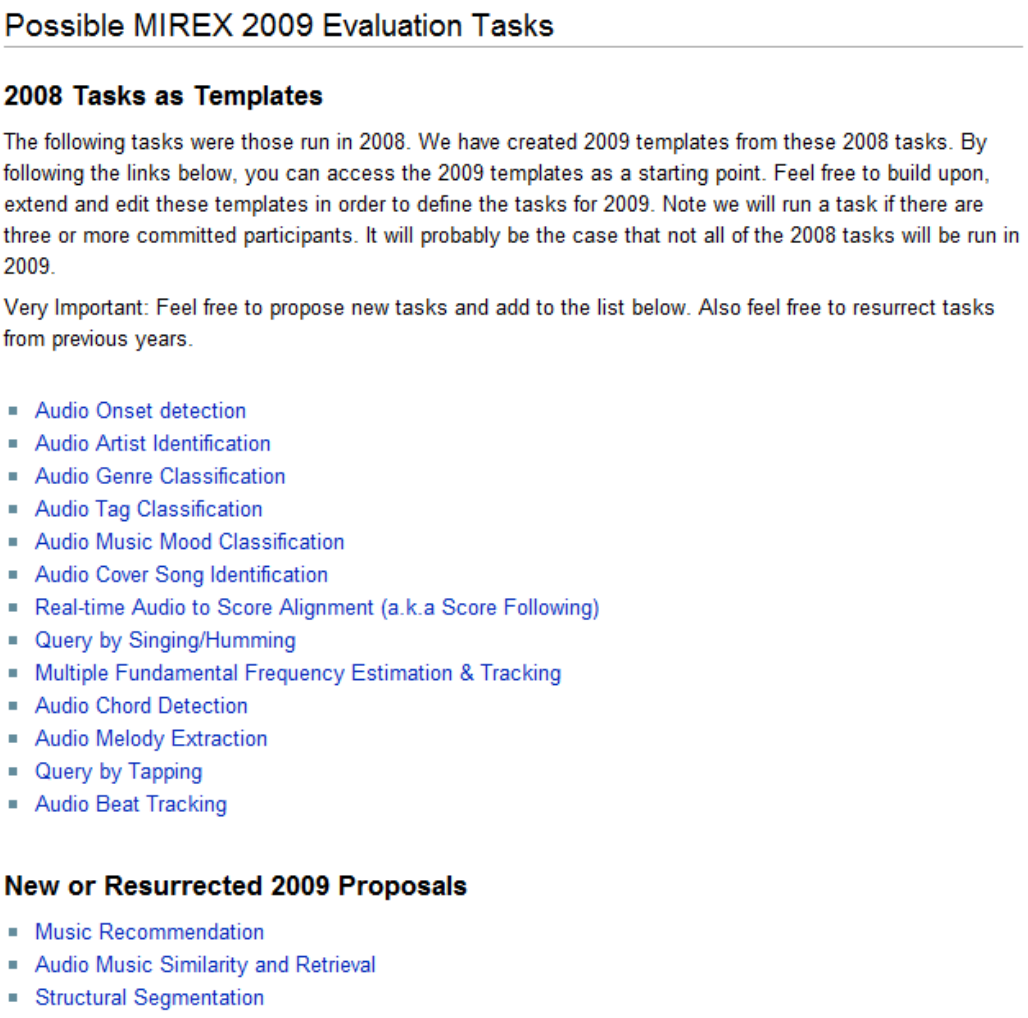
Figura 2-3: Hélice de Shepard y cromagrama (Bartsch y Wakefield, 2005)

Basados en este tipo de sistemas de reconocimiento de acordes se puede derivar un sistema que reconozca la tonalidad de la pieza analizando las notas de los acordes principales. La tonalidad de una pieza musical está caracterizada por la tónica y el modo, por ejemplo una pieza en Do (tónica) mayor (modo) estará compuesta por las notas Do, Re, Mi, Fa, Sol, La y Si. Y en los acordes construidos en esta tonalidad tendrán especial relevancia la tónica, subdominante y dominante. En (Gómez y Herrera, 2004), por ejemplo, se tienen en cuenta piezas en una única tonalidad con 12 tipos de tónicas (croma) y 2 tipos de modos (mayor y menor), dando lugar a 24 tipos de tonalidades.

Tal vez una de las características musicales más buscadas y más útiles en MIR es la melodía principal. Uno de los métodos más referenciados es el empleado en (De

Cheveigné y Baskind, 2003). Simplemente descompone la señal en su componente periódica y aperiódica. La clave del método es que es mucho menos costoso computacionalmente que realizar una transformada de Fourier. El problema no es muy difícil de resolver para melodías aisladas, pero en el caso de grabaciones con varios instrumentos (Paiva et al., 2005) hay que barajar diversas trayectorias de pitch y después eliminar posibilidades siguiendo diversos criterios. Por ejemplo las notas muy cortas, y notas poco destacadas y relacionadas armónicamente con otras son eliminadas, basándose en la hipótesis de que la línea melódica principal no debería tener cambios abruptos, cuando se encuentra un cambio de este tipo, las notas discordantes son sustituidas por otras que cuadren con el contorno de la melodía, etc.

Con el objetivo de mejorar la tecnología existente y favorecer el intercambio de ideas, desde 2004 se vienen realizando anualmente evaluaciones competitivas de MIR, las evaluaciones MIREX (*Music Information Retrieval Evaluation eXchange*) (MIREX, 2009). Una particularidad de estas evaluaciones que las diferencia de otras es que todos los potenciales participantes pueden proponer nuevas disciplinas, los protocolos de evaluación y las bases de datos con las que realizar las evaluaciones, aunque claro está, la última palabra la tienen los organizadores. Como ejemplo se muestran las tareas propuestas por la evaluación MIREX 2009 (Fig. 2-4).



Possible MIREX 2009 Evaluation Tasks

2008 Tasks as Templates

The following tasks were those run in 2008. We have created 2009 templates from these 2008 tasks. By following the links below, you can access the 2009 templates as a starting point. Feel free to build upon, extend and edit these templates in order to define the tasks for 2009. Note we will run a task if there are three or more committed participants. It will probably be the case that not all of the 2008 tasks will be run in 2009.

Very Important: Feel free to propose new tasks and add to the list below. Also feel free to resurrect tasks from previous years.

- Audio Onset detection
- Audio Artist Identification
- Audio Genre Classification
- Audio Tag Classification
- Audio Music Mood Classification
- Audio Cover Song Identification
- Real-time Audio to Score Alignment (a.k.a Score Following)
- Query by Singing/Humming
- Multiple Fundamental Frequency Estimation & Tracking
- Audio Chord Detection
- Audio Melody Extraction
- Query by Tapping
- Audio Beat Tracking

New or Resurrected 2009 Proposals

- Music Recommendation
- Audio Music Similarity and Retrieval
- Structural Segmentation

Figura 2-4: Tareas propuestas para MIREX 09

2.2 Evolución histórica de la síntesis musical

La historia de la síntesis artificial de sonidos musicales, origen de los protocolos actuales de representación musical, se remonta a fechas tan tempranas como 1860, año en que Helmholtz, pionero en el uso de dispositivos eléctricos para generar sonidos, diseña un conjunto de osciladores electromecánicos. Estos dispositivos generaban sonidos simples senoidales, es decir tonos puros.

En 1897, Thaddeus Cahill desarrolla un sistema de generación de sonido electrónico denominado *Dynamophone* o *Telharmonium*. El instrumento se basaba en el uso de dinamos modificadas eléctricamente, produciendo corrientes alternas de diferentes audiofrecuencias. Estas señales pasaban mediante un teclado y un panel de control a una serie de receptores telefónicos provistos de tonos acústicos especiales. El aparato completo pesaba cerca de unas 200 toneladas con una longitud de casi 20 metros.

En 1920, Leon Theremin, inventa y desarrolla un instrumento musical electrónico, el *Theremín*, que se controla mediante la posición de las manos del intérprete. Este instrumento generaba un sonido abstracto, no encontrado en la naturaleza, que cautivó a algunos compositores los cuales incluso llegaron a escribir algunas obras para dicho instrumento en concreto.

El piano *Neo-Bechstein* inventado en 1931, era todavía un piano acústico modificado para capturar las vibraciones producidas naturalmente y someterlas a modificaciones y amplificaciones electrónicas. El primer intento más serio de un instrumento electrónico llega con el Órgano Hammond, que aparece en 1935 de mano de Laurens Hammond y gana rápidamente un puesto de honor en la historia de los instrumentos musicales sintéticos por su peculiar, aunque no enteramente auténtica, calidad de sonido. El principio de generación del sonido se basaba en la rotación de piezas circulares en un campo magnético y utilizaba técnicas de síntesis aditiva.

Como los primeros osciladores de Helmholtz, todos estos sintetizadores pioneros eran sintetizadores analógicos basados en válvulas electrónicas. Con la llegada de los transistores en 1950, se revolucionan las técnicas de síntesis. Los sonidos obtenidos con estos primeros dispositivos eran muy limitados y muy poco naturales, de modo que eran empleados como banda sonora en películas de corte futurista y en determinadas demostraciones de vanguardia. Debido a lo limitado de sus posibilidades en aquel momento, los dispositivos eran empleados únicamente en algunos conservatorios y especialmente en centros de investigación.

Basándose en la tecnología de transistores, Harald Bode desarrolla el *Melochord* en 1961, el primer sintetizador controlado por tensión y en 1964 Robert Moog presenta el primer sintetizador que tuvo impacto comercial, el *Moog*. En 1968 Wendy Carlos publicó su *Switched-on Bach*, que era una pieza de Johann Sebastian Bach interpretada con el Moog. Fué el primer éxito en las listas comerciales íntegramente interpretado de forma sintética.

En los años 70, el aumento de la demanda provoca el nacimiento de nuevas compañías dedicadas a la síntesis de sonido, alguna de las cuales llega hasta nuestros días como *Oberheim, E-mu* y *Roland Corporation*. En 1960, John Chowning comienza a investigar detalladamente las características de los sonidos de frecuencia modulada usando síntesis computacional, pero no es hasta 1973, con la aparición de su famoso trabajo *Síntesis de*

espectros complejos de audio mediante Modulación en Frecuencia, cuando se presta atención a la posibilidad de sintetizar timbres instrumentales con la adecuada combinación de parámetros FM. En 1976 aparecen los primeros instrumentos *Yamaha* basados en esta técnica de síntesis. Estos sintetizadores eran enormes y muy costosos, poco a poco fueron reduciendo su tamaño, pero para realizar las distintas conexiones había que emplear múltiples canales de envío con la consiguiente maraña de cables entre módulos.

En 1975 y paralelamente, la *New England Digital Corporation*, produce el prototipo de un sintetizador digital, comercializado posteriormente con el nombre de *Synclavier*, uno de los más avanzados comparado con sus antecesores. En 1979, se introduce en el mercado el *Fairlight CMI*, primer instrumento con teclado de 6 octavas y controles de pedal. Permitía diferentes tipos de síntesis como síntesis aditiva, sustractiva e incluso PCM. En el mismo año, *Digital Music Systems* produce el primer chip integrado para síntesis de sonido, el DMX-1000, a partir de este momento las tecnologías LSI (*Large Scale Integration*), han ido reduciendo estos DSPs al tamaño de un chip de silicio, apareciendo instrumentos mucho más eficientes y flexibles. La era digital ayudó definitivamente a reducir el tamaño de los sintetizadores, pero el empleo de estos estaba condicionado a bastos manuales de programación, completamente distintos para cada dispositivo en concreto.

Tras varios intentos de muchas marcas por establecer un estándar de comunicaciones para los sintetizadores, fue en la exposición de la NAMM (*National Association of Music Merchants*) de junio de 1982 donde apareció por primera vez el protocolo MIDI fundamentalmente desarrollado por colaboración de marcas japonesas.

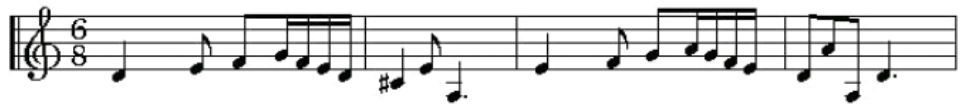
Durante la década de los 80 aparecen los instrumentos electrónicos *Casio* y *Kawai* que acaparan el mercado de los sintetizadores de gama media. En esta década, los grandes avances en tecnología digital producen modelos de síntesis más realistas y baratos. Durante esta década, el avance más significativo en los modelos de síntesis viene marcado por la evolución de las tarjetas de sonido de los PCs. Desde las primeras tarjetas que aparecen en el mercado como la *AdLib* o la *SoundBlaster* que implementan síntesis FM, la tecnología de las tarjetas comienza a orientarse hacia la síntesis PCM como la *Turtle Beach Multisound* o la *Gravis Ultrasound* (GUS), primer intento de fabricar un sampler para el mercado doméstico.

En la década de los 90, los modelos computacionalmente más costosos como los modelos físicos, comienzan a ser realizables y comercializables a nivel de usuario. El primer sintetizador comercial que implementa esta tecnología es desarrollado por *Yamaha* en 1971, tras un acuerdo con la universidad de Stanford, pionera en el desarrollo de la síntesis por guiaonda digital. El sintetizador recibe el nombre de *Yamaha VL1*.

En los años posteriores la mejora en la eficiencia de los microprocesadores para ordenadores personales, el abaratamiento e incremento de la velocidad y capacidad de la memoria de proceso y de los sistemas de almacenamiento digital masivo, acompañada de sistemas software con menor coste computacional, han permitido que todos los complejos procesos que realizaban los aparatosos sintetizadores ahora sean realizados por un ordenador y además mientras maneja otras tareas. Ahora no es necesario obtener un sintetizador, ya que seguro que encontraremos su versión software en forma de plugin (sintetizador software) para los principales programas de edición de audio.

2.3 Estado del arte en representación musical

Las representaciones musicales no sólo se han empleado para fines de composición, también han sido empleadas en MIR. Entre los primeros intentos de representación melódica digital están los que caracterizan simplemente la nota y la duración. Para caracterizar la nota simplemente se muestra la frecuencia fundamental de la misma y para caracterizar la duración se hace en forma de la figura básica del tipo de compás. En el ejemplo de la figura 2-5 el compás es de 6/8, lo que quiere decir que en cada compás hay 6 corcheas, con lo cual cada corchea se representa como "1", las negras como "2", las semicorcheas como "0,5", y así sucesivamente, representando la duración de cada nota como su relación proporcional con la nota patrón del compás. Sin embargo aunque estos parámetros pueden ser válidos para trascripción de partituras a música o de música a partitura, no parecen muy adecuados para realizar comparaciones, ya que si se realiza una consulta de tipo *Query-By-Humming* ésta puede estar desafinada. Por ello se realiza una parametrización que cuantifica las diferencias tonales en semitonos. Esta aproximación fue propuesta en (Nettheim, 1992).



[293.66, 329.63, 349.23, 392, 349.23, 329.63, 293.66, 277.18, 329.63, 220, 329.63, 698.46, 783.99, 440, 783.99, 349.23, 329.63, 293.63, 440, 220, 293.66]

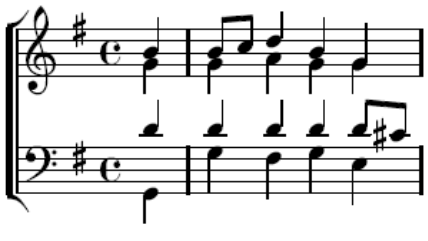
Duration sequence (multiples of the eighth note duration)

[2, 1, 1, 0.5, 0.5, 0.5, 0.5, 2, 1, 3, 2, 1, 1, 0.5, 0.5, 0.5, 0.5, 1, 1, 1, 3]

[2, 1, 2, -2, -1, -2, -1, 3, -7, 7, 1, 2, 2, -2, -2, -1, -2, 7, -1, 2, 5]

Figura 2-5: Representación de melodía propuesta en (Nettheim, 1992)

Por otra parte encontramos representaciones derivadas del protocolo MIDI que pretenden mejorar una serie de carencias que tiene el protocolo, ya no para su representación paramétrica sino para la naturalidad de su sonido. Entre ellas destaca Humdrum (Huron, 1993). Humdrum es un sistema de representación simbólica musical y un conjunto de funciones (Humdrum Toolkit) para realizar operaciones sobre dichas representaciones orientado a musicólogos y estudiosos de la música. Los archivos de datos Humdrum consisten en símbolos que representan figuras musicales (o no musicales). Estos símbolos están dispuestos en columnas denominadas spines. Cada spine representa una línea melódica de la partitura. Por ejemplo las primeras notas de una composición se representan en partitura como en la siguiente figura 2-6 a la izquierda y en Humdrum como en la tabla de la derecha.



**kern	**kern	**kern	**kern
*bass	*tenor	*alto	*soprn
*k[f#]	*k[f#]	*k[f#]	*k[f#]
4GG	4d	4g	4b#
=1	=1	=1	=1
4G	4d	4g	8b#
.	.	.	8cc
4F#	4d	4a	4dd
4G	4d	4g	4b
4E	8d	4g	4g
.	8c#	.	.

Figura 2-6: Partitura y su equivalente Humdrum (Huron, 1993)

Los objetos marcados con asteriscos son metadatos o interpretaciones que aportan instrucciones de cómo se deben leer los spines o marcas que indican la tonalidad y el modo (key) o el tempo y el tipo de compás. La primera linea de cada spine indica el formato de Humdrum en que está escrito el archivo (además de Kern, que es el más popular, hay muchos otros).

Otro tipo de representación es la propuesta por (De Cheveigné, 2000) que sugiere reconocer la melodía mediante los denominados *lattices*, que son grafos que enlazan las notas reconocidas (nodos) con una cierta probabilidad, de modo que si se da una ambigüedad considerable, el esquema soporta diferentes caminos. En cada nodo se guardan numerosos atributos de las notas y características de su relación con notas anteriores y posteriores como ligaduras o trémolos.

Por último, no podemos dejar de hacer referencia al protocolo de descripción melódica de MPEG-7 (Gómez et al, 2003) (Fig. 2-7), en el que se describe el pitch, las variaciones de contorno de éste, el tempo, la métrica, la escala y la tonalidad. El problema es que las variaciones de contorno del pitch han sido descritas con únicamente 5 valores (de -2 a 2), con lo que diferentes melodías pueden tener esta característica idéntica, luego no es especialmente apto para búsquedas de tipo *Query-By-Humming*.

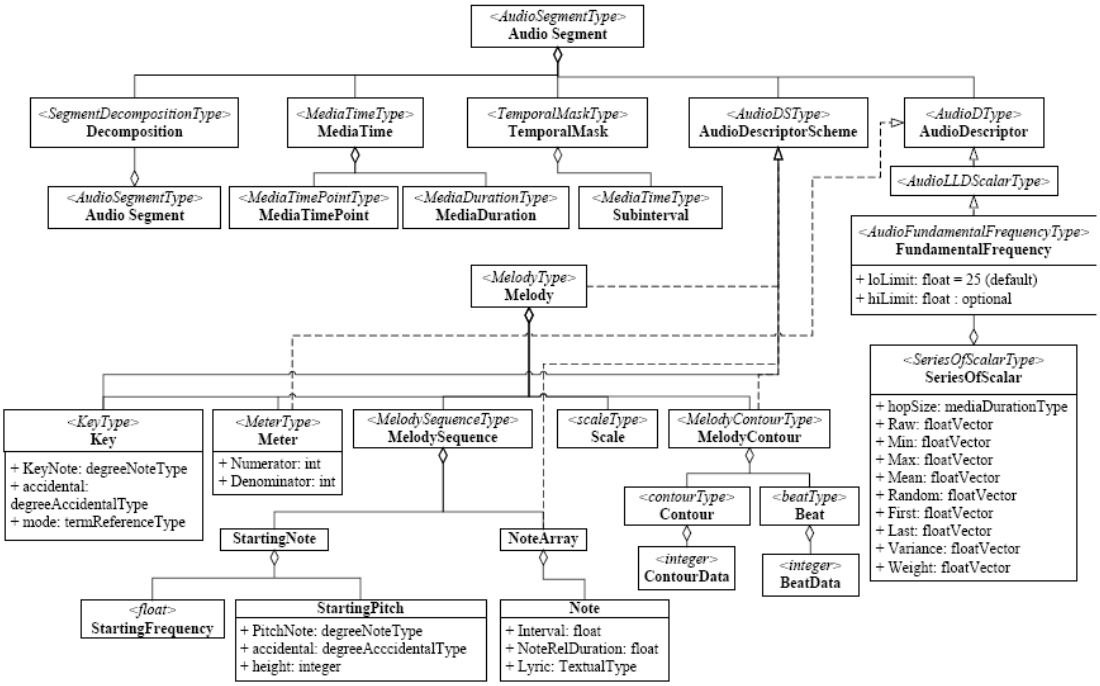


Figura 2-7: Esquema del descriptor de melodía de MPEG-7 (Gómez et al, 2003)

2.4 Estado del arte en controladores musicales

El controlador MIDI más básico es el teclado. Es un dispositivo con las mismas características que el teclado de un piano. Los de características profesionales vienen incluso con las teclas contrapesadas. Hay básicamente de 2 tipos, los que vienen con un sintetizador incorporado y los que simplemente son un controlador MIDI, que normalmente suelen venir además de con las teclas, con una palanca para modificar el la frecuencia fundamental (*pitch bend*) y otra para la modulación (puede asignarse por ejemplo a un vibrato), muchos también permiten controlar el sustain mediante un pedal que

se conecta mediante un jack. Son sensibles a la intensidad con que la tecla es pulsada, lo que traducen en un valor de velocidad de la nota MIDI. Con este tipo de controlador se pueden controlar prácticamente todos los instrumentos software.

Otro tipo de controlador son los pads. Son paneles de botones almohadillados sensibles a la presión (para variar el parámetro de velocidad). Originalmente se sacaron al mercado para controlar cajas de ritmos y otros simuladores de percusión, pero actualmente también se emplean para disparar secuencias programadas, por ejemplo en las sesiones de los DJs.

Algunos tipos de controladores consisten un conjunto de deslizadores (*faders*), potenciómetros, interruptores, secciones de transporte o paneles táctiles, a los que pueden asignarse parámetros de cualquier tipo tanto en los sintetizadores software, como en los plugins e incluso en los propios programas de edición. Por ejemplo, podemos asignar el eje vertical de un panel táctil a la velocidad de redoble de un elemento de percusión, el eje horizontal a la intensidad de la pegada (velocidad), un potenciómetro al balance de un canal del secuenciador, interruptores a los controles de grabación, solo o mute, o un fader a la ganancia de una banda concreta del ecualizador.

También empieza a ser usual ver otros controladores que tienen la forma de instrumentos como violines, guitarras, clarinetes… que son simplemente controladores, pero que por comodidad para el músico que está acostumbrado a su instrumento adoptan esa forma. Incluso se pueden encontrar adaptadores como pastillas MIDI para guitarras u otros instrumentos de cuerda, que se disimulan muy bien y que, como van captando la secuencia de notas pueden emplearse para disparar secuencias programadas, iluminación o pirotecnia en espectáculos.

Algunos ejemplos del tipo anterior incluyen instrumentos modificados como el violín de infrarrojos propuesto por (Chadabe, 1997), la familia de cuerda desarrollada por el MIT (Machover, 1992) o la guitarra (Bongers, 2000), todos modificados para servir de controladores. Sin embargo, instrumentos como la flauta descrita en (Pousset, 1992) o el instrument genérico de viento descrito en (Cook, 2001) son directamente instrumentos construidos para ser controladores, no un original modificado.

En un ámbito más experimental nos encontramos con controladores musicales que se pueden llevar puestos y que son guiados por nuestros movimientos (Paradiso y Hu, 1997). En este caso particular no controlan parámetros MIDI, sino controles de un sintetizador. Como controladores emplean una batuta digital equipada con 3 tipos de sensores. El primero de ellos es un LED infrarrojo en la punta. Al mover la batuta la punta de la misma describe un movimiento recogido por una cámara infrarroja. También lleva tiras de sensores sensibles a la presión en la empuñadura, para detectar la presión del pulgar, índice, corazón, anular y meñique combinados y palma de la mano. Por último lleva 3 acelerómetros que detectan la dirección y velocidad de los movimientos realizados por la batuta. Los 3 tipos de información son combinados para modelar la expresividad. Existen otros modelos de batuta que además incluyen un guante para modelar la expresión de la mano del director (Morita et al., 1991). Ha sido empleada con éxito en el proyecto de orquesta digital *Brain Opera* (Brain Opera, 2010).

Otros (Kapur et al, 2005) han optado por desarrollar sensores “vestibles” para modular depterminados parámetros mientras se toca. Un micrófono captura la señal del instrumento que se está tocando y la envía a un ordenador a la vez que los datos MIDI interpretados del

sensor. En el ordenador los datos son tratados para combinar ambas entradas y producir una salida que supone el audio grabado con el micrófono alterado por un efecto controlado por el sensor "vestible". Este sistema se ha probado inicialmente con el Sítar, incorporando el sensor en una diadema de unos auriculares, de manera que moviendo la cabeza se controlaban los parámetros del efecto. Otro experimento fue situar el sensor en las manos y pies de un batería, lo que además provocaba que los efectos entraran perfectamente sincronizados con la ejecución, ya que ésta se lleva a cabo con manos y pies que llevan el sensor colocado. También se ha probado con percusionistas de la Tabla (un tipo de percusión del norte de la India), obteniendo resultados similares a la batería con sensores en las manos y en la cabeza. Otro experimento a destacar consiste en colocar el sensor en la mano de un DJ realizando scrathching.

Figura 2-8: Distintas aplicaciones y posiciones del sensor

En determinadas aplicaciones musicales, como por ejemplo la danza, no es práctico que el intérprete lleve en las manos objetos como pueda ser la batuta. Hay sistemas inspirados en la filosofía de Merce Cunningham, que decía que la danza debía ser diseñada independiente de la música, incluso proponiendo realizar música para una determinada danza. Para ello hay sistemas desarrollados de telémetro láser basado en triangulación que detecta movimientos de las manos del intérprete y otros basados en sensores colocados en los zapatos (Paradiso y Hu, 1997), o por todo el cuerpo y conectados a un emisor Wi-Fi (Park et al., 2006) de manera que manejan los elementos de la escena en tiempo real. También (Paradiso et al, 1997) propone un sistema de visión artificial basado en segmentación de video de las partes del cuerpo de un intérprete o usuario. Se trata de un sistema barato, ya que solo necesita de una cámara en una habitación. Ha sido probado y aceptado con entusiasmo por parte de bailarines y coreógrafos.

El último tipo de controladores que veremos será el que más relación guarda con el TFM que estamos tratando, controladores basados en audio. En (Paulus y Klapuri, 2003) se emplea una base de datos de sonidos de percusión humana para sintetizar ritmos y realizar comparaciones orientadas a MIR. El problema de este trabajo es que únicamente se contemplan bombo, caja y charles y la base de datos se construyó grabando 7 eventos de cada elemento de batería mediante imitación onomatopéyica y dando golpes, de esta forma estas grabaciones son tomadas como muestras de un sampler disparado por un archivo MIDI. En (Guillet y Richard, 2004) se realiza un estudio parecido, la diferencia es que en este caso las muestras se tomaron después de un estudio on-line de cuáles eran las onomatopeyas más representativas de cada elemento de la batería (Fig. 2-9):

Instrument	Onomatopoeia	Frequency	Onomatopoeia	Frequency
Bass drum	**[pum]**	**36**	**[bum]**	**16**
	[tu]	17	Non-significative	29
Snare drum	**[tʃa]**	**48**	[dum]	11
	[ta]	34	[pfit]	10
	[tu]	18	[pum]	7
	[ts]	14	[bum]	7
	[ti]	12	[tok]	5
	[ʃ]	11	Non-significative	14
Hi-hat or Cymbal	**[ts]**	**48**	[ʃ]	5
	[ti]	**16**	Non-significative	29
Tom or other percussive instrument	**[tom]**	**33**	[tum]	12
	[dom]	**21**	[bum]	11
	[pum]	21	Non-significative	31
Bass drum + snare drum mixture	**[ta]**	**20**	[ts]	7
	[tʃa]	17	Non-significative	12

Figura 2-9: Onomatopeyas más empleadas para los elementos de batería estudiados (Guillet y Richard, 2004)

Una vez realizado este estudio seleccionaron las más representativas y grabaron una base de datos estableciendo dichas onomatopeyas (las que están en negrita) como las correctas para disparar el reconocedor. La base de datos está bastante completa y no es sintetizada a partir de muestras, sino que está grabada directamente del usuario, pero no tiene en cuenta un hecho muy importante, si un usuario no ha aprendido a tocar la batería, ¿porqué debería aprenderse las onomatopeyas a emplear? El sistema deja de ser intuitivo. Esta consideración es tenida en cuenta en la base de datos grabada para este TFM.

2.5 Aplicaciones informáticas comerciales

Fuera del ámbito académico se han encontrado diferentes aplicaciones funcionales similares a las que podrían resultar de los reconocedores propuestos en las conclusiones de este TFM. Básicamente hay 3 de ellas signas de mención:

- Drumtracker. Es un plugin que sirve para sustituir la percusión de una grabación de audio por sus equivalentes MIDI. La aplicación necesita que se etiquete cada uno de los golpes de percusión que aparecen en la grabación, de esta forma los sustituye por eventos MIDI y después con éstos dispara las muestras grabadas de otro plugin. El sistema solicita que se etiqueten los primeros golpes y después hace un reconocimiento del resto de la grabación, pero debe ser revisado entero. Con baterías de pocos elementos funciona bien, pero el nivel de desarrollo que tiene hace que al final se convierta en una herramienta de etiquetado manual más que en un reconocedor automático. Lo bueno que tiene es que funciona con una pista de toda la batería mezclada y es capaz de reconocer cuando suenan varios elementos a la vez (como una clase más).
- Drumagog. Se trata de otro plugin de etiquetado de batería, este funciona mejor y sí se le puede considerar como un reconocedor automático. El problema está en que solo reconoce un elemento de la batería por pista, de modo que hay que pasarle las pistas independientes y simplemente se dispara si se supera un determinado nivel. Esta característica hace que no sea muy útil para grabaciones de batería, ya que

aunque se graben los elementos con micrófonos independientes, estos siempre captan sonidos de elementos cercanos y exige cierta precisión por parte del usuario a la hora de establecer el umbral de disparo. Es muy útil en composiciones que emplean samplers de audio como los que proporciona Logic.

- Pro Trig. Es un software de triggering que permite disparar eventos MIDI a partir de lo que recoge un micrófono. Se debe ajustar el umbral, por lo que presenta el mismo problema que el anterior, pero este tiene un reconocedor espectral que le permite descartar golpes de elementos cercanos. Incluso tiene un modo de reconocimiento con un único micrófono, pero este solo discrimina entre dos clases, bombo y caja, para lo demás debemos emplear tantos micrófonos como elementos, lo cual le resta utilidad en lo que se refiere a percusión humana.

La existencia de este tipo de productos, además de darnos la oportunidad de estudiar su funcionamiento, apoya el interés de un pequeño mercado potencial por el tipo de tecnología que afronta este TFM.

3 Base de datos drhuman

Para poder realizar un estudio de las características acústicas de la percusión humana lo primero es disponer de grabaciones de las señales a analizar. En primer lugar se buscaron bases de datos ya elaboradas de este tipo y se encontraron las confeccionadas para los estudios (Paulus y Klapuri, 2003) y (Guillet y Richard, 2005). La primera de ellas contenía imitaciones onomatopéyicas y mediante golpes, pero está basada en un sampler que toma las muestras y las dispara según un archivo MIDI, con lo que no eran interpretaciones de personas. La segunda sí reunía este requisito, pero por una parte sólo recogía muestras de voz onomatopéyica, y por otra parte no contemplaba los redobles como eventos independientes de percusión. Por ello se concluyó que la mejor opción era adquirir una base de datos propia, ajustada a las posibles necesidades de estudio.

Figura 3-1: Logo de la base de datos drhuman

La base de datos ha sido llamada drhuman (Fig. 3-1) como fusión de los elementos *drum* y *human*, que referencian al concepto de percusión humana (*human drumming*). El diseño de la base de datos contempla la inclusión de 50 usuarios sin restricciones de sexo, edad ni experiencia en la percusión (si bien es cierto que quienes pueden grabar la base de datos con una mínima calidad tienen algún tipo de relación con la percusión o la música), repartidos en 5 escenarios (S0 a S4). Los escenarios de S1 a S4 se reparten todos los fraseos de batería o ritmos de la base de datos. Cada uno tiene un subconjunto de los ritmos disponibles más un ritmo común a todos los escenarios. El escenario S0 lo forman usuarios que han grabado todos los ritmos de la base de datos, es decir, los ritmos de los escenarios S1, S2, S3 y S4. Cada escenario es grabado por 10 usuarios.

El escenario S1 tiene los ritmos más fáciles, está pensado para usuarios que no se dedican ni a la percusión ni a la música, aunque sean aficionados a la misma. El escenario S2 tiene una dificultad media, los usuarios que graban este escenario tienen relación con la música aunque no son bateristas. El escenario S3 está dedicado a percusión de estilo Metal, es un tipo de percusión muy rápida y con abundante doble bombo pero con patrones relativamente simples. Los usuarios que graban esta sesión están relacionados con el mundo del Metal y aunque los ritmos son rápidos, son habituales para ellos. La sesión S4 tiene los ritmos más complejos, está pensada para usuarios que tienen conocimientos de percusión o tocan o han tocado la batería.

Aunque la base de datos está proyectada entera, a fecha de presentación de este TFM sólo está terminado de grabar por completo el escenario S0 y son con estas grabaciones con las que se ha realizado el análisis del siguiente capítulo. A continuación se explican los detalles técnicos de la base de datos, para más información de los aspectos de gestión y planificación de la base de datos puede consultar el anexo B.

3.1 Eventos de percusión y palabras

En esta base de datos se han recogido imitaciones de elementos de percusión de una batería. No se han recogido imitaciones de instrumentos de percusión oriental ni africana ni clásica, simplemente los que suelen formar parte de una batería básica, estos son: bombo (*Kickdrum*), caja (*Snaredrum*), timbales (*Toms*), charles o charleston (*Hi Hat*), distinguiendo si está abierto o cerrado, crash y ride, distinguiendo entre si es golpeado en el filo o en la campana.

Además de cada uno de los elementos de batería de los que se ha hecho imitación se han distinguido dos tipos de eventos: golpe simple y redoble. Cada uno de estos eventos diferenciados recibe el nombre de "palabra". Esto es debido a que la diferencia entre la imitación de un golpe y un redoble, por ejemplo de caja, mediante voz onomatopéyica puede ser sustancialmente distinta. Un golpe simple podría imitarse como /*ta*/ o /*pa*/, mientras que un redoble será más frecuentemente imitado como /*tarrrrrrrra*/. Con los platos ocurre de forma similar a la caja cuando se trata de golpes en la campana del ride o en el charles cerrado, mientras que al golpear el filo del ride, el crash o el charles abierto de forma rápida y repetida, el sonido suena más como un continuo, en lugar de muchos golpes diferenciados. A continuación se presenta una tabla donde se detallan las "palabras" que contiene cada uno de los ritmos a imitar. Los nombres que aparecen a la izquierda son cada uno de los ritmos a imitar

Nombre	Palabras																			
	Bombo		Caja		Tom 1		Tom 2		Tom 3		HH Abierto		HH Cerrado		Ride Campana		Ride Filo		Crash	
	Golpe	Redoble	Golpe	Redoble	Golpe	Redoble	Golpe	Redoble	Golpe	Redoble	Golpe	Redoble	Golpe	Redoble	Golpe	Redoble	Golpe	Redoble	Golpe	Redoble
FV_0	SI		SI		SI		SI		SI		SI				SI				SI	
TV_0																				
FS_0																				
BB_0																				
FV_1	SI		SI		SI		SI		SI						SI				SI	
FV_2	SI		SI									SI	SI					SI		
FV_3	SI			SI													SI	SI	SI	SI
FV_4	SI		SI	SI	SI		SI		SI	SI									SI	
FV_5	SI		SI	SI		SI		SI		SI					SI				SI	
FV_6	SI	SI	SI														SI		SI	
FV_7	SI		SI												SI	SI	SI		SI	
FV_8	SI		SI		SI	SI	SI	SI			SI		SI	SI	SI				SI	
TV_T	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
FS_T	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI	SI
FS_1	SI		SI								SI					SI				
FS_2	SI		SI	SI	SI		SI		SI		SI				SI				SI	
FS_3	SI	SI	SI			SI		SI		SI	SI				SI				SI	
FS_4	SI		SI								SI				SI				SI	
FH_1	SI	SI	SI			SI		SI	SI											
HF_1	SI	SI	SI			SI		SI	SI											
FH_2	SI	SI	SI																	
HF_2	SI	SI	SI																	
FH_3	SI		SI		SI		SI													
HF_3	SI		SI		SI		SI													
FH_4	SI	SI	SI	SI																
HF_4	SI	SI	SI	SI																
FH_5	SI	SI	SI			SI		SI												
HF_5	SI	SI	SI			SI		SI												
FH_6	SI		SI	SI																
HF_6	SI		SI	SI																
FH_7	SI	SI	SI			SI		SI	SI											
HF_7	SI	SI	SI			SI		SI	SI											
FH_8	SI	SI	SI	SI																
HF_8	SI	SI	SI	SI																

Tabla 3-1: Palabras de percusión humana recogidas en cada ritmo

3.2 Tipos de percusión humana recogidos

Con el objetivo de tener variedad de tipos de percusión humana y poder estudiar cada uno de ellos y poder compararlos, hemos recogido gran variedad de modalidades de percusión humana interpretada bajo diferentes condiciones. En la tabla 3-1 la columna "Nombre" hace referencia al ritmo que se debe imitar en cada caso y de cada forma, de esta forma podemos hacer la siguiente clasificación:

- *Imitación por voz*: Son los ritmos cuyo nombre empieza por **FV** (*Free Voice*). Se tiene que imitar la percusión mediante sonidos onomatopéyicos de la voz. La interpretación es libre, es decir cada usuario puede imitar cada elemento de la batería como lo prefiera y además lo puede hacer de formas distintas en cada ritmo. Se realizan varias grabaciones:
    - *Todo*: Se intenta imitar el mayor número de elementos de percusión a la vez (bombo, caja, platos…). Estos son ritmos reales de batería, por lo que suenan a la vez los distintos elementos, se puede optar por imitar los que se quiera en cada momento, de hecho esto puede servir para estudiar qué elementos considera la gente más representativos de cada ritmo.
    - *Por partes*: Se graba el mismo ritmo pero por partes. Se tiene una división de cada ritmo en elementos distintos adaptado a cada ritmo, por ejemplo primero bombo y caja, luego platos y luego toms… pero el usuario puede dividir los elementos a grabar en cada parte como lo prefiera. Este método permite grabar la batería completa sin complicaciones, además es la forma natural en la que se graba cuando se emplean controladores que no sean una batería electrónica, como teclados o pads.
- *Imitación por voz entrenada*: Es el mismo tipo de percusión humana que la que se acaba de explicar, son los ritmos que empiezan por **TV** (*Trained Voice*). La diferencia es que ahora se graba en primer lugar un fichero de entrenamiento **TV_T**, que incluye todos los elementos de la batería tocando diferentes figuras (golpes simples, tresillos, seisillos…) de forma que se pueda registrar como ejecuta el usuario cada uno de ellos. El motivo de generar un archivo de entrenamiento es que este tipo de imitación es fuertemente dependiente del usuario, y conteniendo diferentes figuras se entrena una amplia variedad de casos. La condición que se impone a los usuarios es que hagan de forma distinta cada uno de los elementos de la batería que imitan. A continuación se graba **TV_0**, que es un ritmo en el que deberán respetar la forma en que han imitado cada uno de los elementos en la grabación de entrenamiento. Este último ritmo es grabado, primero con todos los elementos de la batería a la vez y después por partes como en el punto anterior.
- *Beat Box*: Es el ritmo **BB_0**, es opcional, dirigido a los usuarios que sepan realizar *Beat Box* de alguna forma.
- *Estilo libre*: Son los ritmos cuyo nombre empieza por **FS** (*Free Style*). En este caso el usuario elige los elementos con los que se realizará la percusión, como cajas, libros, vasos, cojines, etc. En primer lugar se graba un archivo de entrenamiento **FS_T** y después una serie de ritmos. Para cada uno de los ritmos se realiza una primera grabación con todos los elementos de la batería a la vez y después por partes, de forma análoga como se hace en *imitación por voz*. También se realiza una grabación improvisada del usuario, con la intención de tener una evaluación subjetiva del mismo en un futuro cuando los sistemas de reconocimiento funcionen.
- *Imitación con golpes*: Son los ritmos cuyo nombre empieza por **FH** (*Free Hits*). Sólo tienen bombo, caja y en algunos casos timbales (en grabaciones añadidas a la principal que solo tiene bombo y caja). Se imitan dando golpes con las manos o las

yemas de los dedos por ejemplo en superficies como mesas, carpetas, libros, tu propio cuerpo, etc.

- *Imitación con pies y manos*: Esta es una categoría opcional, dirigida a los usuarios que sepan tocar la batería. Son los ritmos que empiezan por **HF** (*Hands and Feet*). Son los mismos ritmos que los que empiezan por **FH**, pero estos se graban golpeando con las manos a modo de caja y timbales y con los pies como si fuera el bombo, es decir, como si se tocara una batería real.

3.3 Canales

Todas y cada una de las grabaciones programadas en la base de datos han sido realizadas empleando 2 tipos de micrófono para tener cierta variabilidad de canal y así poder comprobar cómo influye éste en los sistemas de reconocimiento.

- Micrófono de bobina o dinámico *AKG D 2002* (Fig 3-2 izquierda). Se trata de un micrófono de mano profesional de gama media. Es un micrófono robusto y resistente, generalmente empleado como micrófono de mano para voz. Al ser dinámico presenta efecto proximidad, esto quiere decir que si se emplea muy pegado a la boca produce un realce en las frecuencias graves, marcando excesivamente los fonemas oclusivos, sobretodo la "p". Tiene un patrón polar cardioide y pierde mucha sensibilidad al crecer la distancia entre el micro y la fuente de sonido.
- Micrófono de condensador (membrana pequeña) *Audio Technica AT4041* (Fig 3-2 derecha). Es un micrófono de condensador de membrana pequeña, patrón polar cardioide, de calidad profesional, de gama alta. Este tipo de micrófono se suele emplear para grabar los ambientes de platos de batería, por lo que viene con un interruptor para activar un filtro paso bajo (no activo durante la grabación de la base de datos). Es un micrófono que recoge sonidos producidos en un amplio rango de distancias de forma muy nítida, por lo que además de tener más nitidez que el otro también introduce más ruido.

Figura 3-2: micrófonos empleados en la grabación de la base de datos

Durante la grabación estos micrófonos estaban quietos, asegurados mediante trípodes para evitar ruidos de movimiento. Por las características de cada micrófono, generalmente el dinámico era colocado de forma más cercana a la fuente de sonido que el de condensador. En algunos casos el dinámico provocaba saturación cuando, en grabaciones de imitación

por voz, el usuario estaba muy cerca. Es un problema real que ocurre frecuentemente con este tipo de micrófonos, de modo que dichas tomas no han sido repetidas, en previsión de que en una aplicación real nos podamos encontrar con este tipo de casos. De igual manera, con el micrófono de condensador, en determinados ambientes se han recogido leves sonidos indeseados, tampoco han sido eliminados. En ningún caso se ha efectuado ningún tipo de filtrado a la señal captada.

Es importante aclarar que se han empleado dos micrófonos, pero durante la grabación estaban posicionados muy cerca y apuntando al mismo punto, es decir que las grabaciones que se pretenden obtener son de un solo canal. No se emplean diferentes micrófonos para captar distintos elementos como es el caso de los programas de *triggering* vistos anteriormente, sino para estudiar la variabilidad de canal. El reconocimiento automático por tanto solo puede hacerse basado en grabaciones de un solo canal.

3.4 Software y Hardware empleados

La grabación se ha realizado con el programa de edición *Sonar 8 Producer Edition*, debido la comodidad de manejo, la capacidad de manejar un enorme número de pistas de audio a la vez (opción más limitada en sus competidores más directos) y al entorno de edición MIDI, muy intuitivo y visual, lo que resulta una condición necesaria de cara a la revisión de la base de datos.

Los ritmos de batería han sido compuestos (o en su caso adaptados) por el autor con el plugin *EZDrummer* de *Toontrack*, que es un plugin que asigna muestras grabadas de batería a notas MIDI, tocadas en este caso con el pad *nanoPAD* de *Korg*, y después ajustadas a mano y cuantizadas.

Los micrófonos anteriormente comentados iban conectados a la tarjeta de sonido *Fast Track Ultra 8R* de *M-Audio*, en concreto el micrófono de condensador iba conectado a una entrada con alimentación fantasma. Dicha tarjeta tiene una sección de previo incorporada en cada canal y digitaliza la señal, en el caso de la base de datos adquirida a 96KHz con una profundidad de 24 bits. Tal vez sea una calidad de audio excesiva, no se debe suponer que cualquier usuario doméstico va a tener esta calidad, pero para realizar el análisis prefiero no limitar la calidad en la fase de grabación y después, con el programa de edición se pueden exportar las pistas con diferentes calidades para simular situaciones de canal diferentes.

3.5 Condiciones ambientales de la grabación

Alguna de las grabaciones de la base de datos coincide en cuanto al lugar de grabación, pero genéricamente cada grabación se ha realizado en un lugar distinto. De esta forma tenemos diferentes ambientes de grabación, simulando condiciones reales. Dichos ambientes van desde una sala de reuniones o un pequeño estudio doméstico hasta un estudio profesional, pasando por un local de ensayo. Esto hace que haya grabaciones con algo de reverberación, otras con nada y otras con ruido de ambiente.

En determinadas grabaciones nos hemos encontrado con usuarios con gran destreza y en otras con usuarios menos hábiles, tanto a la hora de interpretar las secuencias de percusión como a la hora de cuidar el volumen y cercanía al micrófono. Por ello nos encontramos grabaciones con gran riqueza dinámica, otras con saturaciones y otras con niveles bajos.

Salvo que la interpretación no siguiera la marcada por el ritmo se ha dado por buena en cuanto a calidad técnica. La razón es que las aplicaciones potenciales que pueda tener esta investigación se encontrarán con este tipo de fallos.

3.6 Protocolo de adquisición

En primer lugar, tras ponernos en contacto con el usuario, se le envía un correo electrónico en el que se le explica en qué consistirá la grabación (Anexo C). Adjunto a ese correo se le envían unas instrucciones precisas que describen el protocolo de grabación, una serie de archivos de audio en mp3 con los ritmos que deberá imitar en cada modalidad, un enlace a un video tutorial y el consentimiento para la toma de datos que deberá traer firmado (o bien leído y firmarlo en el momento de la grabación). En las instrucciones del correo se especifica con qué modalidad se deberá imitar cada ritmo y se pide al usuario que escuche los ritmos e intente imaginar cómo los imitaría. La grabación se realizará en una única sesión.

Los ritmos que debe imitar el usuario están generados por un sintetizador virtual a partir de un archivo MIDI, por lo que dicho archivo es completamente modificable. La dinámica de grabación consiste en que el usuario escuche un par de veces el ritmo de batería y después lo interprete a la vez que lo escucha por auriculares, de este modo no tiene que memorizarlo y va siguiendo el ritmo de manera natural. Además puede elegir si quiere escuchar también la claqueta (metrónomo) o no. Como en una grabación real, si el usuario se equivoca en una parte, no es necesario volver a grabar todo el ritmo, se puede hacer un "pinchazo" y volver a grabar la parte que no ha quedado bien. Si un usuario tiene dificultades para interpretar un ritmo se le ofrece la posibilidad de bajar el tempo, ya que al tratarse de MIDI esta tarea no supone ningún problema. No obstante se intentará evitar bajar el tempo en la medida de lo posible por la sencilla razón de que además de que hay que crear un nuevo archivo MIDI adaptado a lo que se ha grabado (para el etiquetado automático), los redobles dejan de serlo y son imitados de diferente forma.

La mayor ventaja de que el usuario esté escuchando el archivo mientras lo imita es que al interpretar los eventos a la vez que en el archivo MIDI, luego éste puede ser empleado para etiquetar los eventos detectados tanto de cara a entrenamiento como para evaluar la precisión del reconocimiento automático. La secuencia de grabación de la sesión es como sigue:

1. *Imitación por voz*. Se graba cada uno de los ritmos que se detallan a continuación (primero con todos los elementos juntos y después por partes) en el siguiente orden:

Orden	S0	S1	S2	S3	S4
1	FV_0	FV_0	FV_0	FV_0	FV_0
2	FV_1	FV_1	FV_3	FV_5	FV_7
3	FV_2	FV_2	FV_4	FV_6	FV_8
4	FV_3				
5	FV_4				
6	FV_5				
7	FV_6				
8	FV_7				
9	FV_8				

Tabla 3-2: Secuencia de grabación de imitación por voz

2. *Beat box*. Esta parte es opcional, se hace si el usuario sabe hacer beat box. Consistirá en imitar el archivo **FV_0** mediante beat box.
3. *Entrenamientos*. Se graba primero mediante imitación por voz el ritmo **TV_T** y después volveremos a grabar **FV_0**, ahora denominado **TV_0** (todos los elementos juntos y luego por separado) pero esta vez se deberán imitar los elementos de la misma forma que se ha hecho en el archivo **TV_T**, es decir, si por ejemplo se imitó un golpe simple de caja como /*ta*/ se deberá hacer igual y no /*pa*/ o /*ka*/. Después se vuelve a grabar el archivo **TV_T**, ahora denominado **FS_T**, pero ahora con la batería de estilo libre. Y posteriormente, también con la batería de estilo libre, se grabarán (todos los elementos juntos y luego por separado) el ritmo **FV_0**, ahora denominado **FS_0**.
4. *Estilo libre*. En esta parte con la batería de estilo libre se grabarán (primero con todos los elementos juntos y después por separado) los siguientes ritmos (**FS_I** es un ritmo improvisado por el usuario):

Orden	S0	S1	S2	S3	S4
1	FS_1	FS_1	FS_2	FS_3	FS_4
2	FS_2	FS_I	FS_I	FS_I	FS_I
3	FS_3				
4	FS_4				
5	FS_I				

Tabla 3-3: Secuencia de grabación de estilo libre

5. *Imitación con golpes*. En esta última parte se grabarán, primero con golpes de las manos en una mesa, rodillas, pecho u objetos y después (si el usuario sabe hacerlo, opcionalmente) con las manos para imitar caja y timbales y pisando con los pies en el suelo para imitar el bombo, como si estuviera tocando una batería, los siguientes ritmos:

Orden	S0	S1	S2	S3	S4
1	FH_1/HF_1	FH_1/HF_1	FH_3/HF_3	FH_5/HF_5	FH_7/HF_7
2	FH_2/HF_2	FH_2/HF_2	FH_4/HF_4	FH_6/HF_6	FH_8/HF_8
3	FH_3/HF_3				
4	FH_4/HF_4				
5	FH_5/HF_5				
6	FH_6/HF_6				
7	FH_7/HF_7				
8	FH_8/HF_8				

Tabla 3-2: Secuencia de grabación de imitación con golpes

Los ritmos **TV_T** y **FS_T** son exactamente iguales, así como los ritmos **BB_0**, **FV_0**, **TV_0** y **FS_0** al que denominaremos *ritmo común*. Esto es para poder realizar pruebas cruzadas entre estilos de percusión humana, no se han incluido más estilos porque los demás no permiten imitar tantos elementos como estos. De igual forma los ritmos **FH_x** y **HF_x** también son iguales.

3.7 Protocolo de revisión de la base de datos

Una vez terminada la grabación de la base de datos se procederá a la revisión de la misma. La revisión está orientada a encontrar fallos de interpretación que hagan fallar un sistema de etiquetado automático mediante MIDI, para ello la revisión se efectuará en 3 etapas.

En la primera etapa se editará el audio de igual manera que se editaría en una grabación musical, esto es, recolocando partes del audio que estén fuera de tiempo siempre que esto no afecte a la continuidad natural de la grabación, es decir, si un golpe aislado ha sido imitado fuera del instante en que debería se colocará, pero si el problema está en que toda la grabación o un fragmento de varios golpes está ralentizado o descolocado se dejará como está. Además se arreglarán los "pinchazos" para que no se solapen grabaciones y las transiciones suenen naturales.

En una segunda etapa se limpiarán las pistas de manera que no se oiga nada antes de empezar la grabación ni después y limpiando también partes intermedias de duración considerable. Este también es un proceso habitual en grabaciones musicales.

Por último, las pistas de audio resultantes serán comparadas con los archivos MIDI a los que imitaban y ahora serán los archivos MIDI que no correspondan con la grabación según ciertos parámetros, los que se modifiquen para hacer coincidir los eventos de audio con el archivo MIDI. De esta forma la base de datos quedará completamente etiquetada de forma automática con los archivos MIDI mediante la nota (evento que se está imitando) y el instante en que se imita (parámetro *Note On* de MIDI).

La base de datos completa en su formato de entrega contendrá los archivos MIDI originales, los que han sufrido, para una grabación de un usuario en particular, una alteración del tempo con la especificación de qué grabación es la que tiene la alteración, así como los archivos MIDI completamente adaptados a la grabación con la especificación de a qué grabación de qué usuario está adaptado cada uno. Por otra parte se proporcionarán los archivos de audio grabados a máxima calidad en su formato original, los que han sido editados para recolocar golpes y los que han sido limpiados.

A la hora de determinar en revisión qué correcciones se deben hacer en el archivo MIDI se debe tener en cuenta el tipo de etiquetado que se va a seguir. El etiquetado automático de eventos de percusión se realizará buscando eventos de audio a partir de las etiquetas MIDI. Esto significa que se buscarán los eventos de audio más próximos a la etiqueta MIDI, lo cual implica que si en un archivo de audio que sólo debe tener bombo y caja, encontramos una imitación de charles, no deberemos borrarla en la revisión, siempre que no ocupe el espacio destinado a bombo y caja. Por otra parte los eventos de un mismo elemento (por ejemplo todos los golpes de un redoble de caja) se buscarán, con ciertos márgenes de tolerancia, dentro del espacio de tiempo comprendido entre el instante de comienzo del primer *Note On* hasta el instante final del último *Note Off* de la sucesión de eventos MIDI de un mismo elemento (misma nota MIDI).

Si en mismo archivo de audio encontramos más de una interpretación de un mismo elemento de la batería interpretados de forma suficientemente distinta, o que pueda llevar a confusión con otro elemento del mismo archivo de audio, dicha parte del archivo será considerada como un error de interpretación y será etiquetado como tal. Para archivos de audio distintos de un mismo usuario no nos fijaremos en si se ha interpretado un mismo

elemento de distintas formas, a excepción de los archivos TV_0 con respecto a TV_T y FS_x con respecto a FS_T.

3.8 Posibles pruebas de evaluación

Dado que tenemos diferentes categorías de percusión humana, algunas compartiendo ritmos en común, además de ritmos que comparten varios usuarios, la base de datos permite realizar, entre otras, los siguientes tipos de pruebas:

- Entrenamiento con TV_T y reconocimiento con TV_0 dependiente e independiente de usuario.
- Entrenamiento con TV_T y reconocimiento con FV_x dependiente e independiente de usuario.
- Entrenamiento con cualquier subconjunto de FV_x y reconocimiento con el resto dependiente de usuario.
- Generación de un modelo independiente de usuario con varios FV_x de varios usuarios y reconocimiento con el resto de FV_x.
- Generación de un modelo independiente de usuario con los FV_0 y reconocimiento con el resto de FV_x.
- Entrenamiento con FS_T y reconocimiento con el resto de FS_x dependiente de usuario.
- Entrenamiento con un FS_x y reconocimiento con el resto de FS_x dependiente de usuario.
- Entrenamiento con un HF_x o FH_x y reconocimiento con el resto de FH_x o HF_x dependiente de usuario.
- Reconocimiento de HF_x o FH_x sin entrenamiento.

4 Análisis espectral de los eventos de percusión humana

4.1 Características y variabilidad de palabra

A continuación veremos para cada tipo de percusión humana qué peculiaridades presentan las distintas palabras, qué puntos en común guardan y qué diferencias hay entre diversos estilos de imitación, así como la variabilidad en la imitación de un determinado elemento por un mismo usuario. Se puede considerar cada tipo de imitación por separado, ya que tienen características muy distintas.

4.1.1 Voz onomatopéyica

En el caso de voz onomatopéyica hay dos modalidades, voz entrenada y voz sin entrenar. En el caso de voz entrenada cada usuario tiene su propia grabación de entrenamiento y después realizan una grabación del ritmo común respetando la forma de imitación que han adoptado en el entrenamiento. En (Fig. 4-1a) podemos ver un bombo imitado en ambas grabaciones.Como podemos ver son dos sonidos que presentan características espectrales similares, por lo que se podría establecer un sistema de reconocimiento automático relativamente simple para esta modalidad si no fuera porque se observa cierta variabilidad dentro de un mismo usuario, y esto es común a todos los usuarios y en todos los elementos de la batería, dependiente de la velocidad de ejecución (Fig. 4-1b). Como podemos ver, si los golpes son muy seguidos o se trata de un redoble ya no se emplean las mismas onomatopeyas, estas varían para que el usuario llegue a dar todos los golpes de una forma más cómoda.

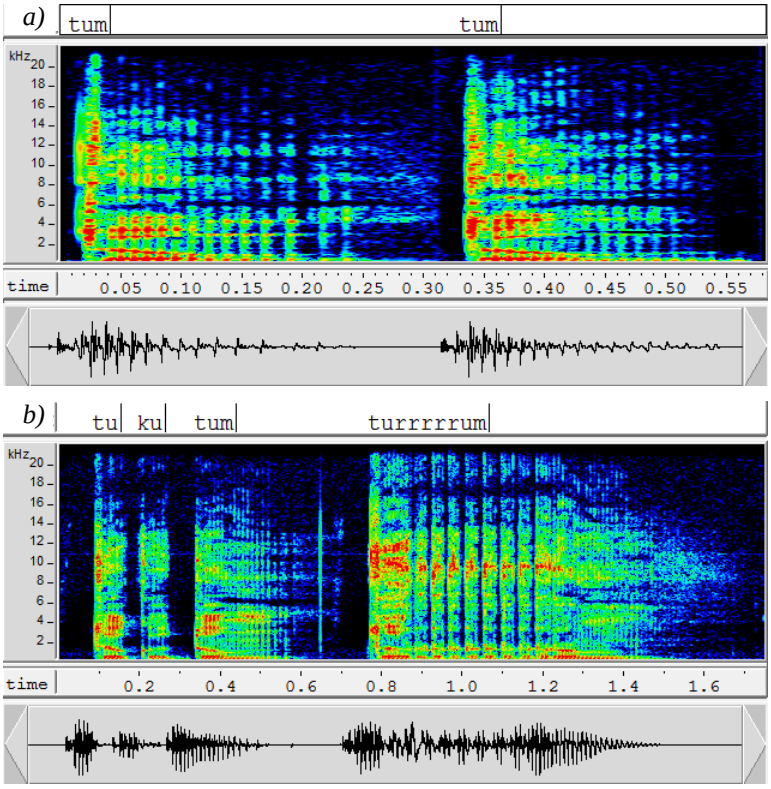


Figura 4-1: a) Bombo en entrenamiento (izq) y en ritmo común (dcha) b) imitación de un mismo elemento (bombo) a distintas velocidades

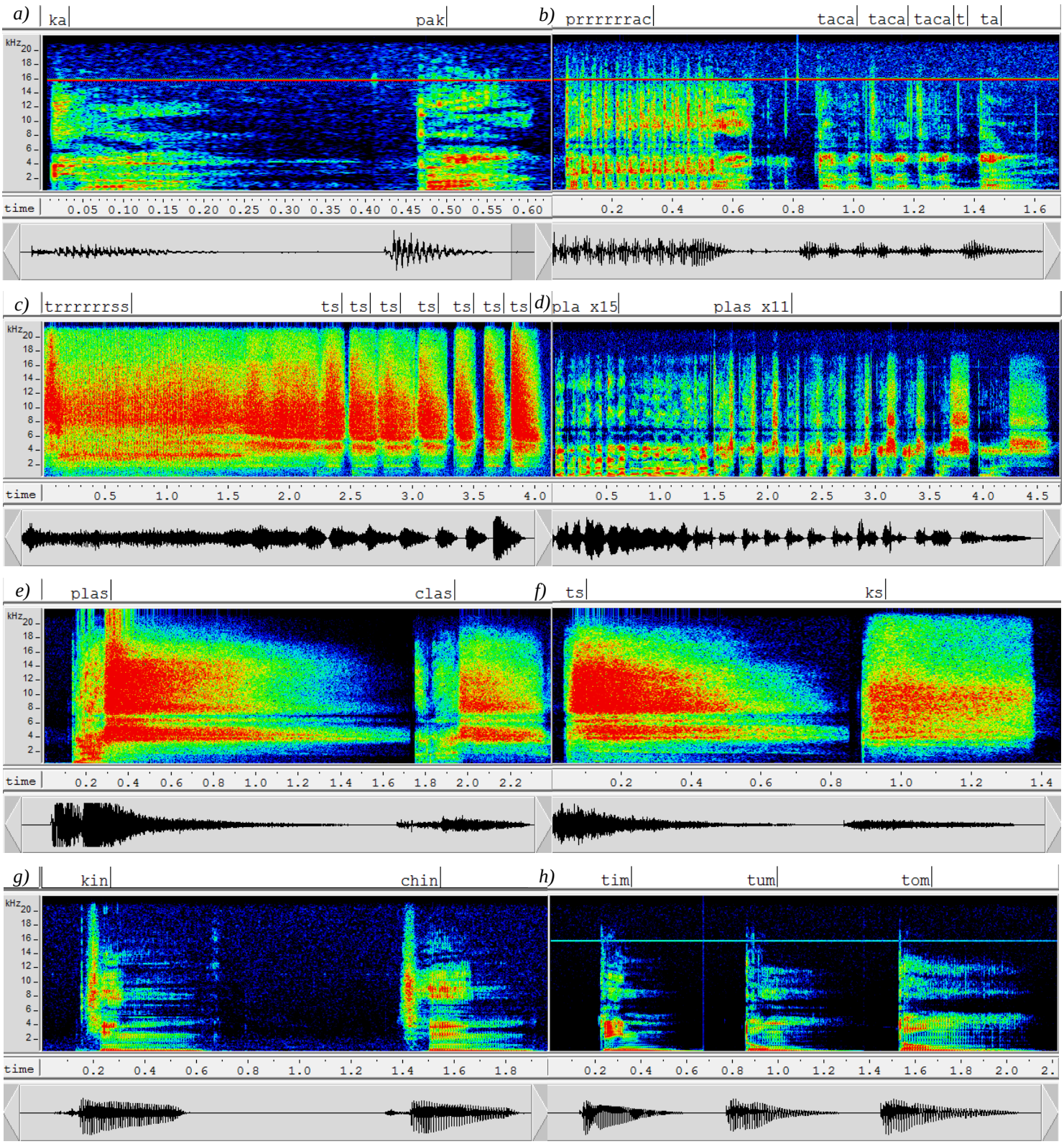


Figura 4-2: a) Imitación de un golpe de caja del mismo ritmo por usuarios distintos. b) Imitaciones con onomatopeyas diferentes de dos usuarios distintos para imitar el mismo redoble. c)y d) Dos formas distintas de imitar una serie de golpes consecutivos al crash, marcando todos los golpes (d) y considerándolos como un sonido continuo (c). e) Un mismo elemento de percusión imitado de dos formas diferentes por un mismo usuario. f) Distintas formas de imitar el charles abierto. g) Imitación del ride en el filo y la campana por un mismo usuario. h) Los tres timbales imitados mediante tres onomatopeyas distintas.

Este fenómeno se observa en todos los usuarios, aunque cada usuario emplea sus propias onomatopeyas según le son más cómodas. En los redobles la variabilidad es todavía más exagerada (Fig. 4-2b). Mientras que, por ejemplo un usuario imita un redoble de caja como /*tarrrrrra*/, otros lo hacen como /*tarararararara*/ y otros como /*tacatatacatata*/ y lo mismo ocurrirá con otros elementos de la batería.

Esto mismo ocurre con el resto de elementos de la batería interpretados a diferentes velocidades, aunque la velocidad sea lenta o se trate de golpes aislados, cada usuario lo interpreta con una onomatopeya distinta. Por ejemplo mientras que unos usuarios imitan varios golpes de platos como un sonido continuo, otros marcan cada uno de los golpes de plato (Fig. 4-2c y d) y mientras que unos imitan un golpe aislado de caja como /*ka*/ otros lo imitan como /*pak*/ (Fig. 4-2a). Incluso (y esto es más preocupante) un mismo usuario en ocasiones imita una misma palabra de diferentes formas (Fig. 4-2e).

Sin embargo para ciertos elementos de la batería, aunque sean imitados de diferente manera encontramos puntos en común. Por ejemplo, hemos observado que el golpe simple de caja es imitado como /*pa*/, /*ta*/, /*ka*/…, el bombo como /*tum*/, /*tu*/, /*pum*/, /*bum*/…, el ride como /*chin*/, /*plin*/, /*tin*/, /*pin*/…, el crash como /*plas*/, /*clas*/, /*chas*/… y el charles abierto como /*chiss*/, /*tss*/, /*kss*/… (Fig. 4-2f). Esto quiere decir que aunque no exista una forma única de imitar onomatopéyicamente un elemento de la batería, sí que hay ciertos elementos que son comunes. En los ejemplos que hemos visto, la caja se imita mediante una oclusiva seguida de una /*a*/ sin nada detrás, el bombo siempre contiene la /*u*/ y el ride la /*i*/ seguida de /*n*/, el crash suele tener una /*a*/ seguida de /*s*/ y el charles abierto termina con una /s/ larga.

El problema llega cuando onomatopeyas parecidas representan elementos distintos de la batería. Por ejemplo el ride en el filo y en la campana se imitan de forma muy parecida por un mismo usuario en muchas ocasiones (Fig. 4-2g), e incluso usuarios distintos imitan de forma idéntica estos dos elementos diferentes. Pasa lo mismo con el charles abierto y cerrado, pero teniendo en cuenta que son el mismo elemento tocado de diferente forma, no parece que el error sea tan trascendental, de hecho si sólo se hubiera introducido una modalidad de charles y otra de ride, probablemente los resultados y la utilidad de posibles aplicaciones futuras no habrían variado demasiado.

Otro problema parecido nos lo encontramos en los timbales. Se han observado dos modos de imitación de timbales. El primero de ellos realizaba un sonido con una vocal distinta para cada uno, por ejemplo /*tim*/, /*tum*/, /*tom*/ (Fig. 4-2h), y el segundo imitaba los tres timbales igual pero variando la frecuencia fundamental (Fig. 4-3a), lo que es bastante lógico teniendo en cuenta que se trata de instrumentos de altura definida. Un problema es que esa alteración de pitch se aprecia cuando un timbal se imita a continuación de otro, pero en grabaciones distintas el pitch de un mismo timbal varía, y mucho más si además son usuarios distintos.

Pero el problema más preocupante es que la mayor parte de las veces el timbal es imitado de forma muy parecida al bombo, es decir mediante la onomatopeya /*tum*/ (Fig. 4-3b). En los casos en los que se hacía notar al usuario este hecho para poder realizar la grabación de entrenamiento, la respuesta más frecuente era imitar el bombo como /*tu*/ y el timbal como /*tum*/ y con otro pitch. Pero no parece una diferencia demasiado notable y más teniendo en cuenta que cuando los golpes de bombo están aislados (o no hay otro golpe muy rápido después) o van al final de un ritmo, la tendencia es a hacerlo como /*tum*/.

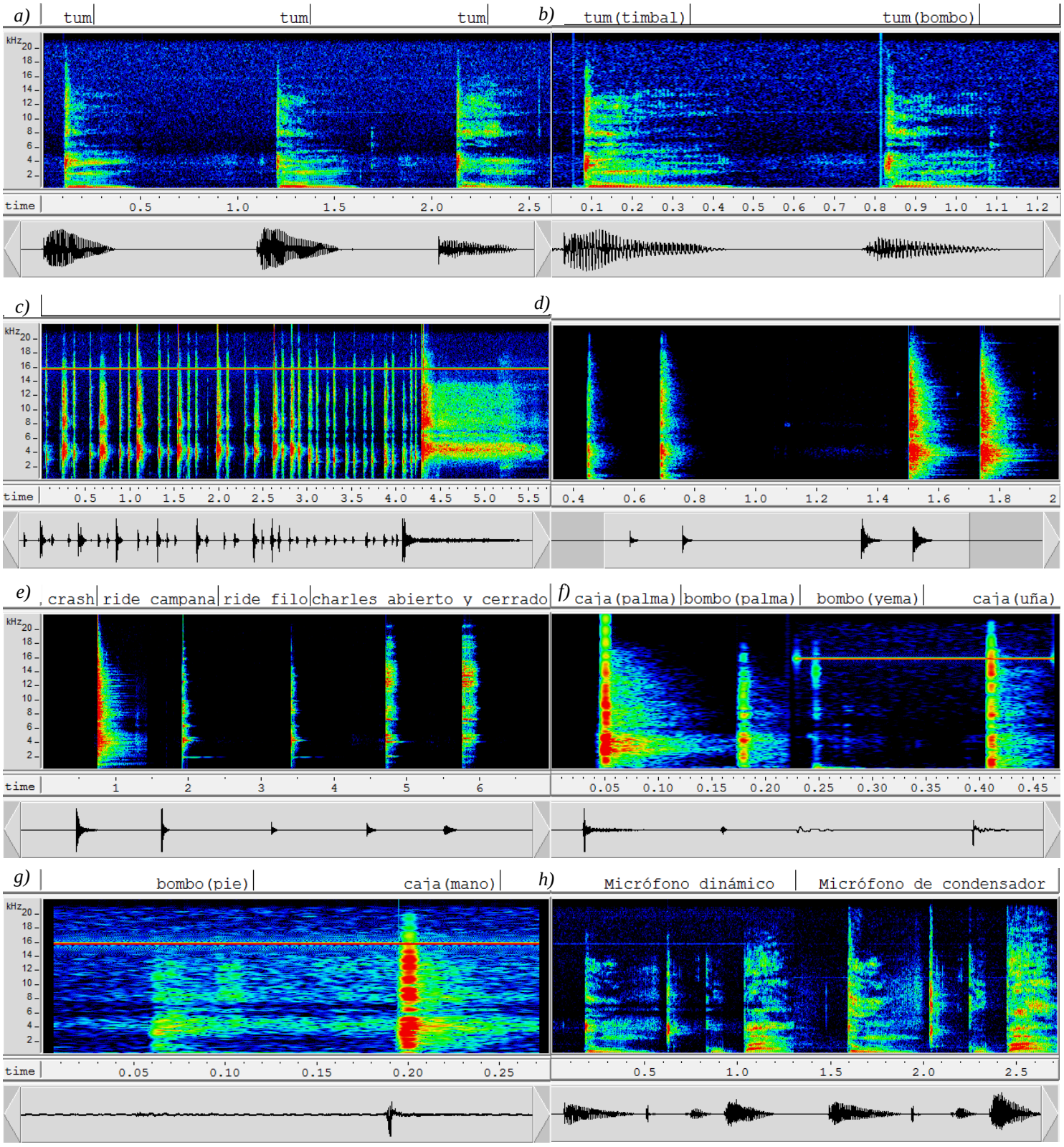


Figura 4-3: a) Imitación los distintos timbales con las misma onomatopeya, pero variando la frecuencia fundamental (mismo usuario). b) Imitaciones muy similares de dos elementos distintos por parte de un mismo usuario. c) Ritmo interpretado mediante beat box. d) Caja imitada por dos utensilios distintos en la modalidad de estilo libre. e) Diferencias de los distintos platos imitados en la modalidad de estilo lire. f) Distintas formas de imitar bombo y caja mediante golpes. g) Diferencia entre los golpes de imitación de bombo (con el pié) y la caja (con la mano). h) Diferencias espectrales de una misma señal grabada con dos micrófonos distintos.

Pasa algo parecido con el charles abierto y el crash. Cuando el crash se imita mediante la onomatopeya plas, clash o chas no hay problema, pero a veces se imita como tss o chss, que es exactamente como se imita casi siempre el charles abierto.

4.1.2 Beat Box

Desafortunadamente la cantidad de usuarios grabados que supieran hacer beat box es muy reducida como para poder realizar un análisis concluyente. Además, clasificando como beat box cualquier tipo de percusión vocal no onomatopéyica, nos podemos encontrar con imitaciones bastante lejanas al beat box tal y como habitualmente se le conoce, con lo que las formas de imitación pueden ser muy amplias (Fig. 4-3c).

4.1.3 Estilo libre

Esta modalidad es completamente dependiente de usuario, dado que cada usuario se construye a su manera su propia batería de estilo libre, por lo que un sonido de caja puede sonar completamente distinto de un usuario a otro dependiendo del objeto que golpee (Fig. 4-3d) y los sonidos que para un usuario suenen a caja para otro sonarán a timbal. Si que se aprecia sin embargo cierta diferencia entre los sonidos que imitan a caja, bombo y timbales de los que imitan a los platos y dentro de los platos el más distinguible es el que imita a la campana del ride (Fig. 4-3e).

Dentro de un mismo usuario, por comparación de unos elementos con otros, si se pueden apreciar diferencias, pero no está tan claro qué sonidos imitan a qué elementos.

4.1.4 Golpes

Las características son muy similares al caso anterior solo que en este caso únicamente hay dos tipos de sonido (o tres en el caso de que haya timbales, pero esto tiene una relevancia secundaria) y son más parecidos. La diferencia de estos sonidos depende bastante de la elección del usuario sobre su forma de ejecución dado que se suele golpear sobre la misma superficie o superficies muy similares. En los casos en los que se ha escogido imitar los diferentes sonidos golpeando con las palmas de las manos la diferencia se consigue empleando distintas partes de la palma, colocándola de distinta forma, pegando con más o menos fuerza, golpeando en sitios distintos (pecho y abdomen)… De esta forma se puede hacer que sonidos distintos imiten a elementos distintos. Otra opción es imitar el bombo golpeando en una mesa con las yemas de los dedos y la caja con las uñas (Fig. 4-3f). Hay muchas formas distintas válidas siempre y cuando se produzcan sonidos diferentes para cada elemento.

4.1.5 Pies y manos

Esta categoría es muy similar a la anterior, con la salvedad de que en este tipo de imitación es más fácil conseguir que los sonidos de bombo y caja sean diferentes (Fig. 4-3g).

4.2 Variabilidad de canal

Como se ha comentado anteriormente al grabar con dos micrófonos distintos hemos recogido la misma señal acústica con dos transductores distintos. En algunos casos se observa cierta diferencia, como más detalle en frecuencias agudas en el micrófono de condensador, tiene mayor rango frecuencial de captación (Fig 4-3h). Además capta más nítidamente ruidos espurios. Pero por lo demás, espectralmente no se observan diferencias notables.

5 Conclusiones y trabajo futuro

5.1 Base de datos

La base de datos drhuman es una base de datos bastante completa en cuanto a estilos de percusión humana, tipos de ritmos y variabilidad de interpretación y canal. El principal defecto que tiene es que se limita a sonidos de percusión de batería. Tal vez se deba realizar un trabajo futuro ampliando los instrumentos de percusión a imitar, pero parece más apropiado resolver de una manera efectiva el problema que se plantea con este conjunto limitado y más tarde expandir el *drumset*. Esta base de datos, sin embargo y por el momento tiene bastante carencia de imitaciones mediante beat box, por lo que se debería realizar un esfuerzo de recopilación de este estilo.

Lo primero sin duda que se debe realizar como trabajo futuro referido a la base de datos es terminar de grabarla, revisarla y etiquetarla.

5.2 Sistemas de reconocimiento

En lo referente a imitación por voz onomatopéyica, podemos establecer ciertos puntos en común, como que todos los usuarios imitan el bombo mediante onomatopeyas que incluyen la vocal *u* e igualmente en la caja con la vocal *a* y gran mayoría coinciden en imitar el ride (sobre todo en la campana) con onomatopeyas que incluyen la vocal *i*.

Sin embargo existe gran variabilidad a la hora de imitar otros elementos como el crash, que es muchas veces imitado mediante la onomatopeya *tss*, por lo que puede ser confundido con el charles abierto o los timbales que pueden ser fácilmente confundidos unos con otros o con el bombo.

Para desarrollar un reconocedor automático de este tipo de eventos, teniendo en cuenta que se trata de voz, lo más apropiado sería un reconocedor de habla espontánea o bien un reconocedor de palabras clave basado en HMM (*Hidden Markov Models*) entrenado con pistas de entrenamiento como la que hay en la base de datos drhuman o bien con otras pistas. Se podría considerar realizar sistemas de entrenamiento con un modelo independiente del usuario o bien adaptados a cada usuario (sin duda daría mejores resultados). Además se podrían explotar las observaciones hechas acerca de las características comunes de imitaciones de determinados elementos de la batería para tener reconocedores más universales. El problema vendría a la hora de diferenciar elementos distintos imitados de formas similares.

Para solucionar dicho problema habría varias posibilidades. La primera de ellas consistiría simplemente en reducir el vocabulario a elementos fácilmente distinguibles, con ellos se podrían construir ritmos fácilmente, pero limitaría la utilidad a aplicaciones de MIR, quedando las orientadas a composición musical muy restringidas. Por otra parte se podría crear un idioma específico que concretara cómo se debe imitar cada elemento de percusión. Si esto se hiciera se podría generar un sistema independiente de usuario, pero eso restaría espontaneidad y naturalidad a las imitaciones, que es precisamente una de las ventajas de los sistemas potenciales que se generarían. Otra posibilidad es que el usuario corrija las primeras transcripciones incorrectas con el fin de que el sistema aprenda, pero si el usuario

corrige transcripciones en las que él mismo se ha equivocado, el aprendizaje sería erróneo, debería hacerse una especie de "corrección responsable". Tal vez esta última posibilidad sería interesante si no se abusara de ella y para esto lo mejor podría ser pasar la primera transcripción obtenida por un sistema basado en n-gramas entrenados con patrones de percusión. De esta forma las transcripciones sería corregidas automáticamente primero y después el usuario podría realizar una segunda corrección.

En cuanto al beat box, dado que se ha recogido muy poca cantidad de grabaciones de este estilo, es difícil extraer conclusiones, lo que es una verdadera lástima, ya que una de las restricciones que se impone a este estilo es que se interpretan todos los sonidos a la vez y la persona que lo hace desarrolla cierta habilidad polifónica. Los buenos beat boxers son capaces de interpretar una melodía a la vez que la percusión, lo que abre un enorme abanico de posibilidades tanto de cara a la síntesis musical como a aplicaciones MIR. Por este motivo no se ha contemplado grabar los elementos de la batería por partes en esta modalidad, así que el conjunto de elementos de percusión a reconocer es bastante más limitado.

Los pocos archivos que se han recogido de beat box en la base de datos drhuman de momento han servido para darnos cuenta de que hay estilos muy diferentes, dado que la única restricción que se impone es que sean sonidos bucales no onomatopéyicos, aunque sí que es cierto que existe una variante de beat box más clásica (si se puede llamar así) que tiene más sonidos comunes. Sin embargo la escasez de datos no permite enunciar una conclusión por el momento.

Para reconocer estilo libre, dada la naturaleza de los sonidos producidos y de la subjetividad de qué sonido imita mejor a qué elemento se podría enfocar el problema desde varios puntos de vista. Se podría implementar un reconocedor como los comentados para estilos de percusión humana anteriores y seguramente funcionaría bastante bien, especialmente un reconocedor de palabras clave que además discriminara otros ruidos. Ahora bien, este tipo de reconocedor sería dependiente de usuario y de *drumset*, por lo que habría que volver a entrenarlo cada vez que se cambiara alguno de los elementos de la batería de estilo libre. Otra alternativa es realizar el reconocimiento con un sistema de clustering automático al que tendríamos que indicar cuántas clases tenemos. Una vez clasificados los sonidos, se podría asignar cuál es cuál mediante conocimientos a priori, como que los platos tienen un espectro más ruidoso o que los timbales y el bombo tienen que sonar respectivamente unos más graves que otros. Con ese tipo de información se podría intentar realizar un sistema sin entrenamiento a tiempo real que podría funcionar bien en baterías de estilo libre que imitaran bien a las reales. Si además etiquetamos la primera clasificación ya tendríamos un sistema entrenado.

En los casos de imitación con golpes y mediante pies y manos, dado que los sonidos producidos son lo suficientemente distintos como para poder clasificarlos correctamente en dos clases se podría intentar realizar el reconocimiento basado directamente en un sistema de clustering automático. El problema de este tipo de sonidos de cara al reconocimiento es que a menos que siempre toquemos en la misma superficie y de la misma manera, entrenar un modelo del tipo que sea por clase sólo tendría validez para una sesión. Sería mucho más útil conseguir un sistema que no necesitara de dicho entrenamiento. Este sistema podría estar basado en un modelo independiente del usuario combinado con información a priori, como que el bombo tiene que sonar más grave que la caja, o se podría combinar con n-gramas de patrones rítmicos. Pero tal vez podría ser más interesante entrenar un sistema

que eligiera qué parámetros del vector de características son más discriminativos para hacer la ponderación del clustering automático.

Vistas las características de los distintos tipos de percusión humana, parece que la modalidad más estable y por tanto la que tiene mayores probabilidades de éxito en tareas de reconocimiento es la imitación por voz onomatopéyica. Queda pues como trabajo futuro experimentar con las distintas propuestas que se han formulado como conclusiones del estudio analítico.

5.3 Posibles aplicaciones

La primera aplicación de un sistema de transcripción de sonidos de imitación de elementos de batería a etiquetas de información acerca de los elementos imitados, que se viene a la mente es MIDI o cualquier otro método de representación musical. Con la transcripción puesta en dicho tipo de formato, la aplicación más inmediata irá encaminada a la interpretación de ritmos para composiciones musicales. Esto puede ser como un programa independiente, o bien embebido en algún tipo de hardware, por ejemplo para videojuegos, o bien como un plugin (VST, RTAS o AU) para programas de edición y grabación de audio. Para programas de edición de audio no sería estrictamente necesario que el sistema funcionara a tiempo real, pero para aplicaciones de interpretación como actuaciones en directo o performances de cualquier tipo sí sería necesario.

Si además de funcionar en tiempo real, el sistema fuera especialmente rápido, podría resultar útil en aplicaciones que requieran interactividad de uno o varios usuarios (en cuyo caso se podría combinar incluso son sistemas de reconocimiento de locutor para distinguir a los usuarios), como aplicaciones web y nuevamente videojuegos.

Tanto en forma de MIDI como en forma de otro tipo de representación musical podría servir para realizar tareas de MIR basadas en patrones rítmicos o como complemento a otro tipo de reconocedores, abriendo la puerta a un nuevo tipo de interfaz de consulta entre personas y sistemas informáticos.

5.4 Otras líneas de investigación relacionadas

El disponer de un sistema de estas características operativo podría abrir más líneas de investigación y aplicaciones. Por ejemplo se podría intentar realizar reconocimiento de batería directamente para disparar eventos MIDI musicales o bien secuencias de luces, video u otros recursos escénicos. O también mejorar las aplicaciones de triggering existentes, que simplemente detectan la intensidad del golpe, intentando que también detecten otros aspectos como si el golpe se ha dado en el centro del parche/plato o en el exterior o si el plato se estaba moviendo cuando se le da un segundo golpe. Con este tipo de información incluso se podría plantear sustituir completamente los sonidos grabados de una batería por otros sintetizados o bien extraídos de una base de datos de muestras lo suficientemente grande, lo que permitiría realizar “pinchazos” en las grabaciones de batería en cualquier punto, manteniendo la naturalidad de la ejecución.

Ya que estamos con la imitación de instrumentos podríamos explorar la imitación de otros instrumentos que al igual que la batería también son imitados con la voz como un bajo (*du dum*…) una guitarra eléctrica (*chann* acorde o *waaarrgghh* un solo con wah wah) o una trompeta (con un peine y una servilleta)… La lista de imitaciones puede provocar alguna

sonrisa al recordar alguna imitación, pero también podía parecer poco seria la idea de imitación de batería y el análisis de la base de datos demuestra que gente que no se conoce de nada coincide en la forma de imitar el instrumento. Además las aplicaciones serían muy similares a las que se han descrito en este TFM.

Finalmente, este tipo de sistemas podrían adaptarse para permitir identificación mediante ritmos, de forma que en lugar de proporcionar al sistema de seguridad una palabra de paso se le podría proporcionar un ritmo de paso o *passrhythm*, tanto mediante un micrófono como mediante una superficie sensible como una pantalla o panel táctil. El problema es que se convertiría en un sistema exclusivo para gente que tuviera un mínimo de agilidad haciendo ritmos, luego su utilidad estaría un poco limitada.

Referencias bibliográficas

Bartsch, M.A. and Wakefield, G.H. - *Audio thumbnailing of popular music using chroma-based representations* - IEEE Transactions on Multimedia, 7(1):96–104, 2005.

Bongers, B. - *Physical Interfaces in the Electronic Arts Interaction Theory and Interfacing Techniques for Real-Time Performance* - M. Wanderley, M. Battier (eds.), Trends in Gestural Control of Music, IRCAM, Paris, 2000.

Brain Opera - http://park.org/Events/BrainOpera/ - 2010

Chadabe, J. - *Electric Sound: The Past and Promise of Electronic Music*. Upper Saddle River, NJ: Prentice Hall, 1997.

Cook, P. - *Principles for Designing Computer Music Controllers* - Proc. New Interfaces for Musical Expression Workshop (NIME´01), Seattle, 2001.

De Cheveigné, A. - *A note-lattice descriptor for melody* - MPEG-7 document number MPEG99/M6086. (2000).

De Cheveigné, A. and Baskind, A. - *F0 estimation* -In Proceedings of Eurospeech, pages 833–836, 2003.

Downie, S. and Nelson, M. - *Evaluation of a simple and effective music information retrieval method* - In Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR), pages 73–80, 2000.

Fujishima, T. - *Realtime chord recognition of musical sound: a system using common lisp Music* - In Proceedings of the International Computer Music Conference, pages 464–467, 1999.

Ghias, A., Logan, J., Chamberlin, D. and Smith, B.C. - *Query by humming: musical information retrieval in an audio database* - In Proceedings of the ACM Conference on Digital Libraries, pages 231–236, 1995.

Gómez, E., Klapuri, A. and Meudic, B. - *Melody description and extraction in the context of music content processing* - Journal of New Music Research, 32(1):23–40, 2003.

Gómez, E. and Herrera, P. - *Estimating the tonality of polyphonic audio files: cognitive versus machine learning modelling strategies* - In Proceedings of the International Conference on Music Information Retrieval, pages 92–95, 2004.

Goto, M. and Muraoka, Y. - *An audio-based real-time beat tracking system and its applications* - In Proceedings of the International Computer Music Conference, pages 17–20, 1998.

Guillet, O., Richard, G. - *Drum Loops Retrieval from Spoken Queries* - Journal of Intelligent Information Systems, pp 159-177, Springer, 2005

Hu, N., Dannenberg, R.B. and Lewis, A.L. - *A probabilistic model of melodic similarity* - In Proceedings of the International Computer Music Conference, pages 509–515, 2002.

Huron, D. - *The Humdrum Toolkit: Software for Music Research* - Center for Computer Assisted Research in the Humanities, Ohio State University, copyright 1993-1999.

Kapur, A., Benning, M. and Tzanetakis, G. - *Query-by-beat-boxing: music retrieval for the DJ* - In Proceedings of the International Conference on Music Information Retrieval, pages 170–177, 2004.

Kapur, A., Yang, E. L., Tindale, A. R. and Driessen P. F. - *Wearable Sensors for Real-Time Music Signal Processing* - Communications, Computers and signal Processing, 2005. PACRIM**.**

Li, Q., Kim, B.M., Guan, D.H. and Oh, D.W. - *A music recommender based on audio features* - In Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR), pages 532–533, 2004.

Machover, T. - *Hyperinstruments: A Progress Report, 1987-1991* - Technical Report, Massachusetts Institute of Tecnology, Boston, 1992.
Melucci, M. and Orio, N. - *Musical information retrieval using melodic surface* - In Proceedings of the ACM Conference on Digital Libraries, pages 152–160, 1999.
MIREX - http://www.music-ir.org/mirex/2009/index.php/Main_Page - 2009
Morita, H., Hashimoto, S. and Ohteru, S. - *A Computer Music System that Follows a Human Conductor* - Computer, Vol 24, pp 44-52, 1991.
Nettheim, N. - *On the spectral analysis of melody* - In Journal of New Music Research, vol. 21, pp. 135-148. 1992.
Orio, N. - *Music Retrieval: A Tutorial And Review* - Foundations and Trends in Information Retrieval, Vol 1, No 1, pp 1-90, 2006.
Paiva, R.P., Mendes, T. and Cardoso, A. - *On the detection of melody notes in polyphonic audio* - In Proceedings of the International Conference on Music Information Retrieval, pages 175–182, 2005.
Paradiso, J. A. and Hu, E. - *Expressive Footwear for Computer-Augmented Dance Performance* - In First International Symposium on Wearable Computers (ISWC '97), 1997
Park, C., Chou, P. H. and Sun, Y. - *A WearableWireless Sensor Platform for Interactive Dance Performances* - Proceedings of the Fourth Annual IEEE International Conference on Pervasive Computing and Communications, p.52-59, March 13-17, 2006.
Paulus, J. and Klapuri, A. - *Model-Based Event Labeling in the Transcription of Percussive Audio Signals* - VI Conference on Digital Audio Effects (DAFx03), London, UK, 2003.
Pousset, D. - *La flute-MIDI: L´historie et quelques Applications* - Master´s Theses, University of Paris-Sorbone, 1992.
Shifrin, J., Pardo, B., Meek, C. and Birmingham, W. - *HMM-based musical query retrieval* - In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, pages 295–300, 2002.
Typke, R., Veltkamp, R.C. and Wiering, F. - *Searching notated polyphonic music using transportation distances* - In Proceedings of the ACM International Conference on Multimedia, pages 128–135, 2004.

Anexos

A Aspectos formales de la base de datos drhuman

A continuación se anexan documentos empleados para adquirir el consentimiento del usario para la toma de datos (B.1), las instrucciones recibidas por el usuario junto con los archivos de audio que contienen los ritmos (B.2) (en este caso solo se anexa el enviado a los usuarios del escenario S0, los demás son similares pero con los nombres de los archivos de audio adaptados) y el mail que recibían los usuarios para concertar cita (B.3).

A.1 Consentimiento de toma de datos

Consentimiento de toma de datos para la base de datos

Firmando el presente documento, entiendo y consiento de forma libre y voluntaria que se registren sonidos producidos por mí de percusión humana (producidos mediante sonidos de mi cuerpo como la voz, respiración, golpes, fricciones o cualquier otro tipo de sonido producido empleando el cuerpo y la ropa, así como sonidos producidos mediante manipulación y golpes de distintos objetos y en distintas superficies), así como datos personales, incluyendo nombre, datos de contacto y mi relación con la música y la percusión, los cuales serán adquiridos, procesados y utilizados por Daniel Hernández López como controlador de la base de datos, de acuerdo con las leyes vigentes y con lo que se expone en el presente documento.

La base de batos drhuman es un proyecto de investigación cuyo fin es analizar las diferentes características espectrales y estadísticas de los distintos eventos de sonido de percusión humana reales grabados y etiquetados, así como la variabilidad de los mismos entre usuarios y entre canales. El motivo de elaborar la base de batos es crear un marco de referencia para la investigación, desarrollo, evaluación y comparación de distintas técnicas y algoritmos de clasificación de eventos acústicos y reconocimiento de patrones rítmicos, de modo que permitan implementar sistemas capaces de reconocer y mapear tales eventos como eventos producidos por instrumentos de percusión, permitiendo así el desarrollo de todo tipo de aplicaciones derivadas de tales sistemas como puedan ser aplicaciones de MIR (*Music Information Retrieval* o Recuperación de Información Musical) como búsquedas, indexación, etc., o aplicaciones que permitan emplear los sonidos de percusión humana recogidos por un micrófono como controladores MIDI.

Dado que la base de batos drhuman pretende servir de marco de referencia en investigación y desarrollo, dicha base de datos podrá ser cedida a licenciatarios internacionales, para investigaciones que se desarrollen en instituciones académicas o por compañías mercantiles, con la finalidad de evaluar sistemas de análisis, clasificación o reconocimiento automático de eventos acústicos o patrones rítmicos, creando un marco de referencia reutilizable de algoritmos y sistemas. Esta cesión se realizará siempre bajo estrictas condiciones de uso siguiendo las directivas europeas y las leyes y regulaciones nacionales de protección de datos. Por ello, se exigirá a licenciatarios externos a la UE el cumplimiento de dicha regulación y de condiciones estrictas para el uso de la base de datos drhuman, que se estipularán en los respectivos contratos de licencia que se suscriban. En este sentido, entiendo y consiento que los sonidos producidos por mí y grabados y procesados en la base de datos drhuman, así como datos relativos a mi relación con la música y la percusión (excluyendo

cualquier otro tipo de datos personales o identificativos, que serán confidenciales) podrán ser cedidos, exclusivamente en dichas condiciones, a licenciatarios de terceros países. Entiendo que si los datos y resultados científicos fueran publicados, mi identidad permanecerá en todo caso confidencial.

Si no estoy de acuerdo con aportar los datos mencionados anteriormente, entiendo que no puedo participar como voluntario en la campaña de adquisición.

Tengo derecho a solicitar acceso a mis datos personales, y a corregir y borrar, si es pertinente, mis datos personales de acuerdo con la legislación vigente. Con este objeto, puedo contactar con el controlador (persona de contacto actual: Daniel Hernández López mail: d.hernandezlopez@gmail.com, nótese que el controlador puede designar otras personas de contacto). Puedo solicitar al controlador que, dependiendo del caso, rectifique, complete, actualice, bloquee o borre mis datos personales que sean imprecisos, incompletos, equivocados, hayan expirado o cuya adquisición, uso, o distribución esté prohibida, teniendo efectos en todas las distribuciones de la base de datos drhuman.

Los datos almacenados por el controlador y cualquier otro dato que se haya transferido a terceros, podrá ser eliminado a petición del controlador después del final del año 2020, a menos que se amplíe dicho plazo como resultado de la negociación del controlador con licenciatarios.

He leído todo lo anteriormente expuesto y entiendo que disfruto de entera libertad para manifestar mi consentimiento con respecto a la adquisición y posterior procesamiento de mis datos acústicos y personales. Mediante la firma del presente documento, manifiesto y doy mi consentimiento con todo lo expuesto anteriormente.

Fecha:

Nombre:

Firma:

Datos de contacto:

Teléfono:
Mail:

Número de Identificación:

Relación con la música y la percusión:

A.2 Instrucciones para los usuarios

Instrucciones para la sesión de grabación de la base de datos

druman

Durante la sesión de grabación estarás escuchando los ritmos y la claqueta. Los ritmos están grabados en los archivos de audio en mp3 que hay junto a este pdf y son de los siguientes tipos:

- Imitación por voz: Son los archivos cuyo nombre empieza por **FV**. Tendrás que imitar la percusión mediante sonidos onomatopéyicos de la voz. Primero grabaremos un archivo en el que intentes imitar el mayor número de elementos de percusión a la vez (bombo, caja, platos…). Estos son ritmos reales de batería, por lo que suenan a la vez los distintos elementos, puedes optar por imitar los que quieras en cada momento. Después grabaremos el mismo ritmo pero por partes, es decir primero bombo y caja, luego platos, luego toms… o como a ti te sea más cómodo.
- Imitación con golpes: Son los archivos cuyo nombre empieza por **FH**. Sólo tienen bombo, caja y en algunos casos timbales. Tendrás que imitarlos dando golpes con las manos o las yemas de los dedos por ejemplo en superficies como mesas, carpetas, libros, tu propio cuerpo, etc. Si has tocado alguna vez la batería y te atreves a intentarlo (esto es opcional) también intentaremos grabarlo golpeando con las manos a modo de caja y timbales y con los pies como si fuera el bombo.
- Estilo libre: Son los archivos cuyo nombre empieza por **FS**. En este caso tú eliges los elementos con los que realizarás la percusión, como cajas, libros, vasos, cojines, lo que quieras. Piensa en qué elementos van a imitar a la caja, timbales (3), bombo, charles, crash y ride y tenlos a mano en la sesión de adquisición para montarte tu propia "batería". Puedes emplear si quieres lápices, bolis u otro tipo de elementos que te sirvan de baquetas. Al igual que con los ritmos que imitarás con la voz, estos ritmos también se grabarán, primero con todos los elementos de la batería y después por partes. Por último podrás grabar el ritmo que a ti se te ocurra (piénsalo antes) con la batería de "estilo libre" para en un futuro probar si se parecía a lo que intentabas imitar.
- Entrenamiento: Es el archivo **TR**. Son distintas formas de tocar los distintos elementos de la batería, tendrás que grabarlo con la voz y con los elementos de tu batería de "estilo libre".
- Groove común: Es el archivo **G_0**. Es un ritmo que grabarás con imitación por voz, la batería de "estilo libre" y si sabes hacerlo (opcional) Beat Box.

Secuencia de grabación

Este es el orden en que iremos grabando cada uno de los ritmos de percusión humana.

1. Imitación por voz. Grabaremos cada uno de los ritmos que se detallan a continuación (primero con todos los elementos juntos y después por separado) en el siguiente orden:
    1. **G_0**
    2. **FV_1**
    3. **FV_2**
    4. **FV_3**
    5. **FV_4**
    6. **FV_5**
    7. **FV_6**
    8. **FV_7**
    9. **FV_8**
2. Beat box. Esta parte es opcional, la haremos si sabes hacer beat box. Consistirá en imitar el archivo **G_0** mediante beat box.
3. Entrenamientos. Grabaremos primero mediante imitación por voz el archivo **TR** y después volveremos a grabar **G_0** (todos los elementos juntos y luego por separado) pero esta vez deberás imitar los elementos de la misma forma que lo has hecho en el archivo **TR**, es decir, si por ejemplo imitaste un golpe simple de caja como “ta” deberás hacerlo igual y no “pa” o “ka”. Después volveremos a grabar el archivo **TR** pero ahora con la batería de estilo libre y posteriormente, también con la batería de estilo libre grabaremos (todos los elementos juntos y luego por separado) el ritmo **G_0**.
4. Estilo libre. En esta parte con la batería de estilo libre grabaremos (primero con todos los elementos juntos y después por separado) los siguientes ritmos:
    1. **FS_1**
    2. **FS_2**
    3. **FS_3**
    4. **FS_4**
    5. **IMPR** (este será un ritmo que se te ocurra, mejor si lo llevas preparado)
5. Imitación por golpes. En esta última parte grabaremos, primero con golpes de las manos en una mesa, rodillas, pecho u objetos y después (si sabes hacerlo, opcionalmente) con las manos para caja y toms y pisando con los pies en el suelo para el bombo, como si estuvieras tocando una batería, los siguientes ritmos:
    1. **FH_1**
    2. **FH_2**
    3. **FH_3**
    4. **FH_4**
    5. **FH_5**
    6. **FH_6**
    7. **FH_7**
    8. **FH_8**

A.3 Comunicación y cita con los usuarios

Hola *Usuario*!!

Lo primero darte las gracias por colaborar con el proyecto drhuman. Se trata de un proyecto de investigación destinado a estudiar las características de los sonidos de percusión humana para poder desarrollar sistemas que los relacionen con los instrumentos de percusión a los que imitan. Una aplicación informática de ejemplo que podría surgir como resultado de esta investigación sería una que identificara sonidos vocales que imitan a una batería y los codificara como una secuencia MIDI de batería. Por si ejemplo si imitáramos una secuencia de bombo y caja diciendo "tu pa tu tu pa", el programa reconocería una secuencia "bombo caja bombo bombo caja" de modo que podríamos emplear la voz como controlador o disparador MIDI.

Se entenderá por percusión humana aquellos sonidos de percusión realizados sin batería ni otros instrumentos musicales de percusión. Para este estudio contemplaremos los siguientes tipos de percusión humana:

- Imitación por voz. Es el ejemplo que hemos visto antes, consiste en imitar de forma onomatopéyica los sonidos de la batería, "tum", "plas", "ding", "ta" serían algunos ejemplos para imitar un tom, un crash, un ride o una caja respectivamente. Tu puedes imitarlos como quieras, no tienes por qué seguir el ejemplo.

- Beat Box. Este método también consiste en imitar la batería con la boca, pero emitiendo sonidos que no tienen porqué ser silábicos.

- Imitación por golpes. Consiste en imitar una secuencia, por ejemplo de bombo y caja dando golpes con las manos sobre una mesa, las piernas o cualquier superficie.

- Imitación con pies y manos. Es un caso similar al anterior, consiste en imitar que estamos tocando la batería dando golpes sobre las rodillas, la tripa o el pecho para imitar la caja o los toms y pisar sobre el suelo para imitar los golpes de bombo. Para los que saben tocar la batería es más natural puesto que los movimientos son muy similares.

- Estilo libre. Consiste en construirte tu propia batería a partir de elementos cotidianos como cacerolas, cojines, libros, cubos o lo que se te ocurra y luego tocarlos como si estuvieras tocando la batería.

Para poder desarrollar el tipo de programa que te he descrito al principio necesitamos primero estudiar las características de este tipo de sonidos, por eso es muy importante adquirir una base de datos con todos estos sonidos y ahí es donde entras tú.

Adjunto a este correo te envío una serie de archivos de audio en mp3 con los ritmos que tendrás que imitar en la sesión de grabación mediante distintos tipos de percusión humana. En esos archivos se escucha la batería y un metrónomo para guiarte. Algunos archivos los imitaremos con un tipo de percusión humana y otros con otro tipo. Por ejemplo mediante imitación por golpes no tendrás que imitar platos, solo bombo y caja y ocasionalmente toms y en los de imitación por voz lo haremos por partes (primero bombo y caja, luego toms y luego platos) para que sea más sencillo.

Además de los archivos de audio te adjunto un pdf de instrucciones en el que se explica en qué orden se grabarán los archivos y con qué tipo de percusión humana tendrás que hacerlo. Si alguno te parece que va muy rápido no te preocupes, se puede bajar el tempo en la grabación.

Además si tienes dudas de cómo se hace algún tipo de percusión humana en este enlace hay colgado un video en que lo explica todo.

Enlace

Durante la grabación estarás escuchando por cascos los archivos que te hemos mandado y la claqueta para que no te pierdas y lo vayas siguiendo, está todo pensado para que la sesión de grabación se haga de la forma más rápida y cómoda posible (no más de media hora), además los ritmos son muy cortos y no son complicados para que sean fáciles de memorizar.

Ahora bien, es MUY IMPORTANTE que tengas familiaridad con los ritmos que tendrás que imitar para que puedas memorizarlos con escucharlos una o dos veces, ya que de lo contrario tendríamos que invertir tiempo en que te los aprendas, habría que repetir más veces hasta que quede bien y la sesión se alargaría. Para evitar eso te he adjuntado los archivos de audio y las instrucciones.

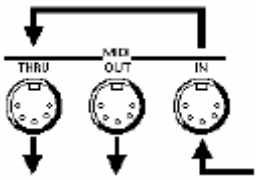
La sesión de adquisición será sobre las *hh:mm* en *lugar acordado*.

Muchas gracias por participar y hasta pronto!
Dani H.

B Protocolo MIDI

El protocolo MIDI proporciona un medio estandarizado capaz de convertir la información de una interpretación musical en datos digitales. Esta información se transmite mediante mensajes MIDI, un conjunto de instrucciones que indican al dispositivo receptor cómo debe interpretar una secuencia musical. Este dispositivo receptor es el que se encarga a su vez de generar, a tiempo real, el sonido propiamente dicho.

El protocolo MIDI también incluye una especificación hardware que consiste, entre otras cosas, en un grupo estandarizado de conectores denominados In, Out y Thru.



El flujo de datos MIDI está compuesto por una serie de bits unidireccional y asíncrona que se transmite a una velocidad de 31,25 Kbps, (10 bits por cada byte; 1 bit de inicio, 8 bits de datos y 1 bit final). Este flujo de datos se emite a partir de un controlador, por ejemplo un teclado o un secuenciador, a tiempo real y a través de conector MIDI Out.

El dispositivo que recibe los datos (por ejemplo, un sintetizador o módulos de sonidos MIDI) a través de su conector MIDI In, responde a los mensajes y emite el sonido mediante sus salidas de audio. Hay que tener en cuenta que muchos teclados MIDI incorporan a la vez el teclado controlador y el sintetizador generador de sonidos, por lo que existe un enlace interno entre ambos dispositivos. Este enlace puede estar activado o desactivado en función de si se está utilizando o no el teclado MIDI con un secuenciador externo (Local On y Local Off). Si se está utilizando el teclado MIDI con un secuenciador externo y se pretende que actúe simultáneamente como controlador y módulo de sonido, es necesario desactivar la función Local (Local Off) para evitar un gasto inútil de polifonía.

El puerto MIDI físico es capaz de alojar hasta 16 canales MIDI independientes gracias a la inclusión del parámetro de 4 bits *Número de canal*. Un teclado, por lo general, puede configurarse para transmitir en cualquiera de estos 16 canales (una excepción notoria a esta regla es la versión original del sintetizador DX7 de Yamaha). Un generador de sonido puede configurarse para recibir en un canal o canales específicos (esto último depende de su capacidad multi-tímbrica, es decir, si es capaz o no de emitir sonidos o "instrumentos" distintos de forma simultánea).

La información recibida en el conector MIDI In de un dispositivo es retransmitida (repetida) mediante el conector MIDI Thru. De esta forma es posible conectar en cadena varios dispositivos y que el flujo de datos llegue a todos por igual (conectando la salida MIDI Thru del primer dispositivo con la entrada MIDI In del segundo). Por ejemplo, se puede utilizar un teclado controlador para introducir datos en un secuenciador, que a su vez se encarga, mediante su conector MIDI Out, de enviarlos hacia distintos módulos generadores de sonido. Así es posible crear música compuesta por distintas partes instrumentales y que cada una de ellas sea interpretada por un instrumento distinto. Por supuesto, el compositor puede introducir las distintas partes de forma independiente desde el teclado ("grabarlas" una a una) y después el secuenciador se encarga de reproducirlas

todas a la vez mediante los módulos de sonido. Cada una de las partes se reproducirá en un canal MIDI distinto, que se corresponderá con el canal de recepción del módulo que disponga del instrumento apropiado.

Un mensaje MIDI se compone de un byte de estado, a continuación del cual aparecen, por regla general, uno o dos bytes de datos. Existen distintos tipos de mensajes MIDI que, en su nivel más alto, se clasifican como Mensajes de canal o Mensajes de sistema. Los mensajes de canal, como su nombre indica, son aquellos aplicables a un canal concreto y contienen el número de canal en su byte de estado. Los mensajes de sistema no están dirigidos a ningún canal concreto por lo que no incluyen número de canal en su byte de estado.

Los mensajes de canal pueden clasificarse, además, en dos tipos: Channel Voice y Mode. Los primeros incluyen datos referentes a la interpretación musical (la inmensa mayoría de los datos de un flujo de señal MIDI típico) y los segundos, datos que alteran la forma en que el instrumento receptor interpreta los datos de los primeros.

Estos últimos se utilizan para transmitir información sobre la interpretación musical. Entre los mensajes de esta categoría están: Note On, Note Off, Polyphonic Key Pressure, Channel Pressure, Pitch Bend Change, Program Change, Control Change, etc.

Note On, Note Off, Velocity
En el protocolo MIDI, los movimientos efectuados al pulsar y soltar una tecla son considerados eventos independientes. Al pulsar una tecla, cuerda, etc. en cualquier instrumento controlador MIDI, el dispositivo emite un mensaje Note On a través del puerto MIDI Out. Si el instrumento está configurado para transmitir por cualquiera de los 16 canales MIDI, el byte de estado del mensaje Note On indicará el número de canal seleccionado. Siguen a este byte de estado dos bytes de datos que especifican el número de tecla (indicando qué nota se ha pulsado) y el valor de velocidad de pulsación (Velocity) que indica lo fuerte que se ha pulsado la tecla.

El generador de sonido que recibe el mensaje utiliza el número de tecla para determinar qué nota debe sonar y el valor de velocidad de pulsación para controlar la amplitud o volumen del sonido. Al soltar la tecla, el controlador emite un mensaje Note Off, que también incluye bytes de datos para el número de nota y la velocidad con que se ha soltado la tecla (este último dato habitualmente se ignora).

Aftertouch
Algunos instrumentos MIDI cuentan con la capacidad de medir la cantidad de presión que se aplica sobre las teclas, cuerdas, etc, una vez han sido pulsadas. Esta información sobre la cantidad de presión aplicada, denominada Aftertouch, se utiliza para controlar algunos aspectos del sonido producido por el generador correspondiente (por ejemplo, vibrato). Si el instrumento controlador dispone de sensores de presión independientes para cada tecla, etc., la información polifónica resultante se transmite como mensajes del tipo Polyphonic Key Pressure. Estos incluyen bytes de datos independientes para el número de tecla y la cantidad de presión.

Lo más habitual es que los teclados incorporen un solo nivel de presión para todo el teclado (no uno para cada tecla). Quizás la excepción más notable a esta regla establecida tácitamente por la mayoría de los fabricantes de sintetizadores sean los instrumentos de la

firma Ensoniq, que si incorporan Polyphonic Key Pressure. Esta información de presión por canal (Channel Aftertouch) se transmite mediante el mensaje Channel Pressure, que sólo necesita un byte da datos para especificar el valor de presión.

Pitch Bend
El mensaje Pitch Bend se envía normalmente a partir del movimiento aplicado sobre la rueda de inflexión tonal que incorporan la inmensa mayoría de los teclados actuales (o a partir de "tirar" de las cuerdas en una guitarra MIDI, por ejemplo). Esta información se utiliza para modificar la altura tonal de los sonidos reproducidos en un canal determinado. Este mensaje incluye dos bytes de datos en vez de sólo uno, con lo que se dispone de una mayor resolución a la hora de definir los movimientos realizados sobre la rueda correspondiente (para que el sonido resultante sea continuo y no dé la sensación de moverse de forma escalonada).

Program Change
Este mensaje se utiliza para indicar el tipo de sonido a emplear en un canal determinado. Sólo precisa de un byte de datos que se encarga de especificar el nuevo número de programa.

Control Change
Estos mensajes MIDI se usan para controlar una gran variedad de funciones de un sintetizador. Como los restantes mensajes de canal, sólo afectan al canal especificado mediante el byte de estado. A continuación del byte de estado aparecen un byte de datos que indica el número de controlador y un segundo byte de datos que indica el valor a aplicar. El número de controlador identifica qué función del generador de sonido se va a controlar. En la Especificación MIDI aparece una lista completa de los números de controlador definidos.

Bank Select
El controlador número 0 (con el valor 32 como LSB) se encarga de la selección de bancos de sonidos. Esta función se utiliza, junto con el mensaje de cambio de programa, para poder acceder a un mayor número de sonidos que los que contiene un solo banco (128). Los sonidos almacenados en bancos distintos al banco 1 se seleccionan anteponiendo al mensaje de cambio de programa un mensaje Control Change que especifica un nuevo valor para los controladores 0 y 32, lo que permite acceder a 16.384 bancos de sonidos con 128 programas cada uno.

Como la Especificación MIDI no describe la forma en que los bancos de un sintetizador deben relacionarse con los mensajes de selección de banco, no existe una forma estándar de seleccionar un banco en todos los sintetizadores del mercado. Algunos fabricantes como Roland (con su estándar GS) y Yamaha (no podían ser menos, ellos cuentan con el estándar XG), han adoptado lo que les ha parecido más apropiado para asegurar una mínima estandarización en cuanto a la selección de bancos de sonidos en sus respectivas gamas.

RPN, NRPN
El controlador número 6 (Data Entry), junto con los controladores números 96 (Data Increment), 97 (Data Decrement), 98 (Registered Parameter Number LSB), 99 (Registered Parameter Number MSB), 100 (Non-Registered Parameter Number LSB) y 101 (Non-Registered Parameter Number MSB), permiten aumentar la cantidad de controladores

disponibles. Los datos se transmiten seleccionando primero el número del parámetro a editar utilizando los controladores 98 y 99 ó 100 y 101. Luego se procede a definir el valor del parámetro utilizando los controladores 6, 96 o 97.

RPN y NRPN se utilizan habitualmente para enviar datos de parámetros a un sintetizador para la edición de sonidos. Los Números de parámetro registrados (Registered Parameter Number - RPN) son aquellos a los que las organizaciones MIDI Manufacturers Association (MMA) y Japan MIDI Standards Committee (JMSC) han asignado alguna función particular. Por ejemplo, existen RPNs definidos para controlar la sensibilidad de pitch bend y la afinación general de un sinte. Por su lado, los no registrados (Non-Registered Parameter Number - NRPN) no tienen asignada ninguna función específica y pueden ser utilizados de forma distinta según el fabricante. De nuevo en este caso Roland y Yamaha, entre otros, han adoptado sus propios estándares.

Mensajes Mode
Estos mensajes (controladores 121 al 127) afectan a la forma en que el generador de sonidos responde a los datos MIDI. El controlador 121 se utiliza para reiniciar todos los valores. El 122, para activar o desactivar la función Local Control (en un sinte MIDI con teclado, es posible independizar las funciones del teclado y el generador de sonidos desactivando esta función, facilitando así su funcionamiento con un secuenciador externo). Los controladores 124 a 127 se utilizan para activar o desactivar el modo Omni y para seleccionar los modos Mono o Poly.

Cuando está activado el modo Omni, el generador de sonidos responde a los mensajes MIDI recibidos por todos los canales. Si está desactivado, el generador sólo responderá a los mensajes recibidos por un canal específico.

En el modo Poly, los mensajes Note On entrantes se reproducen de forma polifónica, lo que significa que cuando se reciben múltiples mensajes Note On, a cada nota se le asigna su propia voz (dependiendo siempre del nùmero de voces disponibles en ese momento en el generador de sonidos). Así, por ejemplo, al tocar un acorde sonarán todas las notas de forma simultánea. Al seleccionar el modo Mono, se asigna una sola voz por cada canal MIDI.

La mayoría de los instrumentos MIDI actuales funcionan por defecto en la configuración Omni On/Poly. El generador reproduce los mensajes de nota recibidos en cualquiera de los canales de forma polifónica. Por otra parte, la combinación Omni Off/Poly puede resultar útil en el caso que varios generadores de sonido estén conectados en cadena mediante sus puertos MIDI Thru, ya que cada uno de ellos recibirá los datos en un canal específico y reproducirá las notas de forma polifónica. Hay que tener en cuenta que cualquier instrumento MIDI dispone de un canal designado como **Canal básico** y que sólo recibirá los mensajes de cambio de modo a través de este canal. Esta asignación puede ser fija o seleccionable por el usuario.

Mensajes System
Los mensajes de sistema se subdividen en mensajes System Common, System Real Time, o System Exclusive. Los primeros están destinados a todos los receptores del sistema; los segundos se utilizan para la sincronización de los elementos que funcionan mediante un reloj temporizador. Por su parte, los mensajes de Sistema Exclusivo incluyen un código de

identificación del fabricante y se utilizan para transferir bytes de datos formateados según una especificación diseñada por ese fabricante.

System Common
Actualmente están definidos los mensajes MTC Quarter Frame, Song Select, Song Position Pointer, Tune Request y End Of Exclusive (EOX). El mensaje MTC Quarter Frame forma parte de la información de código de tiempo MIDI utilizada para la sincronización de equipos MIDI con sistemas de audio y video.

El mensaje Song Select es utilizado por secuenciadores y cajas de ritmo capaces de almacenar varias composiciones distintas. Song Position Pointer se usa para iniciar la reproducción en un secuenciador; siempre en un punto distinto al inicio de la composición. Su valor está relacionado con el número de pulsaciones de reloj MIDI transcurridas desde el punto inicial de la composición. Este mensaje sólo puede ser utilizado con dispositivos capaces de reconocer mensajes System Real Time (MIDI Sync).

Tune Request se utiliza habitualmente para que un sintetizador analógico proceda a recalibrar la afinación de sus osciladores. Este mensaje carece de utilidad en el caso de los sintetizadores digitales.

El mensaje EOX sirve para indicar la finalización de un flujo de datos de Sistema Exclusivo.

System Real Time
Estos mensajes tienen por fin la sincronización de todos los elementos de un sistema MIDI que funcionen mediante reloj MIDI (secuenciadores, cajas de ritmo, arpegiadores, etc.). Para asegurar una correcta temporización, siempre tienen prioridad sobre los otros tipos de mensaje MIDI (los mensajes a tiempo real, de un solo byte, pueden aparecer en cualquier lugar del flujo de datos, incluso entre el byte de estado y el byte de datos de cualquier otro mensaje).

Estos mensajes de sistema a tiempo real son: Timing Clock, Start, Continue, Stop, Active Sensing y System Reset. Timing Clock es el reloj master que define el tempo de reproducción de una secuencia. Se transmite 24 veces por cada nota negra. Los mensajes Start, Continue y Stop se utilizan para controlar la reproducción.

Active Sensing sirve al propósito de eliminar notas "colgadas" que pueden aparecer al desconectar un cable MIDI durante la reproducción. Sin esta función, algunas notas pueden quedar sonando de forma indefinida (han sido activadas mediante un mensaje Note On pero el mensaje Note Off correspondiente no va a llegar nunca).

El mensaje System Reset, se encarga de reiniciar cualquier dispositivo MIDI. Por la importancia de sus efectos, generalmente no se transmite de forma automática; el usuario debe iniciar su emisión de forma manual.

System Exclusive
Los mensajes de Sistema Exclusivo pueden utilizarse para enviar datos como programas de sonido o muestras entre distintos dispositivos MIDI. Cada fabricante define sus propios formatos para este tipo de datos y dispone de un código de identificación único garantizado por la MMA y el JMSC. Esta ID del fabricante, incorporada en cada mensaje de Sistema

Exclusivo, precede a los distintos paquetes de datos, a continuación de los cuales aparece el mensaje EOX mencionado anteriormente, con el que finaliza la transmisión de datos. Los fabricantes están obligados a publicar los detalles que conforman sus formatos de datos de Sistema Exclusivo para que puedan ser utilizados libremente por otros fabricantes o por el propio usuario, siempre y cuando no se alteren o se utilice el formato de forma que entre en conflicto con las especificaciones originales definidas por el fabricante.

Algunos de estos números de identificación están reservados para protocolos especiales, entre los que figuran el Estándar para volcado de muestras MIDI (MIDI Sample Dump Standard), un formato de datos de Sistema Exclusivo dedicado a la transmisión de datos entre muestreadores, así como MIDI Show Control y MIDI Machine Control.

Running Status

Como los datos MIDI se transmiten en forma serial, es muy posible que dos o más eventos musicales que se produzcan en el mismo momento, al ser enviados uno detrás de otro, no se reproduzcan exactamente en el mismo instante. Con una velocidad de transmisión de datos de 31.25 Kbits/seg. y 10 bits transmitidos por cada byte, la transmisión de un mensaje Note On o Note Off de 3 bytes dura aproximadamente 1 milisegundo. Desde luego, generalmente esto es lo suficientemente rápido como para que esos eventos se perciban como simultáneos, aunque a nivel de números no sea así. Por ejemplo, se puede decir que ningún teclista es capaz de percibir ese pequeño desfase al tocar 10 teclas de forma simultánea, siempre y cuando las notas sean reproducidas en un lapso de tiempo inferior a unos 10 milisegundos.

Sin embargo, hay que tener en cuenta que los datos MIDI transmitidos por un secuenciador suelen incluir datos correspondientes a un gran número de pistas y que, en un momento determinado, puede llegar a existir un gran número de eventos que deban reproducirse de forma simultánea. En estos casos el retardo introducido por la transmisión de datos en serie puede llegar a ser claramente perceptible. Para ayudar a reducir la cantidad de datos transmitidos en el flujo de datos MIDI, se emplea una técnica denominada Running Status. Esta técnica toma en consideración el hecho de que es muy habitual que en una cadena de mensajes consecutivos, todos sean del mismo tipo. Por ejemplo, al tocar varios acordes, se generan un montón de mensajes Note On consecutivos, a los que siguen el montón de datos Note Off correspondientes. El truco consiste en que sólo se emite el byte de estado cuando el mensaje correspondiente no es del mismo tipo que el mensaje inmediatamente anterior (en un mismo canal MIDI, claro). Es decir, sólo se transmiten los bytes de datos para todos los mensajes del mismo tipo.

La efectividad de esta técnica mejora mucho más si en vez de enviar mensajes Note Off, se envían mensajes Note On con un valor de velocidad de pulsación igual a 0 (técnica utilizada, por ejemplo, en el secuenciador Logic Audio). De esta forma se producen largas cadenas de menasjes Note On, de transmisión más rápida que las originales cadenas cortas compuestas por mensajes de distintos tipos.

C Currículum Vitae Resumido

Daniel Hernández López, nacido en Madrid en 1983, comenzó los estudios de Ingeniería de Telecomunicación en la Universidad Autónoma de Madrid UAM en Septiembre de 2002 y terminó en Septiembre de 2008.

En 2006 entró a formar parte del grupo de investigación ATVS (Área de Tratamiento de Voz y Señales) de la UAM como coordinador de la adquisición de la UAM de la base de datos *BIOSECURE NoE: Biometrics for Secure Authentication*, siendo más tarde revisor de esta base de datos y de *BiosecurID*. Ha participando como ayudante en la adquisición de una base de datos de huellas de goma (*gummy fingers*) sin cooperación del usuario y en la creación de una base de datos de firma de largo plazo (*long term*) *off line*.

En 2007 empieza a investigar en reconocimiento de locutor dependiente de texto como tema de su Proyecto de Fin de Carrera, estando involucrado en el *Proyecto MARTA* en el marco del programa *CENIT*. Como resultado de dicho periodo de investigación se publican los siguientes artículos:

- D. T. Toledano, D. Hernandez-Lopez, C. Esteve-Elizalde, J. Fierrez, J. Ortega-Garcia, D. Ramos and J. Gonzalez-Rodriguez, "BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition", in *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008.
- D. T. Toledano, D. Hernandez-Lopez, C. Esteve-Elizalde, J. Gonzalez-Rodriguez, R. Fernandez-Pozo, L. Hernandez-Gomez, Luis Hernandez, "MAP and sub-word level t-norm for text-dependent speaker recognition", in *INTERSPEECH-2008*, 1933-1936, Brisbane, Australia, 2008.
- D. Hernández-López, D. T. Toledano, C. Esteve-Elizalde, J. González-Rodríguez, R. Fernández-Pozo y L. Hernández-Gómez, “T-norm y desajuste léxico y acústico en reconocimiento de locutor dependiente de texto”, en *V Jornadas en Tecnologías del Habla*, Bilbao, España, 2008.

A finales de 2008 comienza a cursar el Máster en Ingeniería Informática y de Telecomunicación impartido por la UAM, eligiendo una configuración de asignaturas orientada a conseguir un *major* (especialización) en Tratamiento de Señales Multimedia. En esta nueva etapa comienza con la investigación en temas de audio orientado a música, concretándose dicha investigación en el presente Trabajo Fin de Máster que concluye su etapa de Máster.