Estudio sobre la transcripción automática de instrumentos de percusión

TRABAJO FIN DE MÁSTER

Iván Lozano Blanco

Departamento de Ingeniería del Software e Inteligencia Artificial
Facultad de Informática
Universidad Complutense de Madrid

Junio 2016

Documento maquetado con TEXIS v.1.0.

Estudio sobre la transcripción automática de instrumentos de percusión

Memoria que presenta para optar al título de Máster Universitario en Ingeniería de Sistemas y Control

Dirigida por el Doctor

Javier Arroyo Gallardo

Departamento de Ingeniería del Software e Inteligencia Artificial

Facultad de Informática

Universidad Complutense de Madrid

Junio 2016

A Maite, Aimar y Alain.

Sin música, la vida sería un error
Friedich Nietzsche

Agradecimientos

En primer lugar, dar las gracias a Javier Arroyo por toda la dedicación prestada durante la realización de este trabajo. Su motivación constante, ideas, consejos y aclaraciones durante este tiempo han sido fundamentales para poder llevar el proyecto a buen término.

Agradecer también a mi familia y amigos por estar siempre ahí. En especial, a Aimar y Alain por ser mis dos mejores proyectos, y a Maite, por su apoyo incondicional y por inspirarme siempre para 'no dejar que termine el día sin haber crecido un poco'.

Resumen

La tecnología digital y el auge de las comunicaciones han supuesto un cambio radical en la forma de crear, almacenar y consumir música. Esto, a su vez, ha propiciado el desarrollo de métodos para la búsqueda, extracción y organización de información musical que, entre otras cosas, puede ser utilizado para potenciar el aprendizaje y agilizar la labor de composición por parte de músicos profesionales y aficionados. En este sentido, proveer herramientas software de transcripción automática permitirían, por un lado, acceder de una forma precisa a las crecientes bases de datos de contenido musical, y por otro, generar automáticamente partituras a partir de grabaciones de uno o varios instrumentos tocados simultáneamente. Aunque se han realizado numerosos estudios al respecto, hasta la fecha no se ha encontrado una solución robusta de la problemática que se plantea.

Este trabajo se centra en analizar los estudios realizados hasta el momento en este ámbito, y que conciernan a instrumentos musicales de percusión de altura indeterminada (*unpitched percussion instrument*), es decir, aquellos destinados a mantener el ritmo y proveer acentos de una pieza musical. El ejemplo más claro dentro de la música popular occidental sería la batería, a partir de la cual se ha desarrollado un sistema de detección y clasificación de instrumentos basado en Modelos Ocultos de Markov, que es una técnica ampliamente utilizada en aplicaciones de reconocimiento del habla, pero con poca incidencia en este campo en particular, a pesar de haber obtenido resultados prometedores en estudios previos.

Índice

Agradecimientos IX

Resumen XI

1. Introducción 1
- 1.1. ¿Qué es la transcripción musical automática? . . . 1
  - 1.1.1. Aplicaciones . . . 2
- 1.2. Transcripción automática en instrumentos de percusión . . . 5
  - 1.2.1. El kit de batería . . . 5
  - 1.2.2. Características y singularidades sonoras . . . 7
- 1.3. Motivación . . . 9
- 1.4. Organización del documento . . . 10

2. Estado del arte 11
- 2.1. Descripción y análisis del problema . . . 12
  - 2.1.1. Extracción de características . . . 13
  - 2.1.2. Detección de comienzo de nota . . . 17
  - 2.1.3. Definición de clases . . . 20
- 2.2. Aproximaciones . . . 21
  - 2.2.1. Segmentar y clasificar . . . 22
  - 2.2.2. Separar y detectar . . . 23
  - 2.2.3. Emparejar y adaptar . . . 26
- 2.3. Base de datos *ENST-Drums* . . . 27

3. Introducción a los modelos ocultos de Markov 31
- 3.1. Razonamiento probabilístico . . . 31
- 3.2. Probabilidad a lo largo del tiempo . . . 34
- 3.3. Los Modelos Ocultos de Márkov . . . 35
  - 3.3.1. Elementos de un HMM . . . 36
  - 3.3.2. Aprendizaje del modelo. Algoritmo de Baum-Welch . . . 38
  - 3.3.3. Secuencia de estados más probable. Algoritmo de Viterbi 39
  - 3.3.4. Historia y aplicaciones . . . 40

4. Transcripción de batería mediante HMM **43**
4.1. Consideraciones previas 43
4.1.1. Elección de instrumentos 43
4.1.2. Topología y modo de detección 44
4.1.3. Librerías complementarias 45
4.1.4. Edición de audio 45
4.2. Análisis de los experimentos 46
4.2.1. Recogida de eventos 46
4.2.2. Procesamiento de los eventos 48
4.2.3. Entrenamiento/validación y test del modelo 50
4.2.4. Clasificación con el *HMM* 51
4.3. Resultados 52
4.3.1. Medidas de error utilizadas 52
4.3.2. Evaluación del modelo *off-line* 55
4.3.3. Evaluación del modelo en tiempo real y con acompañamiento 60
4.3.4. Comparativa con otros estudios publicados 63
4.4. Conclusiones y lineas futuras 64

Bibliografía **67**

Índice de figuras

1.1. Extracto de una pieza musical y su transcripción. 2

1.2. Esquema de extracción de características de audio mediante el paquete *Chroma Toolbox* de Matlab 3

1.3. El videojuego *Songs2See* permite aprender a tocar un instrumento de manera interactiva. 4

1.4. 1. caja, 2. bombo, 3. plato hi-hat, 4. tom-tom aéreo, 5,6. tom-toms de suelo (goliat), 7. plato ride, 8,9 platos crash 6

1.5. Señales acústicas y espectrogramas de varios instrumentos. Las zonas rojas del espectrograma denotan alta energía espectral en ese rango de frecuencias (eje ordenadas). 9

2.1. Frecuencias de instrumentos utilizados en la base de datos *ENST-Drums*. Los instrumentos son denotados como: BD (Bombo), CR (cualquier plato crash), CY (otro tipo de plato), HH (plato hi-hat abierto o cerrado), RC (plato ride), SD (caja), TT (cualquier tom-tom) y OT (otros instrumentos). 13

2.2. Banco de filtros MEL. 15

2.3. Representación de una nota y sus distintas partes: "comienzo", "ataque", "transitorio" y "caída". 18

2.4. Representación de la etapa de reducción de la señal. 19

2.5. Frecuencias de combinación de instrumentos utilizados en la base de datos *ENST-Drums*. Los instrumentos son denotados como: BD (Bombo), CR (cualquier plato crash), CY (otro tipo de plato), HH (plato hi-hat abierto o cerrado), RC (plato ride), SD (caja), TT (cualquier tom-tom). 21

2.6. Representación de funcionamiento de los dos tipos de segmentación. En la segmentación fija se observa el desfase que se puede originar si la rejilla temporal no está adecuadamente calculada o el tipo de señal no es apta. 23

2.7. Representación de los vectores de frecuencias S y de ganancias variables con el tiempo A como una factorización del espectrograma X. La imagen ha sido extraída del trabajo de Papanikas (Papanikas, 2012). 25
2.8. Técnica de adaptación del patrón mediante la separación de sonidos utilizando la técnica PFNMF. Las características espectrales son obtenidas de la Transformada de Fourier de Tiempo Reducido (*Short-time Fourier transform, STFT*) . . 26

3.1. Esquema de una red bayesiana ingenua. 34
3.2. Estructura de una red baysiana correspondiente a un proceso de Markov de primer orden (arriba) y de segundo orden (abajo). 35
3.3. Esquema básico de un *HMM*. Los elementos X_i representan la secuencia de estados ocultos y los elementos O_i la secuencia de observaciones. 36
3.4. Esquema ampliado de un *HMM*. 37
3.5. Ejemplo de secuencia de estados más probable producida por una secuencia de observación $O = (O_1, O_2, O_3, O_4, O_5)$. 40

4.1. Esquema de un *HMM* de izquierda-derecha de 5 estados. . . . 44
4.2. Programa *Audacity* desde donde se han generado los archivos de audio fusionando la pista de batería y la pista de acompañamiento musical. 46
4.3. Representación de captura de secuencias de silencio (lineas verdes) y de instrumento (lineas rojas) para una determinada señal sonora. 47
4.4. Representación de captura de características de la señal mediante ventanas solapadas. 49
4.5. Representación de la técnica *K-fold cross validation* para K=10. 51
4.6. Espacio ROC para todos los modelos entrenados de la caja. Cada gráfica indica el resultado de los modelos para los diferentes tiempos de observación. Dentro de cada gráfica los puntos de color verde indican los modelos entrenados para un espacio muestral $m = 96$, los azules para $m = 396$ y los rojos para $m = 798$. Dentro de cada espacio muestral, los símbolos $\star$ son los modelos entrenados con $n = 13$ características, los simbolos $\triangle$ con $n = 26$ y los símbolos $\diamondsuit$ con $n = 39$. 54
4.7. Ejemplo de reconocimiento mediante comparación de probabilidades. La gráfica superior muestra la señal de audio donde se han indicado los golpeos de caja. La gráfica inferior muestra la probabilidad de presencia de los modelos silencio(linea verde) y caja (linea azul). 61

Índice de Tablas

2.1. Número de secuencias y eventos (golpeos) grabados por cada baterista en la base de datos *ENST-Drums*. 28

2.2. Etiquetas utilizadas en los ficheros de anotación de la base de datos *ENST-Drums*. 28

4.1. Resultados de exactitud en % de los modelos *HMM* de caja y silencio sobre los datos de entrenamiento (*Train/val*) y los datos de test (*Test*). 55

4.2. Resultados de la distancia ROC en % de los modelos *HMM* de caja y silencio sobre los datos de test. 56

4.3. Resultados de exactitud y distancia ROC en % de los modelos complementarios de caja y silencio. 56

4.4. Resultados de exactitud en % de los modelos *HMM* de bombo y silencio sobre los datos de entrenamiento (*Train/val*) y los datos de test (*Test*). 57

4.5. Resultados de la distancia ROC en % de los modelos *HMM* de bombo y silencio sobre los datos de test. 58

4.6. Resultados de exactitud y distancia ROC en % de los modelos complementarios de bombo y silencio. 58

4.7. Resultados de exactitud en % de los modelos *HMM* de hi-hat y silencio sobre los datos de entrenamiento (*Train/val*) y los datos de test (*Test*). 59

4.8. Resultados de la distancia ROC en % de los modelos *HMM* de hi-hat y silencio sobre los datos de test. 60

4.9. Resultados de exactitud y distancia ROC en % de los modelos complementarios de hi-hat y silencio. 60

4.10. Comparativa de los resultados de la evaluación de los modelos en *off-line* (sobre el *test set*) y en tiempo real, ambos sin acompañamiento musical. 62

4.11. Resultados de la evaluación de los modelos en tiempo real con acompañamiento musical. 62

4.12. Comparativa de resultados obtenidos para el baterista 3 con el trabajo de Paulus, ambos sin acompañamiento musical. . . 63
4.13. Resultados de otros estudios sin acompañamiento musical. . . 64
4.14. Resultados de otros estudios con acompañamiento musical. . . 64

Capítulo 1

Introducción

Resumen: Este capítulo introduce el concepto de transcripción automática aplicado al ámbito musical, exponiendo de forma breve sus objetivos, aplicaciones y terminología básica. Asimismo, centra la problemática dentro de los instrumentos de percusión, describiendo sus características y peculiaridades sonoras.

Dentro del ámbito musical, se puede definir la transcripción como la notación que se realiza sobre una pieza. Esta "escritura sobre el papel" debe ser tal, que queden reflejadas la tonalidad, el tiempo, la duración y la fuente de cada sonido que ocurra durante la representación. En la cultura occidental, la música escrita utiliza símbolos para indicar estos parámetros. La figura 1.1 muestra la notación realizada sobre una pieza musical representada a través de su señal de audio. A estas notaciones se les llama de forma común partituras.

1.1. ¿Qué es la transcripción musical automática?

El concepto de transcripción musical automática fue usado por primera vez en los años 70, y sirve para denotar aquellos métodos o aplicaciones que tienen como objetivo "analizar una señal de audio para descubrir los instrumentos y las notas que se están tocando, y ser capaz de producir una transcripción escrita de la pieza musical" (Plumbey y Abdallah, 2002). Así pues, se puede dividir la problemática en dos tareas fundamentales: Analizar la pieza musical y transcribir el resultado de ese análisis. El análisis realizado debería permitir extraer los tonos, acordes y acentos, así como la duración de los mismos y el instante de tiempo en el que se producen con el fin de ser trasladados a notación musical. La complejidad de la tarea será mayor cuantos más elementos tomen parte de manera simultánea en la señal, ya que el sistema debería ser capaz de separar las fuentes de procedencia, distinguiendo cada uno de los instrumentos que toman parte en ella.

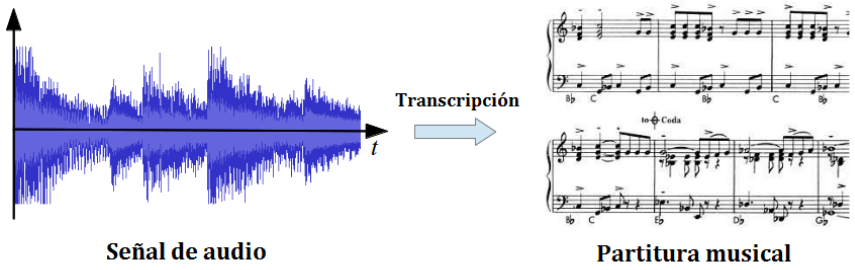

Figura 1.1: Extracto de una pieza musical y su transcripción.

Este trabajo se centra en explorar técnicas específicas para el análisis de la señal musical. Esta tarea se podría dividir en dos apartados:

- Detección del ritmo de la melodía (*beat detection*), determinar el intervalo de tiempo de repetición entre las diferentes notas que conforman la melodía, medido normalmente en golpeos por minuto (*beats per minute*). A pesar de la naturaleza intuitiva que el ser humano posee para llevar el ritmo de una canción, el desarrollo de un algoritmo automático que lo logre no es una tarea trivial (Dixon, 2001).

- Detección del tono (*pitch detection*), es decir, reconocimiento de las notas que conforman la melodía de la canción. Habría que hacer distinción de los métodos de detección de la tonalidad en música de tipo monofónico, donde solo un instrumento toca una nota a la vez, o de música de tipo polifónico, donde varios instrumentos y voces pueden aparecer de manera simultanea. Mientras que en el primer caso la tarea es relativamente sencilla y es prácticamente un problema solucionado (Klapuri, 2004), en el segundo la tarea permanece todavía abierta, debido a posibilidad de solapamiento entre las distintas señales transmitidas por cada fuente.

1.1.1. Aplicaciones

La transcripción automática musical queda enmarcada dentro del campo de la Recuperación de Información Musical (MIR), que estudia métodos y desarrolla aplicaciones para analizar el contenido de una pieza musical desde varios puntos de vista, aglutinando ramas tan diversas como la musicología, la psicología, el estudio académico musical o el aprendizaje automático. Apoyado por la Sociedad Internacional de Recuperación de Información musical (ISMIR (ISM)), entre sus aplicaciones destacan:

- Categorización y clasificación musical por género, artista etc. (jMIR (jMI)).

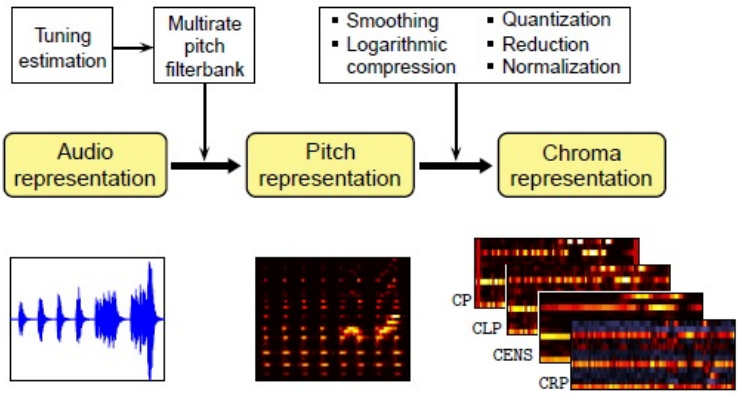

Figura 1.2: Esquema de extracción de características de audio mediante el paquete *Chroma Toolbox* de Matlab .

- Sistemas de recomendación de canciones en función de gustos personales (Pandora (Pan)).
- Generación y mezclado automático de música (VirtualBand (Vir)).
- Busqueda de canciones y melodías en base a patrones, contornos, ritmos o partituras digitalizadas. (Musicpedia (Mus, b), Peachnote (Pea)).
- Técnicas de procesamiento de señal tales como cambiar la instrumentación, el arreglo o la sonoridad de determinados eventos sonoros (Ludwig (Wri)).

En el ámbito puramente analítico de la señal musical, se han desarrollado programas y paquetes para la extracción de características, separación de pistas, determinación del tono dominante, detección de comienzo de nota (*onset detection*) e identificación de instrumentos (Essentia (Ess), Chroma Toolbox (Chr) ver figura 1.2, MSAF (Msa)), aunque varias de estas funcionalidades trabajan con restricciones, principalmente ligadas al grado de polifonía (número de instrumentos involucrados) de la pieza. La mejora de estos desarrollos permitiría proveer herramientas software de transcripción automática de propósito general para músicos profesionales y aficionados.

Asimismo, existen numerosos programas de notación musical para la creación y edición de partituras de forma manual (MuseScore (Mus, a), Denemo (Den)), con una gran variedad en grados de sofisticación. Entre ellos, caben destacar aquellos que aceptan como entrada la señal proporcionada directamente por instrumentos, y en ese sentido, han surgido herramientas completas de transcripción automática eficaces (ScoreCloud (Sco)), aunque

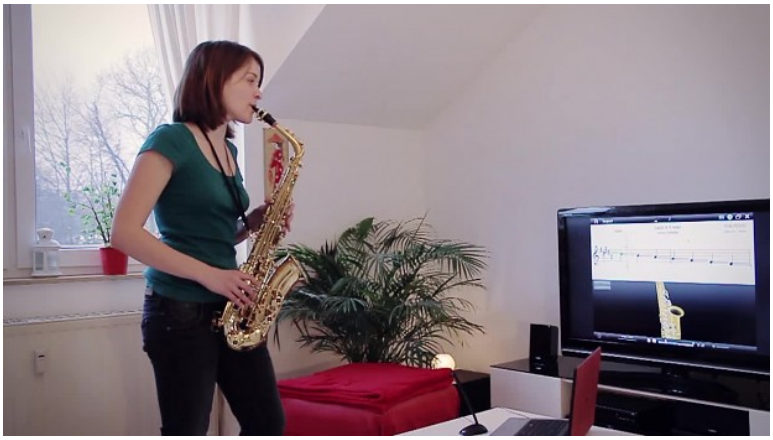

Figura 1.3: El videojuego *Songs2See* permite aprender a tocar un instrumento de manera interactiva.

ofreciendo mejores resultados en interface MIDI [1] que en señal de audio acústica, lo que limita su utilización a instrumentos eléctricos (Scorewriter, 2004).

Otras aplicaciones relacionadas con la transcripción musical incluyen los sistemas interactivos musicales, que son "aquellos cuyo comportamiento cambia en respuesta a una entrada musical" (Rowe, 1993). Estos sistemas son utilizados con fines pedagógicos para crear canciones basada en prototipos (CODES) o a través de juegos musicales educacionales (Songs2See (Son) ver figura 1.3), prácticos y de ensayo para comprobar la precisión de una actuación (good-sounds (Goo)), de ocio en vídeojuegos (Rock Band (Roc), Guitar Hero (Gui)), así como instrumentos de terapia de ayuda a la rehabilitación y cuidado de ancianos (TheBeamz (Bea)), y equipamiento de generación de efectos lumínicos sincronizados musicalmente para actuaciones en directo.

Por último, y centrándose en la transcripción de instrumentos de percusión, mencionar la importancia actual del software de producción musical, donde el post-procesado de los instrumentos de percusión en la mezcla de una canción es imprescindible. A menudo una mala colocación de los micrófonos de grabación en la batería producen interferencias entre sus diferentes piezas, y aunque existen *plug-ins* [2] para atenuar esos problemas (Drumagog (Dru)), sería deseable una aplicación de separación y localización automática de señal recogida desde un solo punto de captación, evitando el consiguiente gasto y dificultades derivadas de una mala grabación de los instrumentos.

[1] Musical Istrument Digital Interface. Es un protocolo estándar para el intercambio de información musical entre dispositivos electrónicos.

[2] Complemento de la aplicación principal que agrega una función nueva y específica.

1.2. Transcripción automática en instrumentos de percusión

La atención recibida en estudios académicos sobre la transcripción de instrumentos de percusión y la batería en particular ha sido considerablemente menor a la de otros instrumentos. Esto puede ser debido al carácter menos "serio" que a este tipo de instrumento se le da, al estar relacionado directamente con la música popular en estilos como el pop o el rock (Paulus, 2009). Aunque dentro de la cultura occidental puede ser cierto, no conviene olvidar que el instrumento de percusión es considerado por los historiadores y antropólogos como el primer artilugio musical desarrollado por el hombre, y el cual ha sido utilizado para crear música desde que ésta es considerada como tal. Otros autores aluden la falta de trabajos a la naturaleza predominantemente melódica y armónica de la música occidental, lo que conlleva centrar sus estudios en los instrumentos que la generan, como el piano o la guitarra (Fitzgerald, 2004).

1.2.1. El kit de batería

El concepto de batería como unidad musical tocada por una sola persona se generó en el siglo XIX. Durante comienzos del siglo XX, con la invención de la música jazz, nació el kit de batería moderno donde los bateristas combinaban el sonido del bombo y la caja al mismo tiempo. Con el aumento de la popularidad del jazz durante la década de los años 30 la batería siguió su evolución añadiendo elementos para llevar el ritmo de las canciones consistentemente (inclusión del plato hi-hat). Durante la década de los 40 nació la música Bebop, permitiendo a los bateristas que la integraban desarrollar nuevas técnicas y composiciones más complejas, debido al carácter de improvisación que este estilo poseía. Asimismo, la reducción del tamaño del bombo y caja, así como la inclusión del plato de ride permitieron a los bateristas generar un ritmo más melódico y ligero. Con la incursión del rock and roll en los años 50 llega la explosión definitiva del instrumento, surgiendo en las dos siguientes décadas multitud de bateristas que revolucionan las técnicas de utilización aportando más velocidad, pegada y dinamismo, dotando al instrumento de personalidad propia más allá del aporte de simple base rítmica. En los 80 surgen las primeras baterías electrónicas, y hoy en día es común ver este tipo de módulos combinados con baterías acústicas dentro del ámbito profesional, tanto en grabaciones en estudio como en actuaciones en vivo.

La figura 1.4 muestra un ejemplo de una batería acústica actual [3], utilizada en multitud de estilos musicales como el jazz, blues, pop o rock y como elemento de enseñanza en escuelas musicales. Entre las piezas que la com-

[3]Modelo *Rock legend drum set* de la marca *Gretsch*.

Figura 1.4: 1. caja, 2. bombo, 3. plato hi-hat, 4. tom-tom aéreo, 5,6. tom-toms de suelo (goliat), 7. plato ride, 8,9 platos crash

ponen se pueden diferenciar dos tipos de instrumentos: los membranófonos, formados por la caja, bombo y tom-toms y los idiófonos, formados por los platos hi-hat, ride y crash.

Los membranófonos son llamados comúnmente tambores, que están compuestos por cascos (cuerpo), parches colocados a ambos extremos de los mismos y aros para sujetarlos. El material de los cascos suele estar fabricado en madera, dispuesto en varias capas superpuestas, aunque existen de otros materiales, como metal (sobre todo para la caja) o fibra de vidrio. Los parches son las membranas sobre las que se golpea y están hechos comúnmente de un material derivado del plástico (mylar), dispuesto en una o varias capas. El parche sobre el que se golpea se denomina "batidor" y el del otro extremo "resonante". El bombo, situado en el suelo, es el tambor más grande de la batería y generalmente se maneja con el pie a través de un pedal con una maza. La caja es el instrumento principal ya que es el encargado de marcar las notas principales de los compases. Se diferencia del resto de tambores porque incorpora una tira metálica, llamada bordonera, situada en contacto con el parche resonante que provee un característico sonido brillante y metálico. Por último, los tom-toms componen el resto de tambores, y proveen a la batería de diferentes sonoridades. La disposición y presencia de éstos varía de una batería a otra, aunque por regla general, dispone al menos de un tom-tom aéreo y otro de suelo, denominado goliat, de mayor tamaño. Tanto el material de los componentes, como la tensión que se aplica a los parches

sobre el casco y sobre todo, el tamaño del mismo (medido comúnmente con el diámetro y profundidad del tambor, en pulgadas) determinará la gravedad o agudeza del sonido que genere cada instrumento. Así, el bombo será el tambor que genere el sonido más grave de todos los elementos.

En cuanto a los idiófonos, llamados comúnmente platillos, son elementos rígidos de una aleación metálica (comúnmente bronce o latón), y sirven para dos propósitos: llevar y acompañar el ritmo mediante golpeos periódicos, caso del plato hi-hat y ride, y para dar énfasis, remarcar o acentuar compases y determinados pasajes, caso del plato crash. El hi-hat está compuesto por dos platillos normalmente separados en reposo, montados sobre un trípode y un pedal en su base, que acciona un mecanismo que los hace chocar. En función de la distancia entre los platillos en el momento del golpeo, el sonido generado cambia considerablemente (ver sección 1.2.2). Junto con el crash, existen otro tipo de platillos de menor tamaño, como el splash, o de distinta forma, como el china, que añaden diferentes efectos especiales y color a la interpretación. Tanto el tamaño de los platillos (medido con el diámetro, en pulgadas) como el grosor de los mismos influirá directamente en la intensidad, potencia sonora o *sustain* [4] que generen.

Por último, comentar que existen otros instrumentos, como cencerros, panderetas o bongos, que pueden ser añadidos a la batería y que su inclusión dependerá del estilo musical que se interprete y del gusto del músico de utilizar esa variedad de sonidos dentro de su interpretación.

Aunque la disposición de algunos elementos pueda cambiar dentro de cada kit, incluyendo o suprimiendo ciertas partes (tom-toms y platos de efectos), hay 3 de ellos que son comunes a todos los estilos y son utilizados para llevar el peso rítmico de la canción y sobre los que se asienta la base de gran parte de las composiciones contemporáneas. Estos serían el bombo (número 2 de la figura 1.4), la caja (número 1) y el hi-hat (número 3). No es de extrañar, por lo tanto, que la mayoría de estudios llevados a cabo se centren en la detección e identificación de sólo estos tres instrumentos, pudiéndose argumentar que si éstos son correctamente transcritos, los mismos principios pueden ser aplicados al resto de elementos (Paulus, 2009).

1.2.2. Características y singularidades sonoras

Tanto los membranófonos como los idiófonos se encuadran dentro de los instrumentos de sonido indeterminado (*unpitched percussion instruments*), lo que significa que las notas que generan no poseen una altura definida. La altura de un sonido viene dada por la frecuencia del mismo, y estos instrumentos, bien carecen de frecuencia fundamental [5] (por no generar una onda

[4] Tiempo de duración del sonido generado.

[5] La frecuencia más baja de una onda periódica.

periódica) o la enmascaran con ruido blanco [6] (caso de las cajas), o bien emiten sonidos inarmónicos [7]. Por supuesto, las baterías requieren ser afinadas (en este caso los tambores) para hacer que todos los elementos encajen sonoramente unos con otros, pero no significa que cada pieza deba conjuntar con alguna tonalidad específica de otros instrumentos musicales. Así pues, no existe un método estándar de afinación de la batería, y dependerá de varios factores, como el tipo de material, el grosor de los parches, la acústica del lugar en el que se toque o el estilo que se quiera interpretar. Se puede decir, por tanto, que la batería constituye un elemento de estudio diferenciado del resto de instrumentos, denominados de sonido determinado [8], por el hecho de carecer de tonos específicos en sus notas.

Cuando se presiona la tecla de un piano, por ejemplo, ésta está asociada a una determinada tonalidad, comienzo y duración (tiempo de pulsación de la tecla), que se deberán detectar para su correcta transcripción. En la batería la duración de la nota tampoco es información relevante, ya que los golpeos sobre una pieza determinada (ya sea tambor o platillo) del instrumento produce sonidos durante más o menos el mismo intervalo de tiempo.

Papanikas reduce en su trabajo la problemática de la transcripción de instrumentos de percusión a dos aspectos: La detección del inicio de la nota para saber cuando acontece y la determinación del instrumento o instrumentos fuente que la generan (Papanikas, 2012). Aunque ambos problemas serán abordados con más detenimiento en el capítulo 2, es evidente que será necesario aplicar técnicas que permitan extraer determinadas características sonoras de los instrumentos para poder detectarlos, establecer diferencias entre ellos y clasificarlos correctamente. La figura 1.5 muestra las señales acústicas de varios instrumentos, y su representación correspondiente en el dominio de la frecuencia (espectrograma). Se puede apreciar la diferencia espectral que generan los membranófonos, como el bombo y la caja, de los idiófonos, como el crash y el hihat. Mientras que los primeros poseen una gran energía espectral a bajas frecuencias (el bombo es el caso más acusado al tener un sonido más grave), en los segundos esa energía se reparte mayoritariamente por frecuencias de rango medio y alto. El caso del plato hihat es especial: mientras sus platillos permanecen unidos, su sonido es muy apagado (parte izquierda) mientras que a medida que se van abriendo, producen un sonido más brillante y duradero en el tiempo (parte derecha).

[6] Señal aleatoria donde sus valores de señal en dos tiempos diferentes no guardan correlación estadística.

[7] Aquellos que se desvían de las los múltiplos enteros de la frecuencia fundamental.

[8] Ejemplos son el piano, violín, guitarra, saxofón o trompeta.

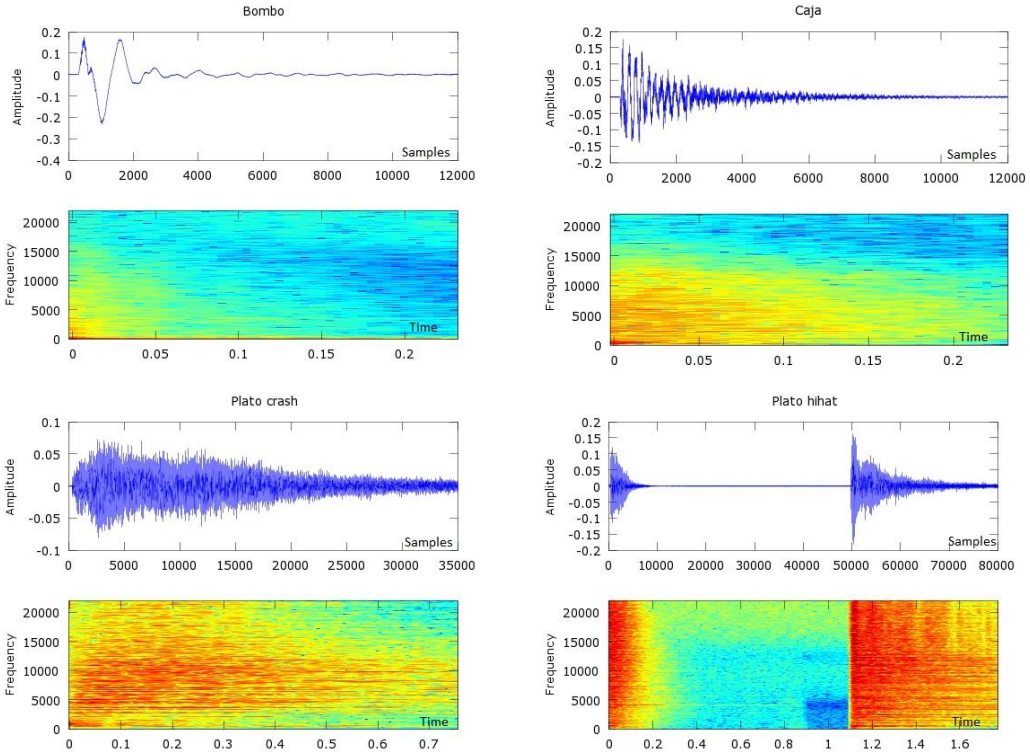

Figura 1.5: Señales acústicas y espectrogramas de varios instrumentos. Las zonas rojas del espectrograma denotan alta energía espectral en ese rango de frecuencias (eje ordenadas).

1.3. Motivación

Aunque muchos educadores coinciden en decir que la transcripción manual es un ejercicio valioso para el desarrollo de la habilidad musical, la motivación para la búsqueda de la transcripción automática permanece vigente. Los músicos con habilidades de transcripción limitadas podrán querer aprender a tocar una canción o pieza determinada, y dado que mucha de la música grabada hoy en día no tiene disponible su notación en papel, utilizarán estas herramientas automáticas de transcripción para poder generar la partitura correspondiente.

Por su naturaleza, el problema de transcripción automática es en muchos sentidos similar al reconocimiento automático del habla, pero no ha recibido un interés académico o comercial comparable. A pesar de ello, se

han llevado a cabo numerosos estudios al respecto, pero a día de hoy no se ha desarrollado un sistema de transcripción de propósito general aplicable de manera práctica (Klapuri, 2004). Actualmente no existe un método de transcripción automática capaz de extraer la información de una pieza de música polifónica de manera eficaz (transcripción completa). Sin embargo, se han obtenido buenos resultados en estudios donde el objetivo es dimensionado para obtener la melodía principal de una canción o los sonidos de percusión más prominentes para obtener el ritmo, que podrían ser definidos como transcripciones parciales (Papanikas, 2012). En otros casos, la señal de entrada es limitada a la presencia de un sólo instrumento.

Entre los trabajos publicados, cabe destacar la falta de aquellos encaminados exclusivamente a la transcripción de música generada por instrumentos de percusión. Este trabajo se centra en analizar las aproximaciones y métodos que los estudios más destacados han desarrollado sobre este tipo de instrumentos y proponer una solución a la problemática mediante un sistema de detección y clasificación de golpeos de batería utilizando la técnica de Modelos Ocultos de Markov (*Hidden Markov Models*, o *HMM*). Estudios anteriores en los que se basa el trabajo han demostrado ser una alternativa para establecer las bases de un sistema de reconocimiento de instrumentos de percusión en tiempo real (Paulus y Klapuri, 2009).

1.4. Organización del documento

El capítulo 2 explica la problemática de la transcripción automática en instrumentos de percusión y analiza diferentes soluciones basándose en los estudios realizados hasta la fecha, clasificándolos en función de la forma de analizar la señal. El capítulo 3 introduce conceptos de razonamiento probabilístico y profundiza en los *HMM*, como ejemplo de técnica utilizada para detectar y clasificar información a partir de señales temporales, como el habla o la música. Finalmente, en el capítulo 4 se aplica esta técnica para la transcripción automática de batería y establece una comparativa con otros estudios realizados previamente.

Capítulo 2

Estado del arte

Resumen: Este capítulo ahonda entre los diferentes estudios realizados en el campo de la transcripción automática. Después de una breve introducción histórica, se describe la problemática enumerando conceptos comunes y técnicas utilizadas dentro del campo de los instrumentos de percusión. Posteriormente se establece una clasificación de los métodos propuestos hasta la fecha en función de la forma de analizar la señal. Por último, se presenta la base de datos de dominio público *ENST-Drums*, como ejemplo de fuente de información sonora para la modelización de sistemas de transcripción en este tipo de instrumentos.

Los primeros intentos de transcribir música polifónica [1] fueron realizados por James A. Moorer en los años 70 en sus trabajos para transcribir composiciones a dos voces (Moorer, 1977). Posteriormente, sus estudios fueron continuados por Chafe et al. (Chafe et al., 1985). Independientemente, Piszczalski y Maher realizaron nuevas investigaciones durante la década de los 80, aunque en general, todos presentaban limitaciones tanto en el número de voces que podían tomar parte, como en los tonos que éstas podían interpretar (Piszczalski, 1986; Maher, 1989). En el campo del análisis del ritmo, los primeros intentos de transcribir instrumentos musicales fueron realizados en los años 80 por Schloss, que desarrollo el primer algoritmo de seguimiento de ritmo (*beat tracking*) en pistas de percusión y en los 90 por Bilmes. Ambos fueron capaces de clasificar diferentes tipos de golpes de conga (Schloss, 1985; Bilmes, 1993) . Durante los 90, Goto y Muraoka realizaron la primera transcripción de música polifónica de percusión (Goto y Muraoka, 1994b) y desarrollaron el primer algoritmo de seguimiento de ritmo para señales de audio completas (Goto y Muraoka, 1994a), basándose en parte en el trabajo previo desarrollado por Schloss.

[1] Referido a la señal donde varios sonidos pueden ocurrir simultáneamente.

A partir de ahí, el interés en este campo ha crecido rápidamente y durante los últimos 20 años han surgido proyectos de largo recorrido llevados a cabo por universidades de todo el mundo (Tampere University of Technology, University of London, Cambridge University). Se puede afirmar, sin embargo, que el rendimiento de los sistemas de transcripción musical son claramente inferiores a los de músicos profesionales en cuanto a eficacia, precisión y flexibilidad. En el ámbito de la transcripción de percusión, se han conseguido buenos resultados pero con limitaciones tanto en el número de instrumentos involucrados (utilizando generalmente bombo, caja y hi-hat) y en la presencia de otros instrumentos de altura determinada. Actualmente no existe un sistema de transcripción de propósito general que sea capaz de transcribir música sin restricciones en el grado de polifonía o del tipo de instrumento (Klapuri, 2004).

2.1. Descripción y análisis del problema

Los trabajos mencionados en este documento se refieren mayoritariamente a los diseñados para ser aplicados a baterías musicales utilizadas en música occidental, aunque muchos de los principios utilizados pueden ser aplicados a otro tipo de instrumentos de percusión. El bombo, la caja y el plato hi-hat son los tres los elementos de la batería que han sido utilizados más a menudo para desarrollar los sistemas de transcripción, ya que son los instrumentos fundamentales a la hora de llevar la base rítmica de la canción (siempre dentro del contexto de la música popular). La figura 2.1 muestra una gráfica denotando la presencia de instrumentos en la base de datos *ENST-Drums* (ENS), utilizada comúnmente en estudios de transcripción de instrumentos de percusión. Se observa que los instrumentos mencionados tienen presencia en más del 70 por ciento de los 80.000 eventos totales almacenados.

La tarea de transcripción más simple considera una señal donde solo un instrumento es tocado a la vez, pero el instante es desconocido. El problema de este tipo de señales es que no es muy realista. Existen sistemas más versátiles capaces de manejar una señal de percusión polifónica, donde varios elementos de la batería pueden ser tocados a la vez, por ejemplo, en la interpretación de un tema musical cualquiera. Por razones prácticas, sería deseable que un sistema de transcripción de batería pueda manejar música polifónica como señal de entrada, aunque la presencia de otros instrumentos puede provocar detecciones erróneas de presencia.

Para dar una estimación de la dificultad de los objetivos planteados por sistemas de transcripción de instrumentos de percusión, se podría comparar con lo que una persona sin educación musical es capaz de realizar: al escuchar una canción o una pista de batería, percibe el ritmo principal y puede marcarlo de forma sincronizada. También es posible que sea capaz de distinguir entre algunos elementos, como caja y bombo. Sin embargo, para poder trans-

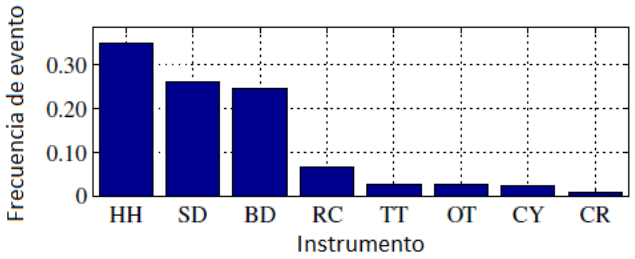

Figura 2.1: Frecuencias de instrumentos utilizados en la base de datos *ENST-Drums*. Los instrumentos son denotados como: BD (Bombo), CR (cualquier plato crash), CY (otro tipo de plato), HH (plato hi-hat abierto o cerrado), RC (plato ride), SD (caja), TT (cualquier tom-tom) y OT (otros instrumentos).

cribir una composición musical, es necesario adquirir una educación previa. Un oyente no entrenado es incapaz de reconocer entre diferentes patrones de batería, silencios, *fills* [2] y arreglos que acontecen en una interpretación, mucho menos traducirlo a un lenguaje musical. A medida que la complejidad polifónica de la percusión aumenta, mayor debe ser el entrenamiento requerido para separar los diferentes instrumentos que intervienen. Si, además, se incluyen otros instrumentos musicales en la señal (como en una canción), la dificultad de reconocimiento aumenta considerablemente. Lo mismo ocurre con un sistema de transcripción, donde la capacidad del sistema para obviar este tipo de información poco valiosa es fundamental.

De acuerdo con Klapuri, es útil estructurar el análisis de la transcripción y dividirlo en problemas de menor alcance (Klapuri, 2004). A continuación se presentan algunas de estas sub-técnicas, utilizadas comúnmente en la mayoría de estudios publicados (explicados en detalle en la sección 2.2). Asimismo, se presentan algunos conceptos utilizados para definir la estructura de clasificación.

2.1.1. Extracción de características

El propósito de la extracción de características es acondicionar la señal de entrada de manera que permita acceder a su contenido, con el fin de obtener valores numéricos que describan segmentos o partes que puedan ser reconocidas o agrupadas. También trata de reducir información no relevante en la señal. Aunque el grupo de características elegido puede ser seleccionado mediante "prueba y error", han sido evaluados algunos algoritmos de selección automático de características y en general presentan mejores resultados que

[2] Secuencia de sonidos rítmicos producido entre diferentes frases o pasajes interpretativos.

escoger estos descriptores de forma aleatoria (Herrera et al., 2002).

Las características usadas en los métodos publicados para instrumentos de percusión son a menudo parámetros acústicos que también han sido utilizados en aplicaciones de reconocimiento del habla y de otro tipo de instrumentos, destacando aquellos descriptores de tipo espectral, que recogen información relevante en el dominio de la frecuencia. Esto es corroborado por estudios donde, una vez analizada la efectividad en el ratio de clasificación, éstos obtienen mejor resultado en comparación con los descriptores de tipo temporal (Herrera et al., 2002). A continuación se describen los más importantes de ambos tipos.

2.1.1.1. Coeficientes Cepstrales en las Frecuencias de Mel

Los Coeficientes Cepstrales en las Frecuencias de Mel (*Mel Frequency Cepstral Coefficients*, o *MFCC*) han sido ampliamente utilizados en técnicas de reconocimiento del habla. Su éxito es debido a su capacidad para representar el espectro de amplitud del habla de una manera compacta. Gracias a investigaciones publicadas como la de Beth Logan (Logan, 2000), se ha podido asumir que esta extracción de características es también apta para el análisis de señales musicales.

Básicamente, se puede dividir la tarea de *MFCC* en 5 apartados:

- Dividir la señal de entrada en ventanas (*frames*). El carácter no estacionario de la señal musical [3], al igual que la del habla, provoca el hecho de tener que escoger un intervalo de tiempo para las ventanas suficientemente corto para tratar esta señal como casi-estacionaria. Generalmente se suele trabajar con ventanas de tamaño alrededor de 20 ms, y el solapamiento se lleva a cabo con desplazamientos entre ventanas de 10ms.

- Obtener el poder espectral de la señal. Para esta fase se aplica la Transformada de Fourier Discreta (*Discrete Fourier Transform*, o *DFT*) a cada ventana. De esta manera se realiza una conversión del dominio del tiempo, representada por la secuencia x_n, al dominio de la frecuencia (secuencia X_k), mediante la siguiente expresión:

$$X_k = \sum_{n=0}^{N-1} x_n e^{\frac{-j2\pi}{N}kn} \qquad (0 \leq k \leq N)$$

  siendo k índice de la frecuencia actual, i la unidad imaginaria y $e^{\frac{-j2\pi}{N}kn}$ la N-ésima ráiz de la unidad [4]. A partir de ese momento se descarta la fase y se trabaja con la envolvente de la señal $|X_k|$.

[3] La energía sonora que transmite varía a lo largo del tiempo.

[4] Números complejos que equivalen a la unidad cuando son elevados a una potencia dada n

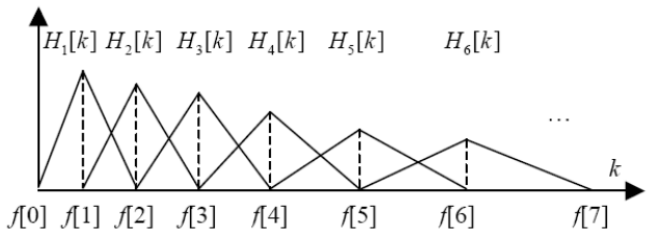

Figura 2.2: Banco de filtros MEL.

- Banco de Fitros. La señal obtenida en el anterior apartado se multiplica por un banco de F filtros triangulares de área unidad (ver figura 2.2). Estos triángulos están espaciados de acuerdo con la escala de frecuencias MEL, que se adecua mejor a la resolución de frecuencias del sistema auditivo humano. La frecuencia en la escala mel puede ser calculada con la siguiente fórmula:

$$F_{MEL} = 2595 log_{10}(1 + F_{Hz}/700)$$

  Una vez que la envolvente del espectro de la señal $|X_k|$ es multiplicada por el banco $H_m[k]$, se calcula la energía correspondiente en cada uno de sus filtros:

$$E_m = \sum_{k=0}^{N-1} |X_k|^2 H_m[k] \qquad (1 \leq m \leq F)$$

  siendo F el número de coeficientes que se quieren extraer.

- Se debe aplicar una compresión logarítmica de todas las energías obtenidas, con el fin de adecuarlas a la respuesta de intensidad no lineal del sistema auditivo. El inconveniente de trabajar en el dominio de la potencia espectral logarítmica es que los espectros de los filtros en las bandas adyacentes presentan un alto grado de correlación, originando coeficientes espectrales estadísticamente muy dependientes entre si.

- Con el fin de eliminar esta dependencia o correlación estadística, se aplica la Transformada Discreta del Coseno (*Discrete Cosine Transform*, o DCT). El DCT lleva los coeficientes espectrales al dominio de la frecuencia convirtiéndolos en coeficientes cepstrales:

$$Cmfcc_m = \sum_{k=0}^{N-1} log(E_k) cos(m(k - \frac{1}{2})\frac{\pi}{N}) \qquad (1 \leq m \leq F)$$

Normalmente, solo son utilizados como parámetros los coeficientes mas bajos (de 5 a 13), y en algunos casos, el coeficiente 0 correspondiente a la energía de la señal es descartado. Asimismo, muchas aplicaciones estiman la derivadas temporales de primer y segundo orden de estos valores ($\Delta MFCC$, $\Delta\Delta MFCC$), así como las medias y varianzas, para usarlas también como características.

2.1.1.2. Otros descriptores espectrales

Otro grupo de descriptores utilizados son aquellos de energía relativa. Estos descriptores dividen el espectro de la señal en un número determinado de bandas de frecuencia para calcular la energía de cada una de ellas. Posteriormente se calcula el porcentaje de energía relativa de cada banda con respecto a la total y se utiliza como característica. El número y el espacio entre bandas depende de la resolución de frecuencias deseada y puede ser elegido empíricamente en función de los resultados que se obtengan (Herrera et al., 2002), aunque a menudo el espaciado entre bandas es cercano a la escala logarítmica, para imitar el sistema auditivo humano (Paulus, 2009).

Otros descriptores más sencillos utilizados en los instrumentos de percusión incluyen valores escalares describiendo la forma del espectro. Para ello, se suele partir de un espectro de magnitud normalizada de la siguiente forma:

$$\widetilde{X_k} = \frac{|X_k|}{\sum_{k=1}^{K} |X_k|}$$

donde $\widetilde{X_k}$ representa el ratio entre la transformada de Fourier para la frecuencia actual k, y el mismo espectro con índices no negativos. Las características espectrales más comunes serían las siguientes:

- Centroide espectral (*spectral centroid*), definido como el centro de masas de la magnitud espectral.
- Dispersión espectral (*spectral spread*), que describe el ancho de banda de la señal espectral.
- Distorsión espectral (*spectral skewness*), que describe la asimetría de la distribución de la frecuencia alrededor del centroide espectral.
- Kurtosis espectral (*spectral kurtosis*), que describe la concentración de la distribución de la frecuencia.

2.1.1.3. Descriptores temporales

En comparación con los anteriores, los descriptores temporales han sido mucho menos utilizados para la clasificación de instrumentos de percusión. Normalmente, para describir la evolución del sonido en el tiempo, éste se

modela utilizando diferenciales de las características espectrales extraídas en ventanas de longitud reducida, que equivalen a sus derivadas de primer o segundo orden.

Entre los descriptores que pueden ser computados en el dominio del tiempo, los dos más comúnmente utilizados son el centroide temporal y el ratio de cruce por cero.

- El centroide temporal, que es un análogo directo del centroide espectral, describe el centro de gravedad temporal de la energía del sonido de la siguiente manera:

$$C_t = \frac{\sum_t tE_t}{\sum E_t}$$

  donde E_t equivale al Valor Cuadrático Medio (*Root Mean Square*, o *RMS*) de la señal en una ventana en el instante t, y el sumatorio es realizado sobre segmentos de longitud fija partiendo del comienzo del evento sonoro. Este descriptor permite discriminar entre sonidos transitorios y aquellos de larga duración.

- El ratio de cruce por cero computa el número de veces que la señal cruza este valor, independientemente del sentido en que lo haga. Este descriptor ha sido utilizado por Gouyon et al., como parámetro clave para clasificar sonidos de percusión (Gouyon et al., 2000). Es definido formalmente de la siguiente manera:

$$ZCT = \frac{1}{T}\sum_{t=1}^{T-1} I\{s_t s_{t-1} < 0\}$$

  donde s es la señal de longitud T en el instante de tiempo t y la función indicador $I\{A\}$ es igual a 1 si su argumento A es cierto y 0 si es falso. Este descriptor está correlacionado con el centroide espectral y el brillo percibido de la señal.

2.1.2. Detección de comienzo de nota

La detección del comienzo de nota (*onset detection*) engloba al conjunto de técnicas que permiten localizar los cambios sonoros en las señales musicales que sean provocados por variaciones tonales (caso de instrumentos de altura determinada) o por la ejecución de golpeos (caso de los instrumentos de percusión como la batería). Una detección correcta de cada evento sonoro abre la posibilidad de desarrollar aplicaciones para el análisis del contenido de la canción, búsqueda y almacenamiento de *loops* [5] similares en bases de datos [6] y por supuesto, la transcripción automática de instrumentos.

[5] Segmentos musicales de corta duración que pretenden servir de base a una composición más compleja.

[6] Un ejemplo de base de datos sería la web soundsnap, que almacena *loops* musicales de todo tipo de instrumentos (Sou).

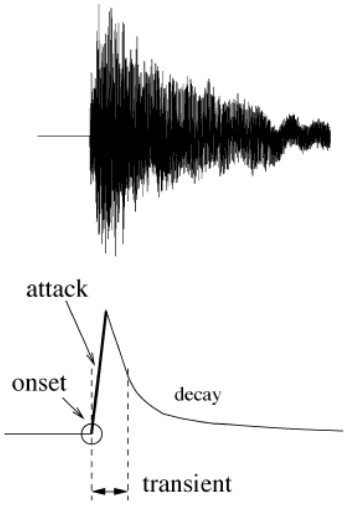

Figura 2.3: Representación de una nota y sus distintas partes: "comienzo", "ataque", "transitorio" y "caída".

A pesar de la existencia de numerosos métodos para abordar el problema, a día de hoy es todavía un área de investigación activa. Prueba de ello es la celebración anual de un concurso de detección de comienzo de nota llevado a cabo por MIREX [7], que es el marco de evaluación formal para aplicaciones de recuperación de información musical (auspiciado a su vez por ISMIR).

La figura 2.3 muestra la representación ideal de una única nota y su división en diferentes zonas. Se puede distinguir un comienzo abrupto de la señal (*onset*), una zona de transición en la que la envolvente varía rápidamente en el tiempo (*transient*) y una zona de decaimiento (*decay*). Tanto la zona transitoria como la de decaimiento no tienen una delimitación precisada y su establecimiento en el momento de implementar el sistema de detección puede dar lugar a diferentes resultados. Asimismo, el caso más realista de entrada es una señal de tipo polifónico, donde varios instrumentos intervienen al mismo tiempo. En el caso de un extracto de batería, es muy común que el plato hi-hat sea tocado a la vez que la caja o el bombo, y debido al carácter aditivo de las señales de audio (los instrumentos se superponen unos a otros) no es posible buscar cambios en la señal analizándola únicamente en el dominio del tiempo. En su lugar, es necesario realizar ciertas operaciones de procesamiento con las que obtener formas simplificadas de la señal que permitan la captura de "picos" sonoros. Así pues, se puede dividir el proceso de detección en varías etapas (Bello et al., 2005):

- Preprocesado de la señal. Esta etapa transforma la señal original para

[7] The Music Information Retrieval Evaluation eXchange (MIR)

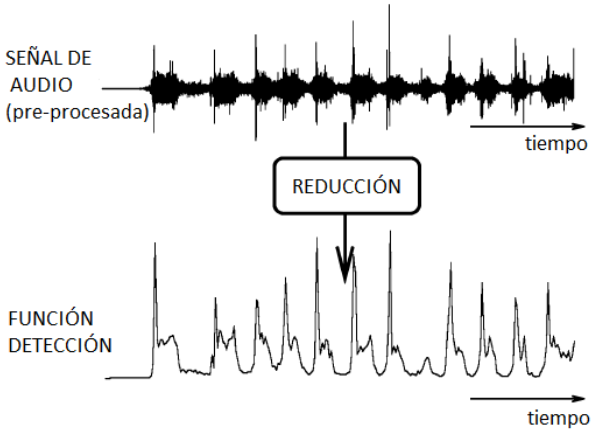

Figura 2.4: Representación de la etapa de reducción de la señal.

acentuar o atenuar diferentes aspectos en función del objetivo que se persiga. Aunque es una etapa opcional, se han desarrollado numerosas aproximaciones al respecto, entre las que destaca la desarrollada por Klapuri (Klapuri, 1999), que ha sido utilizada como base para posteriores implementaciones en trabajos de transcripción de instrumentos de percusión (Fitzgerald, 2004; Paulus y Virtanen, 2005; Papanikas, 2012). En ella, la señal es introducida en un banco de filtros no solapados donde se extraen las características espectrales por separado, para finalmente ser combinadas en la posterior etapa de detección (para una implementación real del algoritmo de Klapuri, ver (Ricard, 2005)).

- Reducción de la señal. Es la etapa más importante y consiste en acondicionar la señal pre-procesada para obtener la llamada función de detección, que pone de manifiesto la existencia de transitorios (variaciones bruscas) en la señal original. Esta reducción se puede llevar a cabo extrayendo características tanto temporales (aunque únicamente consigue buenos resultados en señales de instrumentos de percusión carentes de otros sonidos) como espectrales (mediante técnicas como la Transformada de Fourier de Tiempo Reducido, o *STFT*, y el Contenido de Alta Frecuencia, o *HFC*), o utilizando métodos basados en el supuesto de que la señal puede ser descrita por algún modelo probabilístico, y éste pueda ser cuantificado mediante medidas de probabilidad o criterios de selección Bayesianas. La figura 2.4 muestra el resultado de esta etapa. La señal de audio es convertida a una sucesión de picos que corresponden a los cambios tonales o golpeos.

  Basándose en los resultados conseguidos por Juan Pablo Bello en su estudio para la detección de instrumentos de percusión, se puede destacar

el método espectral de Contenido de Alta Frecuencia (*High Frequency Content, o HFC*) y el método estadístico *Negative log-likelihood* como los más adecuados para generar la función de detección (Bello et al., 2005).

- Captura del pico (*peak picking*). Si la función de detección ha sido correctamente diseñada, los eventos sonoros darán como resultado picos en la señal. La forma más común de aislarlos es mediante funciones de umbralización adaptativa [8], que a su vez generan un suavizado en la señal para evitar detectar falsos positivos, y finalmente, aplicar una función de detección de máximo local.

2.1.3. Definición de clases

Una vez detectado el instrumento o instrumentos que son tocados en cada instante, éstos deben ser etiquetados dentro de algún grupo. Para ello, es necesario definir clases en la que estén contenidos. Los trabajos realizados se centran en la utilización de dos tipos de clasificación:

- Clases de composición de instrumentos. Se deben definir tantas clases como combinaciones posibles de instrumentos puedan ser tocados a la vez. Por ejemplo, para un estudio en el que se tenga en cuenta la presencia de tres instrumentos, bombo, caja y hi-hat, la definición de algunas de las clases serían: clase 1: bombo, clase 2: caja, clase 3: bombo + caja, etc. El número de clases que se definirán será 2^n, siendo n el número de instrumentos a evaluar.

- Clases de detección del instrumento. En este caso, cada clase se utiliza para indicar la presencia o ausencia de un instrumento. Para el mismo estudio que en el anterior caso, se tendría: clase 1: ausencia bombo, clase 2: presencia bombo, clase 3: ausencia caja, clase 4: presencia caja, etc. El número de clases que se definirán será $2n$.

Como se puede observar, para el primer tipo de clasificación, su número crece de manera exponencial con el número de instrumentos considerados, lo que aumenta la complejidad de la aplicación. Por otro lado, puede haber clases para las que no haya ejemplos o instancias suficientes para ser entrenadas con fiabilidad. La figura 2.5 muestra una gráfica con las frecuencias de aparición de las 16 combinaciones de instrumentos más repetidas en la base de datos *ENST-Drums*. Se puede observar que, por ejemplo, la combinación bombo + plato hi-hat + caja, aunque formada por los tres instrumentos básicos, en conjunto no es utilizada con demasiada frecuencia. Es por ello que

[8] Segmentación de la señal y aplicación de un determinado umbral en función de los valores del entorno.

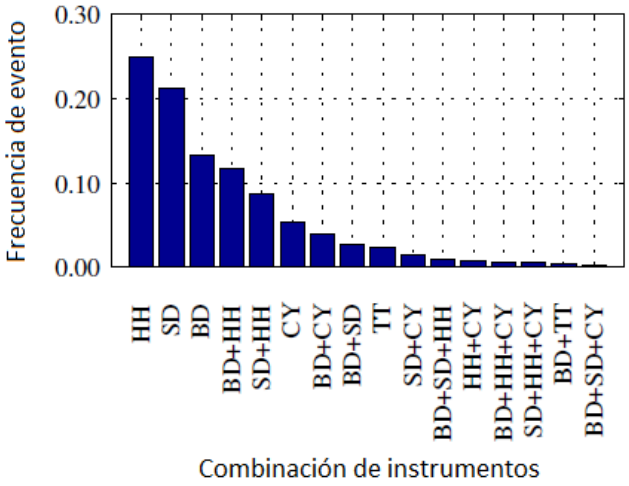

Figura 2.5: Frecuencias de combinación de instrumentos utilizados en la base de datos *ENST-Drums*. Los instrumentos son denotados como: BD (Bombo), CR (cualquier plato crash), CY (otro tipo de plato), HH (plato hi-hat abierto o cerrado), RC (plato ride), SD (caja), TT (cualquier tom-tom).

la mayoría de los estudios recientes se decanta por utilizar un método de clasificación basado en detección de cada instrumento separadamente(Paulus, 2009).

2.2. Aproximaciones

Tal como se indica en la sección 1.2.2, el objetivo de la transcripción es, por un lado, determinar cuando son tocados los instrumentos y de qué instrumentos se tratan. La secuencia y la forma en la que son abordadas estas dos problemáticas determinará el tipo de método aplicado.

De acuerdo con Fitzgerald, el problema de la transcripción de instrumentos de percusión puede ser dividido en dos categorias: "segmentar y clasificar" (sección 2.2.1), y "separar y detectar" (sección 2.2.2) (FitzGerald y Paulus, 2006). Posteriormente, Gillet y Richard añadieron un tercer grupo a la categoría, "emparejar y adaptar" (sección 2.2.3) (Gillet y Richard, 2008). Finalmente, Paulus añade un cuarto método basado en Modelos Ocultos de Markov (*HMM*) (Paulus y Klapuri, 2009). Esta técnica, que será introducida en el tema 3 y evaluada en el tema 4, implementa de manera conjunta las operaciones de segmentación y clasificación.

2.2.1. Segmentar y clasificar

Este método divide primeramente la señal de entrada en segmentos temporales y extrae características de cada uno de ellos por separado. Éstas son utilizadas para implementar un sistema de clasificación, encuadrando cada segmento dentro de alguna de las clases de instrumento o grupo de instrumentos previamente definidas.

La división de los segmentos se puede realizar de dos maneras diferentes (ver figura 2.6):

- La señal es dividida en tramos de longitud fija, creando una malla temporal. Según Bilmes, el período adecuado de estos segmentos debe ser aquel que coincida en mayor medida con todos los comienzos de las notas (Bilmes, 1993). Es un método que puede funcionar bien para extractos musicales con tempo constante y donde patrones de batería no se desvían demasiado de él (Gouyon y Herrera, 2001). En caso de que el período del segmento no esté correctamente ajustado mermará las capacidades de detección del método.

- La señal es dividida en tramos a partir de técnicas de detección del comienzo de nota (*onset detection*). De esta manera, el método asegura que en todos los segmentos divididos hay al menos una nota que se ha detectado. El principal inconveniente de este tipo de detección es que algunas notas pueden no ser detectadas debido a la distorsión de la señal o al nivel bajo en que ésta ha sido tocada. La mayoría de los métodos de "segmentar y clasificar" utilizan este último tipo de segmentación (Gillet y Richard, 2004), reduciendo el riesgo de "saltarse" notas bajando el umbral de detección, lo que también puede provocar la detección de falsos positivos debido al ruido o a la presencia de otros instrumentos.

Una vez extraídas las características sonoras a cada segmento (ver sección 2.1.1 para más detalles), se aplica una técnica de clasificación determinada. Algunas de las utilizadas hasta la fecha son:

- Arboles de decisión (*decission trees*) (Sandvold et al., 2004).

- Máquinas de Vector Soporte (*Support Vector Machines*, o *SVM*) (van Steelant et al., 2004; Tanghe et al., 2005; Gillet y Richard, 2008).

- K-vecinos más Cercanos (*K-Nearest Neighbours*, o *KNN*) (Herrera et al., 2002; Paulus y Klapuri, 2003; Sandvold et al., 2004).

- Comparación de patrones (*template matching*) utilizando la media cuadrática (*RMS*) del vector de características (Sillanpaa et al., 2000).

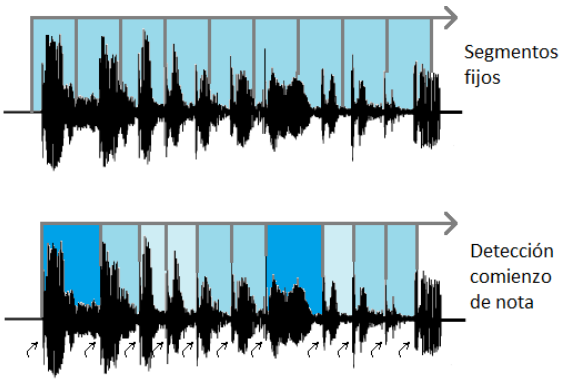

Figura 2.6: Representación de funcionamiento de los dos tipos de segmentación. En la segmentación fija se observa el desfase que se puede originar si la rejilla temporal no está adecuadamente calculada o el tipo de señal no es apta.

- Algoritmos de clustering mediante K-means (*K-means clustering*) (Gouyon y Herrera, 2001; Bello, 2006).
- Métodos estadísticos como Modelos de Mezcla de Gausianas (*Gaussian Mixture Models, GMM*) (Gillet y Richard, 2004).

Uno de los principales problemas ligados a este método tiene que ver con la generalización que se debe realizar para entrenar un modelo de un instrumento, sea tambor o plato. Aunque para una persona es fácil reconocer que diferentes cajas de distintas baterías representan un instrumento de la misma clase, las características acústicas de cada uno de ellos puede variar considerablemente. Esto provoca que en ocasiones sea necesario declarar clases por separado para el mismo instrumento tocado por diferentes modelos de baterías de estilos dispares como electrónico, hip-hop o heavy (Gillet y Richard, 2004).

2.2.2. Separar y detectar

Este método actúa de manera inversa al método "segmentar y clasificar". Primero descompone la señal de entrada en canales o "pistas" conteniendo cada uno de los instrumentos por separado y posteriormente extrae de cada canal los instantes en que el instrumento en cuestión interviene mediante técnicas de detección de comienzo de nota (ver sección 2.1.2).

Aunque existen diferentes técnicas para realizar la separación, hay ciertos requisitos comunes a todas ellas. Previamente, la señal debe ser transformada al dominio de la frecuencia. De esta manera, una señal de entrada X puede

ser definida como la suma de N espectrogramas individuales, cada uno de ellos correspondiente a un componente o instrumento:

$$X \approx \sum_{i=1}^{N} X_i$$

En el contexto de la transcripción de instrumentos de percusión, los espectrogramas individuales X_i son considerados como el producto de dos vectores, uno basado en la frecuencia s_i y otro basado en el tiempo a_i. :

$$X_i = s_i a_i^T$$

Combinando las dos expresiones anteriores se obtiene la siguiente expresión y se muestra su implementación en la figura 2.7.

$$X \approx SA$$

Es decir, se asume que cada instrumento produce un contenido fijo de frecuencias y un conjunto de ganancias que varían con el tiempo. Considerando una señal con tres instrumentos ($i = 3$), bombo, caja y hi-hat, se tendría el vector $S = [s_{bombo} s_{caja} s_{hi-hat}]$, donde $s_i = [s_i^1 s_i^2 s_i^3 ... s_i^m]$, siendo m el número de bandas de frecuencias consideradas y el vector $A = [a_{bombo} a_{caja} a_{hi-hat}]$, donde $a_i = [a_i^1 a_i^2 a_i^3 ... a_i^n]$, siendo n el número de ventanas o *frames* en el que ha sido dividida la señal de entrada.

A partir de estas consideraciones iniciales, existen dos vertientes diferenciadas para abordar el problema de separación de la señal:

- Métodos de separación no supervisado. Tratan de resolver el problema de separación sin conocer información previa de los instrumentos. En otras palabras, el número de instrumentos N, los vectores de frecuencia s_i y de ganancias a_i es desconocido a priori, teniendo como entrada solo la señal de entrada X. La técnica de Factorización de Matrices No Negativas (*Non-negative matrix factorization*, o *NMF*) puede calcular las matrices A y S de manera iterativa a partir de X, con la única restricción de que las tres matrices sean no negativas.

  Una vez obtenidos los canales s_i y a_i de cada componente i se extraen características frecuenciales y temporales para introducirlas en un clasificador con el fin de reconocer los instrumentos. El trabajo de Helen y Virtanen aplica esta técnica en música polifónica y clasifica mediante *SVM* para diferenciar entre pistas de batería y resto de intrumentos (Helen y Virtanen, 2005).

- Métodos de separación supervisada. Este tipo de técnica se vale de información más detallada de las propiedades de cada instrumentos. La técnica *Prior subspace analysis*, o *PSA* propuesta por Fitzgerald para

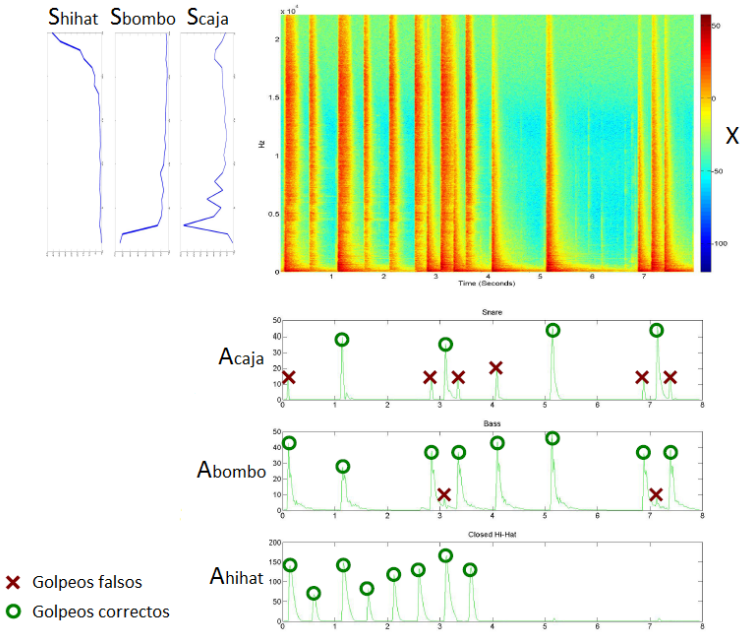

Figura 2.7: Representación de los vectores de frecuencias S y de ganancias variables con el tiempo A como una factorización del espectrograma X. La imagen ha sido extraída del trabajo de Papanikas (Papanikas, 2012).

reconocer golpeos de batería en presencia de otros instrumentos, debe incluir un conocimiento previo del contenido espectral de cada tambor o plato para generar una matriz incial de patrones S_{aprox} (Fitzgerald et al., 2003). Esta información será utilizada para obtener primeramente una aproximación de la matriz de ganancias temporales $\widehat{A}$, y posteriormente la matriz A aplicando la técnica de Análisis de Componentes Independientes (*Independent Component Analysis*, o *ICA*).

Otros estudios que aplican esta técnica muestran un ratio de valor-F [9] de 0.75 para separar caja y bombo en presencia de otros instrumentos (Spich et al., 2010).

Ambas técnicas, *NMF* y *PSA*, han sido combinadas en posteriores estudios como el de Paulus para aplicar *NMF* en técnicas supervisadas (Paulus y Virtanen, 2005). Posteriores trabajos han seguido utilizando este esquema para obtener inicialmente los patrones espectrales de cada pieza de la batería (normalmente bombo, caja y hi-hat) a partir de señales que contienen únicamente el sonido de cada una de ellas (Alves et al., 2009), (Papanikas, 2012). A partir de ahí, otros trabajos proponen modificaciones para obtener

[9] Medida que indica la precisión de un test. Para más información ir a la sección 4.3.1.

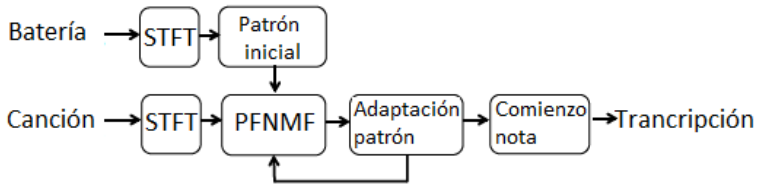

Figura 2.8: Técnica de adaptación del patrón mediante la separación de sonidos utilizando la técnica PFNMF. Las características espectrales son obtenidas de la Transformada de Fourier de Tiempo Reducido (*Short-time Fourier transform, STFT*)

una matriz S actualizada mediante técnicas semi-adaptativas durante la fase de iteración (Dittmar y Gartner, 2014). Esto es útil en aplicaciones donde se requiera recuperar el sonido producido por cada instrumento por separado.

2.2.3. Emparejar y adaptar

Aunque este método podría ser incluido dentro de los dos anteriores, en función de la arquitectura que presente, es separado de ambos por su carácter iterativo. Primeramente, es necesario generar unos patrones iniciales de cada instrumento o combinación de instrumentos a partir de sus características espectrales o temporales.

Aplicado como el método de "segmentar y clasificar", una vez obtenidos los segmentos y extraídas sus características, éstos serían comparados con los patrones para obtener el ratio de "fiabilidad" de que un determinado instrumento ocurra en ese segmento (basado en las distancias espectrales entre ambos). Los segmentos que muestren mayores grados de "fiabilidad" son utilizados para reajustar el patrón y el proceso se vuelve a repetir hasta que éste converja (Yoshii et al., 2007). Aplicado como un método del tipo "separar y detectar", los candidatos para adaptar los patrones son obtenidos mediante técnicas de separación utilizando *NMF* (Dittmar y Gartner, 2010) o variantes como *Partially Fixed Non-Negative Matrix Factorization*, o *PFNMF* (Wu y Lerch, 2015). Estas técnicas comienzan con patrones inicales y gradualmente los adaptan a patrones óptimos, iterando hasta que la diferencia entre el error de dos iteraciones consecutivas no supera un determinado valor o se alcanza un número máximo de éstas. El trabajo de Wu muestra unos resultados de precisión similares a otros estudios de transcripción con los que se ha comparado (Gillet y Richard, 2008; Paulus y Klapuri, 2009). Como principal ventaja de los métodos de descomposición por *NMF*, éstos requieren de pocos datos de entrenamiento para la generación de los patrones iniciales, lo que lo hace más aplicable a un caso de transcripción en el mundo real sin tener

que depender de una base de datos extensa.

2.3. Base de datos *ENST-Drums*

ENST-Drums surge ante la necesidad de proveer una base de datos que suponga el punto de partida para el desarrollo e implementación de técnicas de recuperación de información musical y transcripción de instrumentos de percusión. La falta de acceso libre y amplio a bases de datos de audio ha motivado la grabación de una variada gama de secuencias de percusión con el fin de cubrir tantas técnicas de análisis de señales de percusión como sean posibles. Para este propósito, se ha requerido la participación de tres bateristas profesionales que han realizado las grabaciones tanto acústicas como visuales, interpretando pasajes musicales de diversos estilos, como rock, pop, samba, soul, jazz en tres kits de batería diferentes, con el fin de obtener un espectro sonoro de cada instrumento lo más amplio y heterogéneo posible.

La duración total del material de audio grabado por cada baterista es de alrededor de 75 minutos, repartidos en más de 100 secuencias, almacenadas en ficheros en formato *wav* [10], con una frecuencia de sampleo [11] de 44.100 Hz y una profundidad de bit [12], o *bit depth*, de 16 bits. Estas se pueden dividir en varios tipos, y están resumidas en la tabla 2.1.

- Golpes individuales o *hits*: Son secuencias de grabación en la que el baterista golpea un mismo instrumento (tambor o plato) con varios segundos de diferencia y sin un patrón concreto.
- Fraseos: Compuestas por secuencias de percusión de varios estilos musicales, sin acompañamiento musical. Cada músico eligió interpretar algunos de los estilos propuestos a diferentes tempos (lento, medio y rápido) y dos niveles de complejidad (ritmos simples sin ornamentación y ritmos complejos con ornamentos y/o *fills*).
- Solos: Cada músico interpretó un mínimo de 5 solos en los estilos de su elección. Las instrucciones dadas fueron las siguientes: Un solo típico debe durar alrededor de 30 segundos, se deberían usar todos los instrumentos de los que se compone el kit y contener algunas secuencias complejas (en términos del número de instrumentos involucrados, de contenido rítmico y/o de tempo).
- Acompañamientos: Cada baterista interpreta 17 secuencias, tocadas sobre bases de acompañamiento utilizadas para enseñanza (denominadas *minus one*) y 24 secuencias sobre bases sintentizadas extraídas a partir de ficheros *MIDI*.

[10] *Wave form audio file format*, es un formato de audio digital sin compresión de datos.
[11] Número de muestras de señal recogidas en un segundo.
[12] Representa el número de bits de información por cada muestra.

	baterista 1		**baterista 2**		**baterista 3**	
Tipo	Sec.	Eventos	Sec.	Eventos	Sec.	Eventos
Hits	29	139	31	180	48	283
Fraseos	66	5339	74	9305	68	10467
Solos	7	1420	5	1613	5	1983
minus one	17	8856	17	8788	17	9382
MIDI	24	8224	24	6274	24	7357
Total	142	23978	151	26160	162	29472

Tabla 2.1: Número de secuencias y eventos (golpeos) grabados por cada baterista en la base de datos *ENST-Drums*.

Etiqueta	**Descripción**	**Etiqueta**	**Descripción**
bd	bombo	lmt	tom medio-bajo
sweep	escobillas	mt	tom medio
sticks	baquetas	mtr	*rim shot* tom medio
sd	caja	lt	tom bajo
rs	*rim shot* caja	ltr	*rim shot* tom bajo
cs	*cross stick*	lft	tom de suelo
chh	Hi-hat cerrado	rc	plato *ride*
ohh	Hi-hat abierto	ch	plato *china*
cb	cencerro	cr	plato *crash*
c	otros platos	spl	plato *splash*

Tabla 2.2: Etiquetas utilizadas en los ficheros de anotación de la base de datos *ENST-Drums*.

Para la grabación se han utilizado 8 micrófonos colocados alrededor de la batería con el fin de captar el sonido producido por un determinado instrumento o de recoger el sonido ambiente de la grabación. La señales son almacenadas en diferentes carpetas nombradas con el nombre del instrumento situado más cercano (carpetas *hi-hat*, *kick*, *snare*, *tom_1...*) o zona ambiente determinada (carpetas *overhead_L* y *overhead_R*). Asimismo, el conjunto de señales son editadas y mezcladas para la creación de dos tipos de audio estéreo, almacenados respectivamente en las carpetas *dry_mix* y *wet_mix*, siendo éste ultimo el resultante de una ecualización y compresión de los instrumentos más detallada.

Cada uno de los archivos de grabación, dispone de su consiguiente archivo de anotación, donde se indica el instrumento (tambor o plato) tocado y el instante en el que se produce el impacto. Este archivo es fundamental para recoger los eventos de una grabación y proceder a la extracción acústica de la

señal durante la fase de entrenamiento y testeo de los modelos. La tabla 2.2 muestra las etiquetas de los instrumentos que han sido interpretados durante las grabaciones.

Adicionalmente, *ENST-Drums* incluye dos canales de video con las interpretaciones realizadas por cada baterista tomadas desde ángulos diferentes. Al igual que las secuencias de audio, estas grabaciones pueden ser utilizadas como información para desarrollar un sistema audiovisual de transcripción musical (Gillet y Richard, 2005).

Capítulo 3

Introducción a los modelos ocultos de Markov

Resumen: Aunque los modelos ocultos de Markov (*Hidden Markov Models*, o *HMM*) han sido ampliamente utilizados en las últimas décadas en aplicaciones de reconocimiento del habla, su utilización en la transcripción de pasajes musicales es limitado, más aún en el caso de instrumentos de percusión. Esta falta de estudios realizados al respecto, y el hecho de que éstos hayan obtenido resultados prometedores y puedan ser comparados con las aproximaciones descritas en el apartado 2.2, ha motivado el desarrollo de un sistema de transcripción de este tipo. Este capítulo introduce esta técnica y establece una breve reseña histórica de su desarrollo y aplicaciones.

3.1. Razonamiento probabilístico

La probabilidad establece una manera de resumir la incertidumbre que se posee acerca de un problema, provocada principalmente por la ignorancia o falta de información. Para representar esta información probabilística y aplicarla con fines resolutivos se necesita un lenguaje formal.

Las afirmaciones probabilisticas hablan de los diferentes *mundos*, o situaciones que pueden darse en un problema, y expresan como es de probable que cada uno de estos mundos suceda. Un modelo de probabilidad plenamente especificado asocia una probabilidad numérica $P(m)$ con cada posible mundo de la siguiente forma:

$$(0 \leq P(m) \leq 1) \quad y \quad \sum_{m \in \Omega} P(m) = 1$$

Estableciéndose los axiomas básicos de la teoría de probabilidad, donde todo posible mundo tiene una probabilidad entre 0 y 1 y la probabilidad total del

conjunto de posibles mundos es 1. Las aseveraciones probabilisticas no vienen habitualmente descritas como probabilidades de que un determinado mundo ocurra, sino de que lo haga simultaneamente un determinado conjunto de ellos. Estos conjuntos son llamados eventos, y expresamos su probabilidad conjunta a través de las llamadas proposiciones ϕ:

$$P(\phi) = \sum_{w \in \phi} P(w)$$

Existen probabilidades no condicionadas o probabilidades a priori, que se refieren al grado de creencia de una proposición en ausencia de cualquier otra información adicional. En el supuesto caso de lanzar dos dados, la probabilidad de obtener en ambos el mismo valor sería una probabilidad a priori. La mayor parte del tiempo, sin embargo, se posee otra información relevante, a la que se suele llamar evidencia, que ha sido revelada con anterioridad y que sirve para definir probabilidades condicionales o a posteriori de una determinada proposición. Siguiendo con el mismo ejemplo, si se lanza primero un dado y sale tres, podemos calcular la probabilidad de obtener en ambos dados el mismo valor dado este supuesto. Esta probabilidad se escribe como $P(dobles|dado1 = 3)$ y se describe como la probabilidad de obtener el mismo valor en los dados $dado$ [1] que el primer dado lanzado tenga valor tres.

Matemáticamente, las probabilidades condicionales son definidas en términos de probabilidades no condicionales de la siguiente manera, dadas dos proposiciones a y b:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

siendo:

- $P(a|b)$ representa la probabilidad de que ocurra a habiendo b.
- $P(a \wedge b)$ representa la probabilidad de que ocurran a y b simultáneamente.
- $P(b)$ representa la probabilidad a priori de que ocurra b.

La probabilidad $P(a \wedge b)$ puede ser renombrada de maneras diferentes a través de la regla del producto:

$$P(a \wedge b) = P(a|b)P(b) \; o \; P(a \wedge b) = P(b|a)P(a)$$

Esta equivalencia viene del hecho de que, en la primera ecuación, para que a y b sean ciertas, se necesita que b sea cierta, y también que a lo sea habiendo b. Análogamente en la segunda, se necesita que a sea cierta, y también que b lo sea habiendo a. Introduciendo esta última expresión en la anterior ecuación

[1] Expresado a traves de "|".

se obtiene el denominado Teorema de Bayes, que subyace la gran mayoria de sistemas de inteligencia artificial con inferencia probabilística.

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$

Una manera sencilla de representar las relaciones de independencia y dependencia condicional entre las distintas variables que definen el funcionamiento de un sistema es a través de las redes bayesianas. Una red bayesiana es un grafo acíclico dirigido [2] donde:

- Cada nodo corresponde con una variable, que puede ser discreta o continua.
- Un conjunto de uniones dirigidas o flechas que conectan pares de nodos. Si hay algún nodo del nodo X al nodo Y, se dice que X es padre de Y.
- Cada nodo X_i tiene una distribución de probabilidad condicional que cuantifica el efecto que tienen los padres sobre el nodo (variable) a través de la expresión $P(X_i|Padres(X_i)$.

Probablemente, el modelo de red bayesiana más comúnmente utilizada dentro de aplicaciones de aprendizaje automático sea el modelo bayesiano ingenuo, o *naive bayes*, donde la clase variable C a predecir constituye la raiz de la red, mientras que las variables atributo X_i son los *hijos* de ésta (ver figura 3.1). Este modelo simplifica el problema de clasificación mediante la hipótesis de que la presencia o ausencia de una característica particular (atributo) no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable. Es decir, considera los atributos condicionalmente independientes entre si:

$$P(X_1, X_2, ..., X_n|C) = P(X_1|C)P(X_2|C)...P(X_n|C)$$

De esta manera, con los valores atributo observados $X_1, X_2, ..., X_n$, se puede predecir la probabilidad de que se dé la clase C:

$$P(C|X_1, X_2, ..., X_n) = \alpha P(C) \prod_i^n P(X_i|C)$$

siendo α un factor de escalado constante dependiente de las variables atributo.

[2] Grafo que no posee un camino directo que empiece y termine en un mismo nodo.

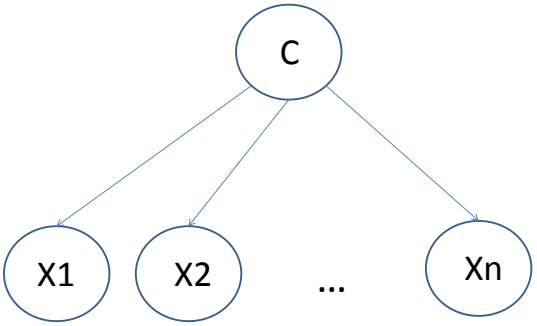

Figura 3.1: Esquema de una red bayesiana ingenua.

3.2. Probabilidad a lo largo del tiempo

Un subgrupo de las redes bayesianas son aquellas que modelan secuencias de variables, como señales del habla o música. Son llamadas redes bayesianas dinámicas y constituyen un caso especial en el que las variables incluyen una dimensión temporal. En este caso, el valor de una variable (o nodo) en el momento t puede ser calculado en función de un conjunto de regresores asociados al valor de la misma variable en instantes anteriores. Este concepto es definido como el modelo de transición, y especifica la distribución de probabilidad sobre los últimos estados de las variables, esto es, $P(X_t|X_{0:t-1})$. El mayor problema de este planteamiento es que el número de estados previos aumenta indefinidamente con el paso del tiempo, haciendo inviable su gestión en procesos largos o indefinidos. La propiedad de Markov, desarrollada por el matemático de origen ruso Andrei Markov, trata este problema mediante la afirmación de que el estado actual de una variable depende únicamente de un número finito y fijo de estados anteriores. Los procesos que satisfacen esta propiedad son denominadas cadenas de Markov, pudiendo haber de varios tipos. La cadena de Markov de primer orden es la más simple de todas. En ella, el estado actual de la variable depende únicamente del estado de la variable en el instante previo y no de ningún estado anterior. Se tiene por lo tanto, suficiente información para establecer un futuro condicionalmente dependiente del pasado de manera que el modelo de transición resultante sea $P(X_t|X_{t-1})$. El modelo de transición para una cadena de Markov de segundo orden será la distribución condicional $P(X_t|X_{t-1}, X_{t-2})$. La figura 3.2 muestra las estrucutras de ambos procesos.

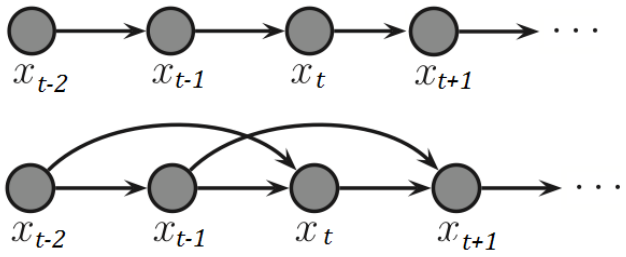

Figura 3.2: Estructura de una red baysiana correspondiente a un proceso de Markov de primer orden (arriba) y de segundo orden (abajo).

Este tipo de procesos establecen la diferenciación entre variables de estado X, no observables directamente, y variables de evidencia E, que si lo son. Para definir la relación entre ambos grupos de variables, se utiliza el modelo de observación, que establece la probabilidad de que se produzca una determinada evidencia observable en el tiempo t en función de las variables de estado en ese mismo instante, $P(E_t|X_t)$. A continuación se describe un modelo concreto de este tipo de arquitectura como son los modelos ocultos de Markov, o *HMM*.

3.3. Los Modelos Ocultos de Márkov

En las cadenas de Markov más simples, el estado es directamente visible al observador, y de esa manera, las probabilidades de transición entre los estados son parámetros del sistema. En los *HMM* el estado no es directamente visible, pero las variables de salida (evidencias u observaciones), dependientes del estado, si lo son (ver figura 3.3). Se considera a los *HMM* las redes bayesianas dinámicas más simples, donde el estado del proceso es descrito por una sola variable discreta. Los posibles valores de la variable son los posibles estados del mundo. Se trata de un proceso estocástico, es decir, está formado por variables aleatorias que evolucionan en función de otra variable, generalmente el tiempo. Cada estado tiene definida una distribución de probabilidad sobre los tipos de salidas, por lo que la secuencia de salidas generadas por un *HMM* da información sobre la secuencia de estados.

Los *HMM* pueden ser divididos en discretos o continuos. En los *HMM* discretos el espacio de estados de las variables ocultas y de observación es discreta, es decir, sólo pueden adoptar valores catégoricos. Por el contrario, los *HMM* continuos permiten adoptar valores reales para el espacio de estados y las observaciones, modelando las probabilidades comúnmente mediante distribuciones gaussianas.

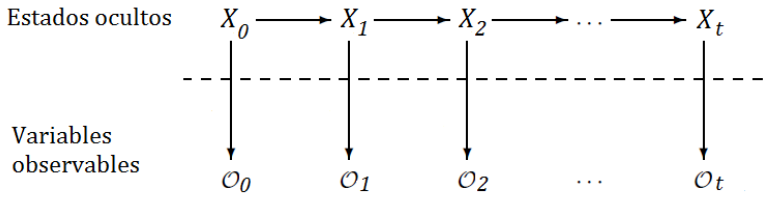

Figura 3.3: Esquema básico de un *HMM*. Los elementos X_i representan la secuencia de estados ocultos y los elementos O_i la secuencia de observaciones.

3.3.1. Elementos de un HMM

La notación utilizada para describir las partes que definen un *HMM* ha sido directamente rescatada del tutorial de Lawrence Rabiner, uno de los principales precursores para la utilización de los *HMM* dentro del campo de reconocimiento del habla (Rabiner, 1989). Así pues, un *HMM* está formado por los siguientes elementos:

- N, representa el número de estados del modelo. Aunque los estados estén ocultos, es decir, no se conozca su valor y por ello parezca una decisión trivial, el número de estados que componen un modelo a menudo tiene un significado o analogía física sobre el acontecimiento que se quiera detectar y predecir.
- Q, representa el conjunto de los distintos estados: $Q = (q_0, q_1, ..., q_N - 1)$
- M, representa el número de los distintos símbolos de observación. Este conjunto corresponde a la salida física del sistema que está siendo modelado.
- T, es la longitud de la secuencia de observación.
- O representa la secuencia de observación: $O = (O_0, O_1, ..., O_T - 1)$
- A, representa la matriz de distribución de probabilidad de transición de estados. Esta matriz $A = [a_{ij}]$ tiene un tamaño NxN, donde:

  $$a_{ij} = P(q_j|q_i)$$

  Es decir, cada valor de la matriz representa la probabilidad de estar en el estado q_j en el instante $t + 1$, habiendo estado en el estado q_i en el instante t.
- B, representa la matriz de distribución de probabilidad de los símbolos de observación. Esta matriz $B = [b_j(O_t)]$ tiene un tamaño de NxM,

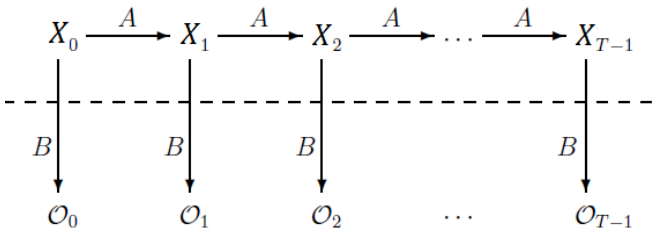

Figura 3.4: Esquema ampliado de un *HMM*.

donde:

$$b_j(O_t) = P(O_t|q_j)$$

Es decir, cada valor de la matriz representa la probabilidad de que se produzca una observación O_t (en el instante t) estando en el estado q_j.

- π, representa distribución de estados inicial, donde π_i es la probabilidad de que el primer estado sea N_i.

El esquema expuesto representa el caso donde las observaciones son caracterizadas como símbolos discretos elegidos de un alfabeto finito, por lo que sólo se podría utilizar densidades de probabilidad discreta dentro de cada estado del modelo. Este trabajo se centra en modelar una serie de *HMM*s con secuencias de observación formadas por señales continuas, cuyo dominio es el conjunto de números reales. Aunque es posible cuantizar y redimensionar la señal continua para conseguir observaciones discretas, se produciría una seria degradación asociada, por lo que la posibilidad de utilizar *HMM* con observaciones continuas es el camino más aconsejable (Russell y Norvig, 2014). De esta manera, se puede computar cada valor $b_j(O_t)$ mediante la siguiente distribución continua:

$$b_j(O_t) = \mathcal{N}[O; \mu_j, \Sigma_j]$$

Siendo $\mathcal{N}[O; \mu_j, \Sigma_j]$ un modelo de probabilidad formado por una gaussiana multivariante con vector de medias μ y matriz de covarianza Σ, esto es:

$$\mathcal{N}[O; \mu_j, \Sigma_j] = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(O-\mu)'\Sigma^{-1}(O-\mu)}$$

Asi pues, un *HMM* es definido por las matrices de distribución A y B y por el vector de probabilidad de estados inicial π. El conjunto se denota por $\lambda = (A, B, \pi)$. A continuación se describen brevemente dos de los algoritmos que se aplican en el contexto de los *HMM* y que son utilizados en el presente trabajo.

3.3.2. Aprendizaje del modelo. Algoritmo de Baum-Welch

Uno de los problemas relacionados con los *HMM* es el de encontrar un modelo que maximice la probabilidad de una secuencia de observaciones O, es decir, determinar el modelo que mejor explica esa secuencia. Esto se puede ver como el proceso de entrenar el modelo que mejor se ajuste a los datos observados. Para ello, se utiliza el llamado algoritmo esperanza-maximización o algoritmo EM, que es utilizado en modelos probabilísticos que dependen de variables no observables. Un caso especial de EM es el algoritmo de Baum-Welch, desarrollado por Leonard Baum y Ted Petrie y utilizado específicamente para determinar los parámetros que definen los modelos *HMM* (Baum y Petrie, 1966).

A continuación se describe esquemáticamente la secuencia de pasos a seguir:

- Inicializar los parámetros del modelo $\lambda = (A, B, \pi)$.
- Calcular las probabilidades de observación parcial $\alpha_t(i)$ y $\beta_t(i)$ para cada estado q_i a partir del algoritmo de avance-retroceso [3].

$$\alpha_0(i) = \pi_i(O_0); \quad \alpha_t(i) = [\sum_{j=0}^{N-1} \alpha_{t-1}(j)a_{ji}]b_i(O_t)$$

$$\beta_{T-1}(i) = 1; \quad \beta_t(i) = [\sum_{j=0}^{N-1} a_{ij}b_j(O_{t+1})]\beta_{t+1}(j)$$

- Calcular $\gamma_t(i,j)$ y $\gamma_t(i)$ para cada estado q_i a partir de las probabilidades de observación parcial obtenidas en el punto anterior. $\gamma_t(i,j)$ es la probabilidad de estar en el estado q_i en el instante t y trasladarse al estado q_j en el instante $t+1$. $\gamma_t(i)$ se define como la probabilidad de estar en el estado q_i en el instante t.

$$\gamma_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$$

$$\gamma_t(i) = \sum_{j=0}^{N-1} \gamma_t(i,j)$$

[3] Este algoritmo permite obtener una secuencia de observación O óptima dado un modelo $\lambda = (A, B, \pi)$

- Re-estimar los parámetros del modelo $\lambda = A, B, \pi$ y computar $P(O|\lambda)$.

$$\pi_i = \gamma_0(i)$$

$$a_{ij} = \frac{\sum_{t=0}^{T-2} \gamma_t(i,j)}{\sum_{t=0}^{T-2} \gamma_t(i)}$$

$$b_j(k) = \frac{\sum_{t \in 0,1,...,T-1; O_t=k} \gamma_t(i)}{\sum_{t=0}^{T-1} \gamma_t(i)}$$

La re-estimación es un proceso iterativo, que se repetirá desde el segundo punto si $P(O|\lambda)$ incrementa, o si no converge por debajo de un determinado umbral establecido previamente.

3.3.3. Secuencia de estados más probable. Algoritmo de Viterbi

El algoritmo de Viterbi permite hallar la secuencia más probable de estados ocultos (llamado camino de Viterbi) que produce una secuencia observada de sucesos. Es decir, dada una secuencia de observación determinada $O = (O_0, O_1, ..., O_T - 1)$, se quiere encontrar la secuencia de estados $Q = (q_0, q_1, ..., q_N - 1)$ óptima que mejor explica la secuencia de observaciones. Para ello se definen las variables $\delta_t(i)$, que representa la máxima probabilidad de encontrarse en el estado q_i de la secuencia parcial de observación hasta el instante t: $\delta_t(i) = \text{máx}_{q_1,q_2,...,q_{t-1}} P(q_1, q_2, ..., q_t = i, O_1, O_2, ..., O_t|\lambda)$. Puesto que el objetivo es obtener la secuencia de estados más probable, será necesario almacenar el argumento que hace máxima la ecuación anterior en cada instante de tiempo t y para cada estado j y para ello se utiliza la variable $\varphi_t(j)$. A continuación se detallan los pasos a seguir:

- Inicialización de la variable $\delta(i)$ en el instante inicial $t = 1$:

$$\delta_1(i) = \pi_i b_i(O_1) \qquad (1 \leq i \leq N)$$

- Recursión de la variable para los siguientes instantes de observación:

$$\delta_{t+1}(j) = [\max_{0 \leq i \leq N} \delta_t(i) a_{ij}] b_j(O_{t+1}) \qquad (1 \leq j \leq N)$$

  donde $t = 1, 2, ..., T - 1$.

  Obtención del máximo argumento que máximiza la función anterior y que se almacenará en la variable $\varphi_{t+1}(j)$

$$\varphi_{t+1}(j) = arg \max_{0 \leq i \leq N} \delta_t(i) a_{ij} \qquad (1 \leq j \leq N)$$

  donde $t = 1, 2, ..., T - 1$.

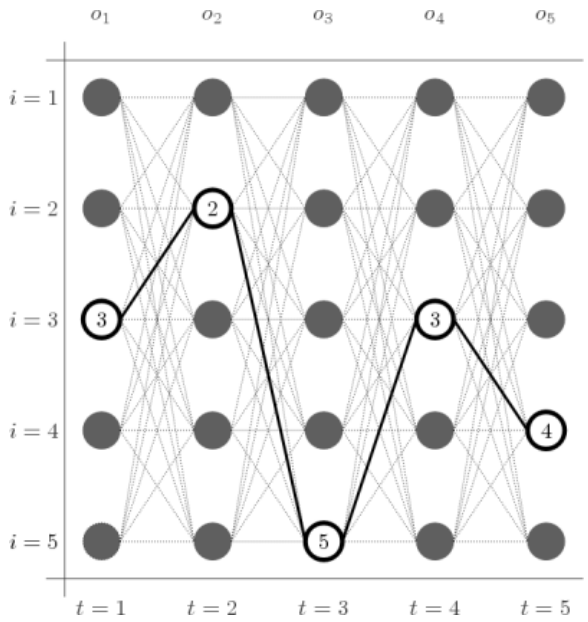

Figura 3.5: Ejemplo de secuencia de estados más probable producida por una secuencia de observación $O = (O_1, O_2, O_3, O_4, O_5)$.

- Obtención del estado más probable en el instante de observación $t = T$:

$$q_T^* = arg \max_{0 \leq i \leq N} \delta_T(i)$$

- Reconstrucción de la secuencia de estados más probable en los instantes de observación anteriores.

$$q_t^* = \varphi_{t+1}(q_{t+1}^*)$$

donde $t = T-2, T-1, ..., 1$.

3.3.4. Historia y aplicaciones

El modelo oculto de Markov y sus algoritmos asociados para inferencia y aprendizaje, incluyendo el algoritmo de avance-retroceso, fueron publicados en una serie de artículos por el matemático Leonard E. Baum junto a otros autores a finales de la década de 1960 y principios de 1970 (Baum y Petrie, 1966; Baum et al., 1970). El algoritmo de avance-retroceso fue uno de los precursores de la formulación general del algoritmo EM (Dempster et al., 1977). El algoritmo de Viterbi apareció en 1967 (Viterbi, 1967). A mediados de la década de los 70 empezaron a surgir los primeros reconocedores del habla basados en *HMM* (The DRAGON system, 1975).

A día de hoy, los *HMM* son especialmente utilizados para el reconocimiento de señales y formas temporales. Las aplicaciones abarcan campos diversos como reconocimiento del habla (Rabiner y Juang, 1993), reconocimiento de escritura manual (Kundu et al., 1988; Bharath y Madhvanath, 2007) y de gestos (Wilson y Bobick, 1999), traducción automática (Och y Ney, 2003), economia financiera (Bhar y hamori, 2004), biologia computacional (Krogh et al., 1994), criptoanálisis (Karlof y Wagner, 2003) o etiquetado gramatical (Kupiec, 1992).

Dentro de la transcripción musical, los *HMM* han sido motivo de estudio en aplicaciones de clasificación de pasajes musicales y estilos de música (Shao et al., 2004). Algunos de ellos se han centrado en la transcripción automática de instrumentos como el piano (Emiya et al., 2008). En cuanto a transcripción de instrumentos de percusión con esta técnica, se puede destacar el trabajo de Jouni Paulus por los buenos resultados obtenidos (Paulus y Klapuri, 2009) y el de Umut Simsekli (Simsekli et al., 2010). Este último desarrolla un algoritmo de detección y clasificación en tiempo real de sonidos de percusión monofónicos. Para ello, realiza experimentos sobre distintos modos de palmeo de manos y dos tipos de tambores turcos en varios escenarios, como una cámara anecoica [4] o una sala en silencio. Propone distintas consideraciones, como tratar el modelo de observaciones a partir de una distribución de Poisson o elegir un tamaño de ventana de extracción de características y un número de estados determinados para cada *HMM* en función de la reverberación del escenario. Los resultados muestran una precisión cercana al 85 % para un retraso permitido en la detección del golpeo de 20-30 ms, que puede ser no apto para determinados sistemas que requieren una respuesta inmediata, pero si para sistemas interactivos musicales (ver la sección 1.1.1 de aplicaciones para más información).

El trabajo de Paulus, por su parte, es la base sobre la que se ha cimentado las pruebas y experimentos de este trabajo, ya que se va a aplicar la misma técnica basada en *HMM* a partir de la base de datos *ENST-Drums*, detallada en el anterior capítulo (ver sección 2.3). Presenta dos tipos de detectores, basado en combinación de instrumentos y en presencia/ausencia, determinando que el último caso es el que mejor resultados ofrece. Para detectar una secuencia de silencio utiliza un modelo de mezcla de gaussianas con un estado y para los golpeos un modelo oculto de Markov de cuatro estados. Ofrece un resultado en tiempo real que es equiparable a otros estudios de diferentes metodologías realizados con anterioridad y deja entrever un camino de desarrollo en esta dirección. Con el fin de probar la fiabilidad del sistema *HMM* sin técnicas adicionales, se ha optado por variar su proposición con un sistema de presencia/ausencia con una distribución multivariante gaussiana para modelar las secuencias de observación (ver sección 3.3.1).

[4] Sala diseñada para absorber en su totalidad las reflexiones producidas por ondas acústicas o electromagnéticas.

Capítulo 4

Transcripción de batería mediante HMM

Resumen: A continuación se describe el conjunto de experimentos de transcripción realizados y basados en *HMM*. Se detalla y argumenta la parametrización de los distintos modelos *HMM*, eligiendo su topología, número de estados de que constan, tiempos de observación utilizados y características acústicas recogidas. Se enumeran asimismo las librerías adicionales que ha sido necesario añadir al trabajo, desarrollado en lenguaje Matlab y validado a través de la plataforma Octave. Por último, se analiza el proceso de selección de los modelos mediante pruebas de entrenamiento/validación y test y se presentan los resultados obtenidos en lecturas de audio en continuo.

4.1. Consideraciones previas

Para la implementación del sistema de transcripción, es necesario establecer unas premisas o condiciones fijas sobre las que modificar parámetros del sistema, con el fin de estudiar y encontrar el modelo con que mejor respuesta se obtenga. A continuación se enumeran las características que se mantendrán fijas a lo largo de la implementación.

4.1.1. Elección de instrumentos

El estudio se ha centrado en la detección de los tres instrumentos principales que forman parte de la base de datos *ENST-Drums* y que a su vez, representa el sonido básico de una batería para interpretar los patrones de la gran mayoría de estilos musicales contemporáneos. Estos instrumentos son la caja, el bombo y el plato hi-hat.

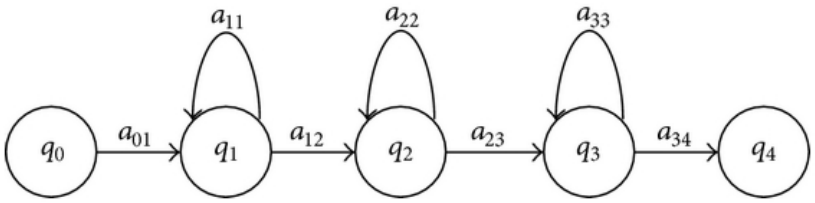

Figura 4.1: Esquema de un *HMM* de izquierda-derecha de 5 estados.

Independientemente de la fuerza con que se golpee cada instrumento, algunos de ellos pueden ofrecer un sonido diferente en función de su configuración. En el caso de la caja, está puede ser golpeada con o sin bordonera, ofreciendo peculiaridades acústicas distintas. Con el fin de simplificar el sistema implementado e intentar unificar ambos sonidos, se ha modelado un solo tipo de *HMM* que englobe ambas configuraciones (etiquetados en *ENST-Drums* como 'sd' cuando la caja es tocada con bordonera y 'sd-' cuando es tocada sin ella). El mismo criterio se ha seguido para el plato hi-hat, que tal como se explica en el apartado 1.2.2, produce sonidos muy diferentes en función de la distancia de los platos que lo forman. La base de datos *ENST-Drums* recoge dos tipos de etiquetas en función si el hi-hat es tocado con los platos separados ('ohh') o juntos ('chh'). El estudio de Paulus no hace distinción entre las distintas configuraciones, y modela un sólo *HMM* por cada instrumento (Paulus y Klapuri, 2009).

4.1.2. Topología y modo de detección

El primer paso para el entrenamiento de *HMM* es definir el prototipo de los modelos. Herramientas comerciales de reconocimiento del habla como HTK Toolkit (HTK) permiten utilizar una topología de izquierda-derecha para modelar los fonemas que componen las palabras. Este tipo de *HMM* es adecuados para modelar secuencias temporales ya que su arquitectura guarda analogía con este tipo de señales, asociando sucesivos segmentos con sucesivos estados. Para ello, el *HMM* se puede configurar de manera que sólo esté permitida la transición al mismo estado o al siguiente, sin posibilidad de volver a un estado anterior (ver figura 4.1). Por lo tanto, los modelos presentados en el trabajo se han diseñado en base a este tipo de arquitectura.

Asimismo, la forma de uso del *HMM* puede dar lugar a resultados diferentes en función de la información con la que se modele. Por un lado, se puede configurar el modelo por cada combinación de instrumentos tocada simultáneamente. Es decir, habría que modelar un *HMM* para la caja y otro para el bombo, así como otro para caja+bombo. Una opción menos confusa

establece que se modele un solo *HMM* por cada instrumento, y paralelamente, un *HMM* para las secuencias de silencio. La principal ventaja de esta opción es la drástica reducción de modelos totales necesarios para realizar el estudio. Asimismo, con este modo de funcionamiento se puede utilizar más eficientemente el instrumento a modelar durante la fase de entrenamiento que con los modelos de combinaciones. Otra diferencia es que cada instrumento tiene un modelo de silencio diferente, es decir, habrá un conjunto de dos modelos {instrumento, silencio} por cada instrumento a detectar (Paulus, 2009). Además, los modelos combinatorios pueden suponer un problema de *overfitting*, es decir, habrá ciertos modelos que no tengan suficientes muestras dentro de la base de datos que puedan ser entrenadas con fiabilidad, por lo que el modelo se ajustará en exceso a los pocas muestras existentes y previsiblemente no generalizará bien para secuencias no incluidas en la fase de entrenamiento/validación (ver sección 2.1.3). Por estos motivos, se ha decidido modelar los *HMM* como detector de cada instrumento por separado.

Una vez definida la arquitectura del modelo, todavía es necesario describir otras características del *HMM*, como son el número de estados que se compone y cual es la naturaleza de la transición entre los mismos. Las pruebas preliminares realizadas sobre la base de datos, en la que se ha testeado varios modelos con diferentes tamaños, ha determinado que el número de estados que formarán los modelos de silencio será dos y el que formarán los modelos de cada instrumento cuatro. Asimismo, se ha modelado la probabilidad de las secuencias de observación mediante una distribución gaussiana multivariante (ver sección 3.3.1). Para simplificar el modelo y los cálculos, se han utilizado matrices de covarianza Σ diagonales, considerando que las características que definen las observaciones no tienen una correlacion directa entre si.

4.1.3. Librerías complementarias

Se han necesitado dos librerías externas a la funcionalidad básica de *Octave* para la realización de los experimentos. Por un lado, la librería Rastamat (Ras), que permite recoger (entre otras características sonoras) los coeficientes cepstrales y sus derivadas de primer y segundo orden, pudiendo parametrizar para ello el número de componentes, el tamaño de la ventana temporal de extracción o el salto entre ventanas. Por otro lado, la librería H2M (H2M) ofrece la funcionalidad completa para entrenar y validar *HMM* en lenguaje Matlab.

4.1.4. Edición de audio

Para evaluar el rendimiento de los *HMM* en tiempo real (ver sección 4.3.3), se han utilizado las grabaciones de batería de *ENST-Drums* que vienen pro-

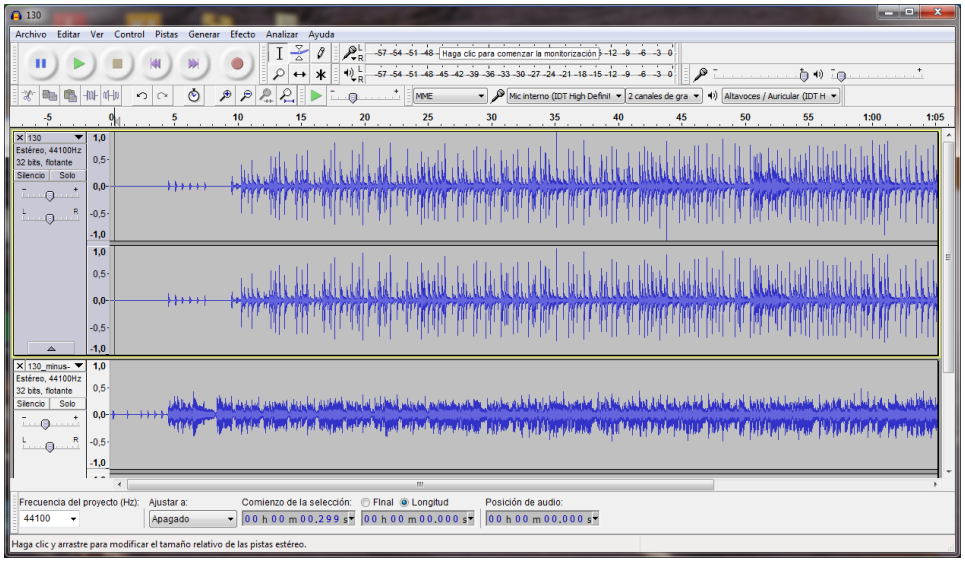

Figura 4.2: Programa *Audacity* desde donde se han generado los archivos de audio fusionando la pista de batería y la pista de acompañamiento musical.

vistas con sus correspondientes archivos de acompañamiento musical, con el fin de probar los sistemas en ambos escenarios. Para generar los archivos de audio con acompañamiento, se ha utilizado el programa *Audacity*, que es un software de grabación multipista y edición de audio.

4.2. Análisis de los experimentos

A continuación, se presenta un desglose de los experimentos realizados, detallando los procesos de recogida de eventos, extracción de características y el entrenamiento de los modelos.

4.2.1. Recogida de eventos

El proceso de recogida de eventos consiste en almacenar los instantes iniciales y finales de lo que se considera una secuencia temporal de golpeo y de silencio. Se considerará silencio a cualquier secuencia donde no haya presencia del instrumento en cuestión. Como se explica en la sección 2.3, *ENST-Drums* dispone de ficheros donde están almacenados los instantes de golpeo de cada instrumento para cada secuencia de grabación. Esos instantes representarán el instante inicial de la secuencia de un golpeo. El instante final vendrá determinado por el tiempo máximo de observación T_{obs} con el que se quiere entrenar el modelo. Así, si un instante de golpeo es susceptible de ser recogido (porque el instrumento en cuestión coincide con el tipo que se está

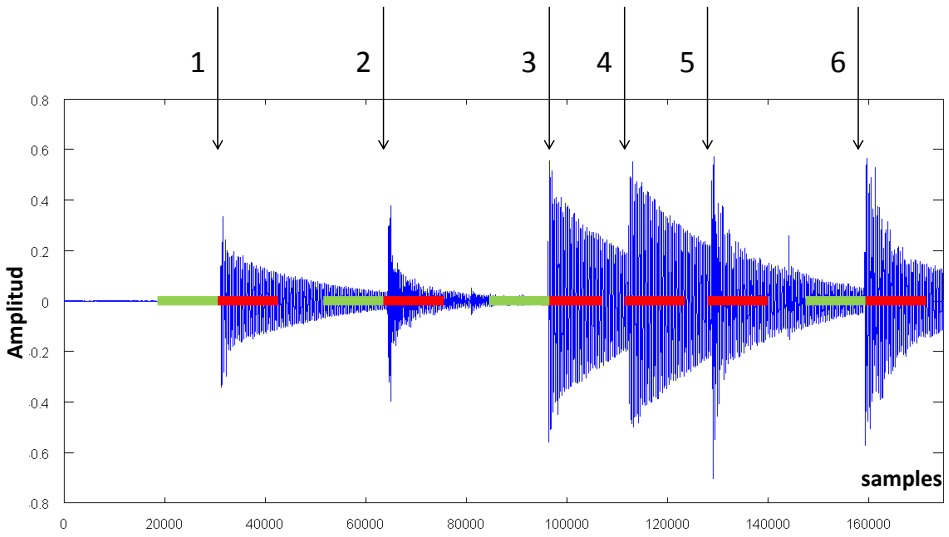

Figura 4.3: Representación de captura de secuencias de silencio (lineas verdes) y de instrumento (lineas rojas) para una determinada señal sonora.

entrenando), se comprueba que el siguiente golpeo no se produce a menos del T_{obs} utilizado. Si es así, y hay un golpeo temporalmente más cercano, el instante final se reducirá hasta el inicio del siguiente golpeo. De esta manera se asegura que la secuencia pertenece únicamente al sonido generado por un único golpeo del instrumento.

Para recoger las secuencias de silencio, se ha considerado recoger los instantes anteriores al inicio de un golpeo. Si ese golpeo en cuestión no tiene golpeos previos a una distancia temporal de $2 * T_{obs}$ se recogerá la secuencia inmediatamente anterior al golpeo, y de tamaño T_{obs}. Si por el contrario, un golpeo tiene otros golpeos previos cercanos, esa secuencia será descartada y no será considerada como silencio. La figura 4.3 explica gráficamente lo expuesto. Las lineas rojas representan las secuencias temporales de golpeo y las lineas verdes las de silencio, que se extraerán posteriormente del fichero de audio en base a los instantes iniciales y finales recogidos de cada secuencia. Como puede observarse, algunos golpeos que se encuentran temporalmente cercanos (eventos 3,4 y 5) no permiten extraer secuencias de silencio de entre ellos.

La recogida de eventos se ha programado de forma que tanto el número

total secuencias, así como la proporción del tipo golpeo con respecto a silencio sea parametrizable. Esta proporción se ha fijado en un 50 % para ambos tipos de secuencias, es decir, para todos los conjuntos de eventos (*event set*) que servirán para entrenar los modelos, el número de tipos de secuencias (instrumento y silencio) será el mismo. El *event set* estará dividido a su vez en tres subgrupos de tamaño similar, pertenecientes a cada uno de los tres bateristas que componen la base de datos. De esta manera, nos aseguramos que cada modelo es entrenado con información proveniente de todos los bateristas a partes iguales, evitando posibles problemas de falta de variabilidad. Para aumentar la diversidad de las muestras (o secuencias) recogidas en los *event set*, se ha establecido que el número máximo de secuencias por fichero de grabación quede limitado a un número de seis. De esta forma, se podrá conseguir conjunto de datos que abarque un mayor número de ficheros por cada baterista.

El número de muestras, que representa el número total de ejemplos de secuencias del conjunto de datos, con las que se entrenará cada modelo está preestablecido y será de $m = 96$, $m = 396$ y $m = 798$ para cada tiempo de observación T_{obs} (ver sección 4.2.2). Una de las cosas que se quiere analizar es si el tamaño que se usa en las muestras influye para mejorar los resultados o no. El objetivo es conseguir un modelo entrenado con un número de muestras suficiente que permite generalizar bien pero limitado, de manera que se puedan implementar sistemas fiables con relativamente pocos ejemplos.

Por otro lado, se ha establecido una semilla fija para obtener idénticos *event set* para un mismo tamaño de muestras. Es decir, para hacer pruebas comparables en modelos con un mismo conjunto de tamaño m, éste es siempre el mismo, extrayéndose de los mismos golpeos y silencios de la base de datos. Se trata de disminuir la variabilidad del sistema y encontrar el modelo óptimo para las mismas condiciones.

4.2.2. Procesamiento de los eventos

Cada secuencia de observación se va a descomponer en ventanas solapadas que serán procesadas por el *HMM*. Tal como se explica en el apartado anterior, los eventos se recogerán en base al tiempo máximo de observación T_{obs}, que representa el máximo valor que puede tener una secuencia de observación que es susceptible de ser entrenada y validada. Su valor se establece de la siguiente manera:

$$T_{obs} = W_{size} + (T \times H_{size})$$

Siendo,

- W_{size} el tamaño de la ventana de extracción de características.
- T el número máximo de ventanas que componen una secuencia de observación.

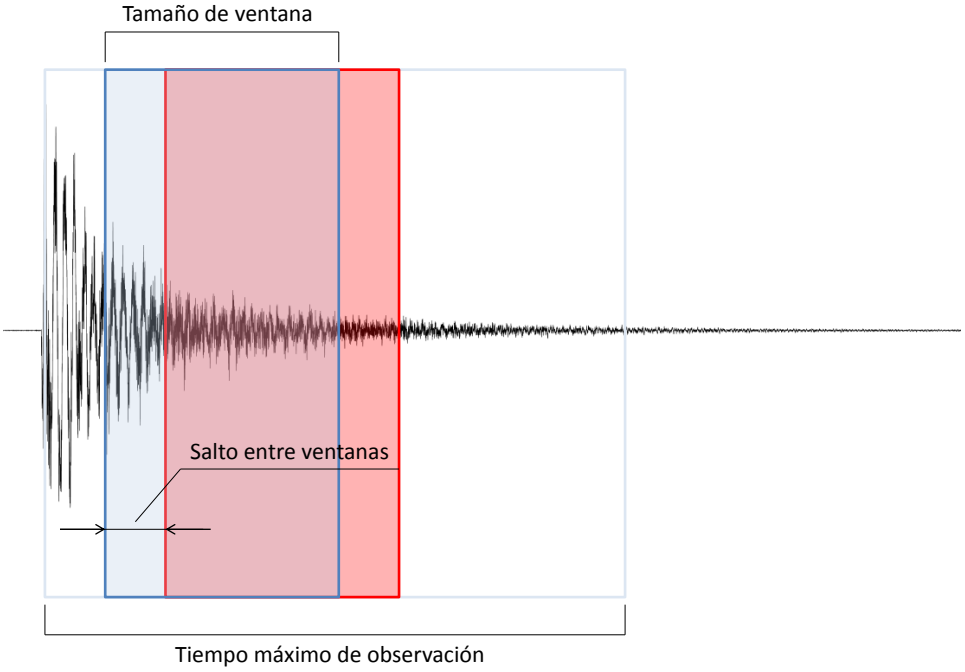

Figura 4.4: Representación de captura de características de la señal mediante ventanas solapadas.

- H_{size} el salto de tiempo entre ventanas consecutivas.

En definitiva, las secuencias de observación están formadas por ventanas de extracción solapadas entre si (ver figura 4.4). Una vez capturada la señal a partir de la información del *event set*, ésta es dividida en ventanas o *frames* de las que se extraerán sus características sonoras.

De esta manera, se entrenarán modelos con tres T_{obs} diferentes, definidos en función del tamaño de las ventanas y el salto entre ellas. El número máximo de ventanas T que se extraerán de una secuencia será siempre diez, independiente del T_{obs} considerado. Se han elegido los valores de salto de manera que el solape entre ventanas tenga siempre la misma proporción, en este caso del 80 %. Así pues, se tendría:

- Tiempo de observación de 150 ms, con tamaño de la ventana de 50 ms y salto de 10 ms.
- Tiempo de observación de 120 ms, con tamaño de la ventana de 40 ms y salto de 8 ms.
- Tiempo de observación de 75 ms, con tamaño de la ventana de 25 ms y salto de 5 ms.

El siguiente paso es extraer las características a partir de las ventanas que conforman las secuencias de observación. Estas características extraídas serán las observaciones que alimentarán el *HMM*. El trabajo se ha centrado en la extracción de los coeficientes cepstrales y sus derivadas deltas y delta-deltas. Estudios similares deshechan la incorporación de otras características espectrales al no encontrar mejoras en los resultados (Paulus y Klapuri, 2009). Se han realizado pruebas adicionales para corroborar esta afirmación (ver sección 4.3.2). Así, para cada conjunto m, se ha recogido un número de características $n = 13$, que equivale a los 13 primeros *MFCC*, $n = 26$, que engloba al anterior grupo y sus deltas y $n = 39$, que engloba el anterior conjunto además de sus delta-delta.

La naturaleza de estas características es de tipo continuo, lo que hace que el *HMM* tenga que modelar la matriz de distribución de las secuencias de observación, definida a partir de las medias y varianzas de cada tipo de característica.

El resultado final de la extracción de características es el conjunto global de datos (*data set*) que servirá para entrenar, validar y testar el modelo. Se entrenarán por lo tanto, un total de 27 modelos por instrumento (y su correspondiente modelo de silencio) combinando los valores de tres tiempos de observación T_{obs}, tres tamaños muestrales m y tres conjuntos de características n.

4.2.3. Entrenamiento/validación y test del modelo

Una vez generado el *data set* de tamaño muestral m, éste se divide en 2 grupos: los datos de entrenamiento (*training set*) y los datos de test (*test set*), con una proporción fija del 80 % y 20 % del total respectivamente. El *training set* servirá para aprender las características de las secuencias observadas y validar su efectividad. Para ello, se procederá a realizar validación cruzada con K=10 iteraciones (*10-fold cross validation*) sobre el conjunto total de los datos de entrenamiento. Los datos son divididos en 10 subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba o validación (*validation set*) y el resto (K-1) como datos de entrenamiento. El proceso es repetido durante K iteraciones, con cada uno de los posibles subconjuntos de datos de prueba, para finalmente, realizar la media aritmética sobre el error obtenido en cada iteración (ver figura 4.5).

Esta acción se repite 10 veces para el mismo *training set* con semillas aleatorias diferentes, obteniendo finalmente una media de todas las validaciones cruzadas. El objetivo es conseguir un resultado que limite la variabilidad y pueda desvirtuar el rendimiento del modelo entrenado. Para estudiar la tendencia del *data set* recogido, se realizará esta acción para diferentes tamaños del *training set*. Esto permitirá ver la efectividad del modelo entrenado para un tamaño muestral determinado, y extraer conclusiones respecto a la necesidad de incluir más muestras y características para el diseño del modelo.

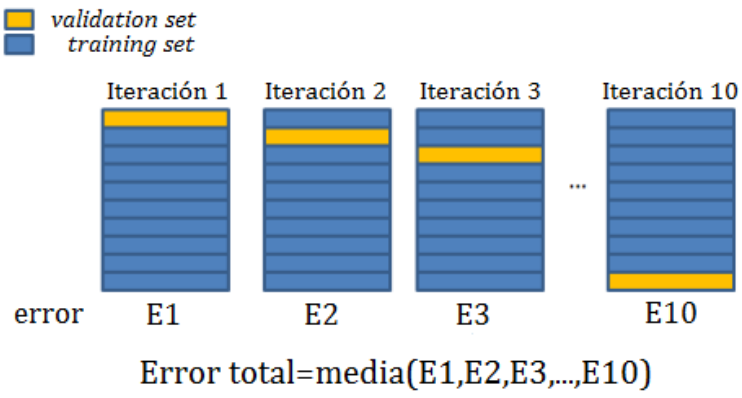

Figura 4.5: Representación de la técnica *K-fold cross validation* para K=10.

Finalmente, los 27 modelos entrenados y validados para cada instrumento se probarán en el *test set*, formado por muestras no utilizadas para el entrenamiento. El modelo que mejor ratio de predicción ofrezca, será el utilizado en la fase de evaluación en tiempo real, o continuo.

4.2.4. Clasificación con el *HMM*

Para clasificar las secuencias entre un modelo instrumento y un modelo silencio, se aplicará el algoritmo de Viterbi (ver sección 3.3.3) y se compararán los valores máximos de $P_i(O_1, O_2, ..., O_T|\lambda)$ de los estados i en ambos modelos en el instante T, que indican la máxima probabilidad de que se dé una secuencia de observación completa $O = (O_1, O_2..., O_T)$, dado un modelo λ:

$$\delta_T(i) = \max_{0 \leq i \leq N} P_i(O_1, O_2, ..., O_T|\lambda)$$

Donde $P(O_1, O_2, ..., O_t|\lambda)$ equivale a la función de verosimilitud calculada a partir de la función de densidad gaussiana [1] f dependiente del modelo λ:

$$P(O|\lambda) = \mathcal{L}(\lambda|\mathcal{O}) = f_\lambda(O)$$

Comparando el mayor valor de $\delta_T(i)$ en ambos modelos, se determinará a que tipo pertenece la secuencia en cuestión. Esto se aplicará de igual manera para la fase de evaluación en continuo 4.3.3.

[1] Describe la verosimilitud relativa para que una variable aleatoria tome un determinado valor dentro de la función.

4.3. Resultados

En esta sección se presentan los resultados obtenidos para los distintos modelos *HMM* planteados, con las topologías, configuraciones y parámetros establecidos en las anteriores secciones. Se elegirá el conjunto de modelos instrumento y silencio que mejor resultado ofrezcan en los datos de test.

4.3.1. Medidas de error utilizadas

Las matrices de confusión es una herramienta utilizada para evaluar el rendimiento y efectividad de un modelo de predicción en técnicas de aprendizaje supervisado, donde los datos de entrenamiento son presentados en pares de objetos: un componente de entrada formado por las características que definen el sistema y un componente de salida formado por el tipo de clase en el que está clasificado. Durante las fases de entrenamiento/validación y test de los modelos la información acerca de la capacidad de predicción es almacenada en estas matrices, las cuales están compuestas por los siguientes parámetros:

- VP: Número de verdaderos positivos, es decir, golpeos del instrumento que han sido detectados como tal.
- VN: Número de verdaderos negativos, es decir, secuencias de silencio que han sido detectados como tal.
- FP: Número de falsos positivos, es decir, secuencias de silencio que han sido detectados como golpeos del instrumento.
- FN: Número de falsos negativos, es decir, golpeos del instrumento que han sido detectados como silencio.

A partir de estos datos, se pueden obtener distintos estimadores que indican la capacidad del modelo en diferentes aspectos. Para medir la efectividad de los modelos se ha utilizado como parámetro la *exactitud*, que establece la proporción de resultados verdaderos entre el número total de muestras examinadas:

$$exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

Otra manera de examinar el rendimiento del clasificador es a través del espacio ROC, que representa gráficamente el ratio de verdaderos positivos o *sensibilidad* del sistema frente al ratio de falsos positivos o *(1-especificidad)*:

- RVP es el ratio de verdaderos positivos, o *sensibilidad*:

$$RVP = \frac{VP}{VP + FN}$$

- RFP: Es el ratio de falsos positivos:

$$RFP = \frac{FP}{FP + TN}$$

La *sensibilidad* mide hasta qué punto el clasificador es capaz de detectar o clasificar los casos positivos (en este caso, detectar correctamente los golpeos del instrumento), de entre todos los casos positivos disponibles. Por otro lado, el ratio de falsos positivos define cuántos resultados positivos son incorrectos de entre todos los casos negativos disponibles. Cada resultado de predicción de un modelo representa un punto en el espacio ROC. El mejor modelo posible de predicción se situaría en la coordenada (0,1) del espacio ROC, representando un 100 % de *sensibilidad* y un 100 % de *especificidad*. El análisis ROC proporciona herramientas para seleccionar los modelos posiblemente óptimos y descartar modelos subóptimos independientemente del coste de la distribución de las dos clases sobre las que se decide. La figura 4.6 muestra las 3 gráficas ROC obtenidas de evaluar los distintos modelos planteados para el instrumento caja.

Una manera de comparar los puntos del espacio ROC es a través de la ecuación que mide la distancia de cada uno de ellos con la coordenada (0,1):

$$Roc_d = 1 - \sqrt{W * (1 - RVP)^2 + (1 - W) * RFP^2}$$

donde W es el peso que se le quiera dar a uno u otro parámetro. En este caso se ha establecido en 0.5.

Aunque no se ha utilizado la distancia ROC para seleccionar el modelo final para la fase de evaluación en continuo, se ha calculado en todos los modelos con el fin de extraer conclusiones en cuanto a la actuación de cada uno de ellos, ya que al ser una magnitud que agrega dos magnitudes permite sacar conclusiones de forma más sencilla.

La evaluación de un sistema de transcripción en modo continuo es relativamente sencilla, una vez se han establecido las tipos de instrumentos sobre los que se trabajarán. Para cada instrumento, la anotación real y los resultados son comparados entre si. Cada evento detectado en el sistema debe ser emparejado solo con otro evento de la anotación real. Para dar validez a este emparejamiento, se debe establecer una desviación de tiempo que delimita la máxima diferencia de desfase en la que ambos eventos ocurren dentro de la canción. No hay una regla establecida para determinar este valor de desviación, pudiendo ir desde los 25ms (Yoshii et al., 2007), hasta los 50 ms (Gillet y Richard, 2008). Una vez todos los eventos han sido emparejados, es posible calcular el ratio de rellamada (*recall rate*), o *exhaustividad*:

$$R = \frac{Eventos\ Correctos}{Eventos\ Anotados} = \frac{VP}{VP + FN}$$

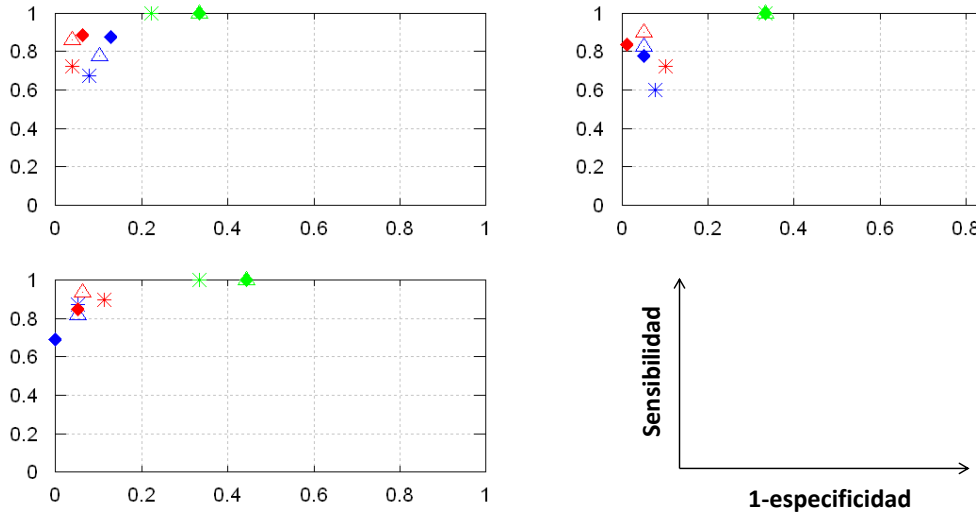

Figura 4.6: Espacio ROC para todos los modelos entrenados de la caja. Cada gráfica indica el resultado de los modelos para los diferentes tiempos de observación. Dentro de cada gráfica los puntos de color verde indican los modelos entrenados para un espacio muestral $m = 96$, los azules para $m = 396$ y los rojos para $m = 798$. Dentro de cada espacio muestral, los símbolos $\star$ son los modelos entrenados con $n = 13$ características, los simbolos $\triangle$ con $n = 26$ y los símbolos $\diamondsuit$ con $n = 39$.

y el ratio de *precisión* (*precision rate*):

$$P = \frac{Eventos\ Correctos}{Eventos\ Transcritos} = \frac{VP}{VP + FP}$$

A partir de estos dos ratios se puede obtener el Valor-F (*F-score o F-measure)* como un valor único ponderado de ambos, y que se puede utilizar como medida de precisión del estudio:

$$F = \frac{2RP}{(R + P)}$$

caja		T_{obs} 75ms		T_{obs} 120ms		T_{obs} 150ms	
		Train/val	*Test*	*Train/val*	*Test*	*Train/val*	*Test*
m=96	n=13	90.9	**89.4**	93.9	84.2	92.0	**84.2**
	n=26	99.8	84.2	99.0	84.2	94.5	78.9
	n=39	100	84.2	98.3	84.2	93.8	78.9
m=396	n=13	91.3	79.7	88.2	75.9	89.4	**91.0**
	n=26	92.4	83.5	90.9	**88.6**	93.6	88.4
	n=39	91.9	**87.3**	91.2	86.0	93.7	84.6
m=798	n=13	89.7	84.2	85.8	81.1	87.8	89.2
	n=26	91.6	**91.1**	91.1	**92.4**	92.2	**93.6**
	n=39	92.5	**91.1**	91.5	91.1	91.8	89.8

Tabla 4.1: Resultados de exactitud en % de los modelos *HMM* de caja y silencio sobre los datos de entrenamiento (*Train/val*) y los datos de test (*Test*).

.

4.3.2. Evaluación del modelo *off-line*

A continuación se presentan las tablas con los resultados en la fase de entrenamiento/validación y test de los tres instrumentos evaluados.

4.3.2.1. Caja

La tabla 4.1 muestra los resultados obtenidos para la caja. Dentro de la fase de entrenamiento/validación se observa que, para un mismo tiempo de observación T_{obs} y para el mismo conjunto de datos m, la exactitud del clasificador mejora considerablemente si utilizamos un número de características $n = 26$ y $n = 39$. Es posible que para $n = 13$ el modelo no sea capaz de clasificar convenientemente debido a un problema de sesgo, o *bias*. Es decir, le faltan características para ajustar convenientemente las observaciones al tipo de modelo que pertenece. Los resultados para $n = 26$ y $n = 39$ en el entrenamiento/validación son similares para todos los escenarios, por lo que se evaluará el rendimiento de ambos tipos de modelos en la fase de test.

Los resultados de clasificación en el *test set* muestran que para un mismo tiempo de observación se observa una mejora sustancial a medida que se aumenta el conjunto de datos. Esto puede ser debido a que el modelo generaliza mejor con más datos durante la fase de entrenamiento y su efectividad aumenta durante la predicción del *test set*. Asimismo, se observa una mejora en la predicción a medida que se aumenta el tiempo de observación T_{obs} para un mismo conjunto de datos m. Los picos máximos de *exactitud* se obtienen para un número de características $n = 26$ en todos los tiempos de observación. El modelo con $T_{obs} = 150$, $m = 798$ y $n = 26$ obtiene el mejor

caja		T_{obs} 75ms	T_{obs} 120ms	T_{obs} 150ms
		ROC_d	ROC_d	ROC_d
m=96	n=13	**84.2**	76.4	**76.4**
	n=26	76.4	76.4	68.5
	n=39	76.4	76.4	68.5
m=396	n=13	76.3	71.1	**90.2**
	n=26	82.5	**87.1**	86.8
	n=39	**87.3**	83.6	78.2
m=798	n=13	80.3	79.2	89.2
	n=26	89.9	**92.0**	**93.6**
	n=39	**90.8**	88.4	88.6

Tabla 4.2: Resultados de la distancia ROC en % de los modelos *HMM* de caja y silencio sobre los datos de test.

.

caja	*Test*	ROC_d
$m = 798, n = 30$	82.2	79.5
$m = 1596, n = 26$	88.9	88.1
$m = 1596, n = 30$	91.4	91.3

Tabla 4.3: Resultados de exactitud y distancia ROC en % de los modelos complementarios de caja y silencio.

rendimiento.

Es posible que el rendimiento de los modelos disminuya para $n = 39$ debido a un problema de *overfitting* de los datos de entrenamiento: El hecho de recoger muchas características ajusta demasiado los parámetros del modelo a los datos entrenados y no los generaliza bien en el *test set*. En ese sentido, se observa que el número de falsos negativos aumenta de $n = 26$ a $n = 39$ para el mismo *test set* y para tiempos de observación $T_{obs} = 120$ms y $T_{obs} = 150$ms. Esto se puede observar en la tabla 4.2, que muestra la distancia ROC de todos los modelos en el *test set*.

Se ha complementado el estudio realizando variaciones en el conjunto de datos y en el número de características para el modelo elegido, y comprobar si es posible mejorar el ratio de predicción. En concreto, se han creado modelos con un tamaño muestral $m = 1596$ y se ha aumentado el número de características en cuatro más, añadiendo las características espectrales detalladas en la sección 2.1.1.2. Los resultados obtenidos no logran superar el ratio de exactitud del modelo original, por lo que éste último será el elegido para la fase de evaluación continua (ver tabla 4.3).

bombo		T_{obs} 75ms		T_{obs} 120ms		T_{obs} 150ms	
		Train/val	*Test*	*Train/val*	*Test*	*Train/val*	*Test*
m=96	n=13	98.9	**100**	96.0	**100**	96.7	94.7
	n=26	98.5	**100**	97.0	94.7	99.4	94.7
	n=39	98.6	94.7	97.7	94.7	98.9	94.7
m=396	n=13	95.3	91.1	92.8	84.8	92.7	94.7
	n=26	95.4	**94.9**	93.5	93.6	96.6	98.7
	n=39	95.1	93.6	94.4	**94.9**	96.8	**100**
m=798	n=13	94.8	93.7	93.2	91.1	92.3	93.7
	n=26	96.5	**95.5**	95.0	97.4	95.8	96.2
	n=39	96.0	**95.5**	96.5	**98.1**	94.6	**98.7**

Tabla 4.4: Resultados de exactitud en % de los modelos *HMM* de bombo y silencio sobre los datos de entrenamiento (*Train/val*) y los datos de test (*Test*).

.

4.3.2.2. Bombo

Los modelos generados para el bombo muestran los mejores resultados de los tres instrumentos evaluados tanto en la fase de entrenamiento/validación como en la fase de test. Esto puede ser debido a que es el instrumento que presenta un sonido más grave con respecto al resto de instrumentos, a que su sonido no cambia demasiado a diferentes niveles de intensidad y tampoco es configurable, a diferencia de la caja tocada sin bordonera y el plato hi-hat con los platos abiertos o cerrados. La tabla 4.4 muestra los resultados. Al igual que ocurre con la caja, los resultados en la fase de entrenamiento/validación son mejores para un número de características $n = 26$ y $n = 39$ dado un mismo tiempo de observación y mismo conjunto de datos.

Los resultados de clasificación en el *test set* muestran que para un mismo tiempo de observación se observa una mejora sustancial a medida que se aumenta el conjunto de datos. Mención especial tienen los resultados obtenidos de 100 % de exactitud obtenidos para $T_{obs} = 75$ms y $m = 96$, debidos a la poca cantidad de muestras con la que se ha entrenado el modelo. Comparando con respecto a los tiempos de observación para un mismo tamaño muestral, los mejores resultados se obtienen para $T_{obs} = 150$.

Al contrario de lo que ocurre en los modelos de caja, el número de falsos negativos disminuye a medida que se aumenta el número de características, mientras el de falsos positivos permanece constante. Esto se refleja en una mejora de *sensibilidad* del modelo que, como se ha visto, afecta directamente sobre la distancia ROC, mostrada en la tabla 4.5. Esto ocurre principalmente en los modelos con $T_{obs} = 120$ y $T_{obs} = 150$.

bombo		T_{obs} 75ms	T_{obs} 120ms	T_{obs} 150ms
		ROC_d	ROC_d	ROC_d
m=96	n=13	**100**	**100**	**92.9**
	n=26	**100**	92.1	92.1
	n=39	92.9	92.1	92.1
m=396	n=13	87.6	80.4	92.9
	n=26	**92.9**	91.1	98.2
	n=39	91.1	**94.3**	**100**
m=798	n=13	93.6	90.8	91.9
	n=26	**95.5**	96.4	96.0
	n=39	95.2	**98.0**	**98.2**

Tabla 4.5: Resultados de la distancia ROC en % de los modelos *HMM* de bombo y silencio sobre los datos de test.

.

bombo	*Test*	ROC_d
$m = 798, n = 43$	96.8	96.3
$m = 1596, n = 39$	97.5	97.3
$m = 1596, n = 43$	95.4	94.0

Tabla 4.6: Resultados de exactitud y distancia ROC en % de los modelos complementarios de bombo y silencio.

Aunque se ha obtenido un 100 % de aciertos para $T_{obs} = 150$, $m = 396$ y $n = 39$ se ha optado también por incluir en la siguiente fase el mismo modelo entrenado con más muestras $T_{obs} = 150$, $m = 798$ y $n = 39$, por considerarse que pueda generalizar mejor durante la evaluación en continuo. Se presentará el resultado del modelo que mejores ratios de detección ofrezca.

Al igual que con la caja, se ha ampliado el estudio sobre modelos incrementando el tamaño muestral y el número de características a partir del modelo $T_{obs} = 150$, $m = 798$ y $n = 39$ (ver tabla 4.6). Aunque los resultados son muy buenos en la predicción del *test set*, no logran mejorar los ratios de predicción obtenidos previamente, por lo que se mantendrá el criterio de probar en tiempo real los dos modelos descritos anteriormente. El modelo $T_{obs} = 150$, $m = 798$ y $n = 43$ presenta un número de falsos positivos superior al modelo $T_{obs} = 150$, $m = 798$ y $n = 39$. Asimismo, se observa que el número de falsos positivos también aumenta de manera acusada del modelo $T_{obs} = 150$, $m = 1596$ y $n = 39$ al modelo $T_{obs} = 150$, $m = 1596$ y $n = 43$ lo que hace indicar que las cuatro características espectrales añadidas introduce información no relevante para la clasificación.

hi-hat		T_{obs} 75ms		T_{obs} 120ms		T_{obs} 150ms	
		Train/val	*Test*	*Train/val*	*Test*	*Train/val*	*Test*
m=96	n=13	94.9	**78.9**	90.5	78.9	87.0	63.1
	n=26	94.5	73.6	92.4	78.9	91.5	73.6
	n=39	92.5	**78.9**	92.4	78.9	91.2	**78.9**
m=396	n=13	86.7	81.0	82.2	73.4	84.5	88.4
	n=26	86.3	81.0	85.1	**89.8**	88.0	85.8
	n=39	86.3	**82.2**	84.0	88.6	89.4	**89.7**
m=798	n=13	85.8	86.1	81.8	86.7	81.8	85.3
	n=26	85.7	87.4	83.5	86.7	85.1	88.5
	n=39	84.9	**88.0**	84.3	**88.0**	84.8	**90.4**

Tabla 4.7: Resultados de exactitud en % de los modelos *HMM* de hi-hat y silencio sobre los datos de entrenamiento (*Train/val*) y los datos de test (*Test*).

.

4.3.2.3. Plato Hi-hat

Tal y como se explica en la sección 1.2.2, el plato hi-hat produce sonidos muy diferenciados en función de las distancia entre los platos que lo componen. Con el fin de simplificar el estudio, y comprobar el ratio de predicción de un único modelo *HMM*, se ha modelado el mismo para que sea capaz de detectar tanto el golpeo del hi-hat con los platos juntos ('chh' en la base de datos *ENST-Drums*) como abiertos ('ohh'). Los resultados se muestran en la tabla 4.7. Excepto para un tiempo de observación de $T_{obs} = 75$ms, los mejores resultados en los datos de entrenamiento/validación se obtienen generalmente para un número de características $n = 26$ y $n = 39$ dado un mismo tiempo de observación y mismo conjunto de datos.

Se ha observado que para un tiempo de observación de $T_{obs} = 120$ y $T_{obs} = 150$, el número de falsos positivos disminuye a medida que aumenta el número de características, mientras que el número de falsos negativos permanece constante. Esto se refleja en una mejora del ratio de falsos positivos, y tiene incidencia directa sobre la distancia ROC, mostrada en la tabla 4.8.

En los datos de test, el modelo $T_{obs} = 150$, $m = 798$ y $n = 39$ ofrece el mejor ratio de exactitud, aunque realmente las diferencias entre los modelos entrenados para un conjunto $m = 798$ son relativamente pequeñas.

Al igual que con la caja y el bombo, se ha ampliado el estudio con la creación de modelos incrementando el tamaño muestral y el número de características a partir del modelo $T_{obs} = 150$, $m = 798$ y $n = 39$. El hecho de aumentar el tamaño muestral si podría ayudar a la hora de generalizar mejor el modelo, ya que como se ha explicado, el espectro sonoro del plato hi-hat

hi-hat		T_{obs} 75ms	T_{obs} 120ms	T_{obs} 150ms
		ROC_d	ROC_d	ROC_d
m=96	n=13	**75.3**	**78.8**	63.1
	n=26	72.5	75.3	72.5
	n=39	**75.3**	75.3	**75.3**
m=396	n=13	80.7	73.4	87.8
	n=26	80.9	**89.8**	84.5
	n=39	**82.0**	88.5	**89.4**
m=798	n=13	85.7	86.1	85.2
	n=26	86.9	86.7	88.5
	n=39	**87.9**	**86.8**	**89.4**

Tabla 4.8: Resultados de la distancia ROC en % de los modelos *HMM* de hi-hat y silencio sobre los datos de test.

.

hi-hat	*Test*	ROC_d
$m = 798, n = 43$	87.8	87.8
$m = 1596, n = 39$	85.5	85.0
$m = 1596, n = 43$	83.5	83.2

Tabla 4.9: Resultados de exactitud y distancia ROC en % de los modelos complementarios de hi-hat y silencio.

produce características muy diferenciadas. Sin embargo, tal como se observa en la tabla 4.9 los resultados, aunque con ratios de acierto aceptables, no logran superar los obtenidos previamente.

4.3.3. Evaluación del modelo en tiempo real y con acompañamiento

Para evaluar los modelos finales seleccionados por instrumento, se utilizarán las grabaciones de *ENST-Drums* que incorporan acompañamiento musical a la sección rítmica. Esto permitirá evaluar los modelos de dos maneras distintas: sin acompañamiento, es decir, utilizando solo la pista de batería, y con acompañamiento, añadiendo el resto de instrumentos. Conviene comentar que las grabaciones utilizadas para la evaluación en continuo contienen mucha información musical. Se trata de secuencias con patrones de batería complejos y ritmos medios y altos, donde normalmente el número de instrumentos que actúan a la vez es múltiple. Es de esperar que el ratio de acierto de los modelos se vea deteriorado por el hecho de incorporar otros instrumentos que puedan "tapar" o camuflar tambores y platos objeti-

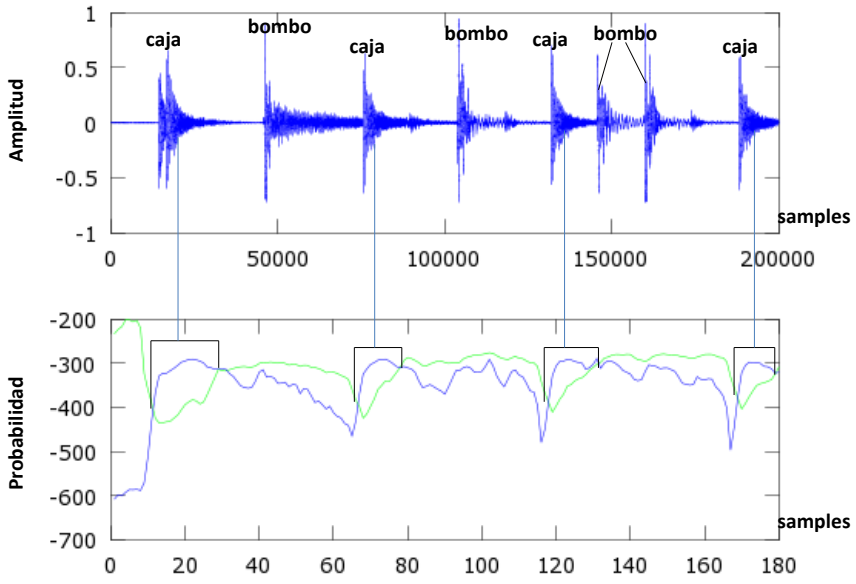

Figura 4.7: Ejemplo de reconocimiento mediante comparación de probabilidades. La gráfica superior muestra la señal de audio donde se han indicado los golpeos de caja. La gráfica inferior muestra la probabilidad de presencia de los modelos silencio(linea verde) y caja (linea azul).

vo. Para generar grabaciones con el acompañamiento, se utiliza el editor de audio *Audacity*, que permite solapar pistas y editar su volumen. Los modelos seleccionados finalmente han sido:

- Caja: modelo entrenado con $T_{obs} = 150$, $m = 798$, $n = 26$.
- Bombo: modelos entrenados con $T_{obs} = 150$, $m = 798$, $n = 39$ y $T_{obs} = 150$, $m = 396$, $n = 39$. Los resultados mostrados en las siguientes tablas pertenecen al modelo con $m = 798$, que obtiene los mejores ratios entre ambos.
- Hi-hat: modelo entrenado con $T_{obs} = 150$, $m = 798$, $n = 39$.

Durante el proceso de evaluación continua, cada archivo de audio es escaneado cada 25ms, que entra dentro de los márgenes de desviación utilizados en trabajos similares de detección en tiempo real. En cada escaneado, se extrae el tiempo de observación definido para el modelo elegido y se calculan las probabilidades de que esa secuencia pertenezca al modelo instrumento y silencio. Las áreas aisladas donde la probabilidad del modelo instrumento supere a la del modelo silencio se considera como golpeo del instrumento y

	Offline			Tiempo real		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Caja	93.6	93.6	93.6	74.7	32.8	45.6
Bombo	100	97.5	98.7	92.8	46.9	62.3
Hi-hat	94.3	85.8	89.9	58.2	27.1	37.0

Tabla 4.10: Comparativa de los resultados de la evaluación de los modelos en *off-line* (sobre el *test set*) y en tiempo real, ambos sin acompañamiento musical.

	P(%)	R(%)	F1(%)
Caja	59.8	27.5	37.7
Bombo	79.2	6.8	12.5
Hi-hat	36.0	4.6	8.1

Tabla 4.11: Resultados de la evaluación de los modelos en tiempo real con acompañamiento musical.

se marcará como tal (ver figura 4.7). Una vez escaneado todo el fichero, se comparan los resultados con los eventos reales y se contabilizan los errores (falsos negativos y falsos positivos) y aciertos (verdaderos positivos).

Por último, después de escanear todos los ficheros de audio se calcula la *precisión* P, el ratio de *exhaustividad* R y el Valor-F. La tabla 4.10 muestra los resultados sin acompañamiento musical. Se observa que los datos de *precisión* son aceptables, si se comparan con respecto a otros estudios publicados (ver sección 4.3.4), es decir, la relación de verdaderos positivos, o golpeos del instrumento en cuestión, con respecto al de los falsos positivos muestra que los golpeos que el sistema identifica tienen una alta probabilidad de ser verdaderos. Destacar en ese sentido una *precisión* del 92,8 % en la detección del sonido del bombo. Por el contrario, los resultados de *exhaustividad* del sistema no son tan buenos, principalmente porque este se deja muchos golpeos sin detectar, es decir, el número de falsos negativos con respecto al de verdaderos postivos es considerablemente mayor. Esto hace que la medida *F1*, basada en la media armonizada de ambos parámetros se reduzca considerablemente.

Los resultados finales de evaluación con respecto a esos ficheros de audio con acompañamiento musical generados a través del programa *Audacity* se muestran en la tabla 4.11. Para poder hacer una comparativa equitativa con otros estudios, la intensidad de la pista de batería supone 2/3 de la grabación final mientras que la pista de acompañamiento supone el tercio restante. En este caso, los problemas de detección de golpeos presentado en la evaluación

		P(%)	R(%)	F1(%)
(Paulus y Klapuri, 2009)	Caja	68.7	57.7	62.7
	Bombo	95.7	88.1	91.8
	Hi-hat	82.7	80.9	81.8
Baterista 3	Caja	65.7	43.6	52.4
	Bombo	96.8	70.3	81.4
	Hi-hat	57.7	36.3	44.6

Tabla 4.12: Comparativa de resultados obtenidos para el baterista 3 con el trabajo de Paulus, ambos sin acompañamiento musical.

de los sistemas sin acompañamiento se ven acentuados sobremanera. Sobre todo para el bombo y el plato hi-hat el sistema deja la mayoría de golpeos sin detectar. Como punto positivo, comentar que el ratio de *precisión* no ha bajado tan considerablemente, por lo que el sistema mantiene en parte su capacidad para acertar cuando detecta un posible golpeo.

4.3.4. Comparativa con otros estudios publicados

A continuación se presentan resultados obtenidos en anteriores trabajos de transcripción que han utilizado la misma base de datos *ENST-Drums*. El trabajo de Paulus y Kapluri (Paulus y Klapuri, 2009) de reconocimiento por *HMM* es el que guarda mayor relación con el presente estudio, aunque la forma de evaluarlo sea distinta, presentando los resultados sobre un baterista en particular y utilizando los otros dos para el entrenamiento de los modelos. La tabla 4.12 muestra una comparativa entre sus resultados y el mejor ratio por baterista obtenido en el presente trabajo, ambos sin acompañamiento musical. Aunque el ratio de *exhaustividad* se mantiene por debajo, la *precisión* del modelo para la caja y el bombo mantienen un ratio similar o superior.

El resto de trabajos utilizan técnicas del tipo 'separar y detectar' mediante *non-negative matrix factorization* y *Prior subspace analysis* (Paulus y Virtanen, 2005), tipo 'segmentar y clasificar' mediante máquinas de vector soporte (Tanghe et al., 2005), y tipo 'emparejar y adaptar' mediante la técnica de *Partially Fixed Non-Negative Matrix Factorization* (Wu y Lerch, 2015).

Comentar que los datos de *precisión* tanto con y sin acompañamiento se mantienen al mismo nivel, o incluso superan, las prestaciones de algunos de los trabajos. Es en el ratio de *exhaustividad* donde las prestaciones de los modelos caen considerablemente, sobre todo en el análisis con acompañamiento.

		P(%)	R(%)	F1(%)
(Paulus y Virtanen, 2005)	Caja	75.6	38.1	50.7
	Bombo	85.0	80.1	82.5
	Hi-hat	57.1	67.7	61.9
(Tanghe et al., 2005)	Caja	62.9	37.9	47.3
	Bombo	95.4	54.0	69.0
	Hi-hat	61.1	72.3	66.2
(Wu y Lerch, 2015)	Caja	82.5	45.3	58.5
	Bombo	88.6	93.8	91.1
	Hi-hat	91.8	70.5	79.7

Tabla 4.13: Resultados de otros estudios sin acompañamiento musical.

		P(%)	R(%)	F1(%)
(Paulus y Virtanen, 2005)	Caja	57.0	16.7	25.9
	Bombo	69.9	57.9	63.4
	Hi-hat	58.2	53.5	55.8
(Tanghe et al., 2005)	Caja	65.9	14.2	23.4
	Bombo	80.9	38.4	51.1
	Hi-hat	47.1	69.5	56.1
(Wu y Lerch, 2015)	Caja	68.4	46.4	55.2
	Bombo	71.4	86.2	78.1
	Hi-hat	90.2	70.6	79.2

Tabla 4.14: Resultados de otros estudios con acompañamiento musical.

4.4. Conclusiones y lineas futuras

Se ha presentado un estudio sobre reconocimiento de instrumentos de percusión mediante modelos ocultos de Markov. Se ha utilizado una técnica por clases de detección de instrumento, donde cada modelo se utiliza para detectar la presencia o ausencia de un único tambor o plato. También se ha definido previamente la topología de red a utilizar y el número de estados de los que consta cada modelo, siendo dos para el silencio y cuatro para el instrumento. Los instrumentos sometidos a estudio han sido la caja, el bombo y el plato hi-hat. Se han probado diferentes tiempos de observación, conjuntos de datos y número de características para evaluar la predicción de los modelos y compararlos entre si. Se ha concluido que los modelos que han sido entrenados con tiempos de observación mayores a $100ms$ y un número de características $n = 26$ y $n = 39$ obtienen mejores resultados de clasificación *off-line* que el resto.

El sistema en continuo ofrece un ratio de acierto elevado cuando detecta

un instrumento, principalmente la caja y el bombo. Los ratios de *precisión P* de los modelos son alentadores de cara a probar nuevas configuraciones para la misma topología de *HMM* utilizada. Así, aunque se hayan elegido cuatro estados para describir la secuencia espectral emitida por los instrumentos, es posible que otro valor pueda mejorar los resultados. Otra consideración sería modelar *HMM* independientes para el plato hi-hat tocado cerrado y abierto, debido a las grandes diferencias acústicas que producen entre si y que un solo modelo no pueda asumir de forma efectiva.

Como principal inconveniente, la capacidad total de detección del sistema es limitada, considerando una gran cantidad de eventos como secuencias de silencio. Es precisamente el criterio de consideración de silencio el que dificulta la correcta clasificación de la señal. En este caso, el silencio es considerado la ausencia de cualquiera de los instrumentos sometidos a estudio, incluyendo también el remanente de sonido generado instantes previos por un instrumento objetivo, con la variabilidad y dificultad de modelado que ello conlleva. Cualquier otro sonido en la grabación, del tipo que sea, debe ser considerado como silencio, por lo que su caracterización debe ser amplia. Se ha probado que una red *HMM* de dos estados para modelar el silencio abarca demasiada información e 'invade' las características espectrales del tambor o plato que se quiere detectar, considerándolo como silencio un número elevado de veces.

Una posible alternativa sería considerar la matriz de covarianzas Σ que define las relaciones entre las características de las observaciones como una matriz completa, en lugar de la matriz diagonal que se ha utilizado para realizar los experimentos. De esta forma se establecerían correlaciones entre las características recogidas, pudiendo ajustar mejor la gaussiana que define la distribución de las observaciones. Un inconveniente a esta alternativa sería el mayor coste computacional que ello conlleva.

Otra opción sería reemplazar la distribución multivariante gaussiana por una mezcla de gaussianas (*Gaussian Mixture Model, o GMM*) de m componentes para modelar la distribución estado-observación:

$$b_j(O_t) = \sum_{m=1}^{M} c_{jm} \mathcal{N}[O; \mu_j, \Sigma_j]$$

Siendo c_{jm} la probabilidad a priori de la componente m (una función de densidad de distribución gaussiana) del estado q_j, que debe cumplir las siguientes restricciones:

$$\sum_{m=1}^{M} c_{jm} = 1 \quad y \quad c_{jm} \geq 0$$

Usar *GMM* incrementa el coste computacional, además de requerir nuevos cálculos, como determinar el número de componentes por estado en el

modelo *HMM*. Este paso se puede simplificar considerando el mismo número para todos ellos.

En definitiva, pese a ser la primera aproximación a la problemática de la transcripción musical en instrumentos de percusión con una técnica tan poco trabajada en este ámbito como los *HMM*, los resultados obtenidos *off-line* son muy buenos. La evaluación del sistema en continuo sin acompañamiento arroja resultados buenos, comparables a otras aproximaciones. Sin embargo, la evaluación en tiempo real con acompañamiento deja ver que hay margen de mejora, y que hay ideas para continuar sobre ello. Aunque los resultados hayan mostrado deficiencias en la capacidad de detección de los instrumentos en condiciones adversas de solapamiento con otros instrumentos o de acompañamiento musical, las alternativas disponibles para aumentar la complejidad de los modelos *HMM* hacen pensar que una exploración más profunda pueda mejorar los resultados obtenidos.

Bibliografía

Y así, del mucho leer y del poco dormir,
se le secó el celebro de manera que vino
a perder el juicio.

Miguel de Cervantes Saavedra

Chroma, Matlab tool box. Disponible en `http://resources.mpi-inf.mpg.de/MIR/chromatoolbox/`.

Denemo, notation software. Disponible en `http://www.denemo.org/`.

Drumagog, drum replacer plug-in. Disponible en `http://www.drumagog.com/`.

ENST-Drums, drum database. Disponible en `http://perso.telecom-paristech.fr/~grichard/ENST-drums/`.

Essentia, software library. Disponible en `http://essentia.upf.edu/`.

Good Sounds, software. Disponible en `https://good-sounds.org/`.

Guitar Hero, video game. Disponible en `https://www.guitarhero.com/es/`.

H2M, EM estimation of Hidden Markov Models. Disponible en `http://perso.telecom-paristech.fr/~cappe/Code/H2m/`.

HTK Speech Recognition Toolkit. Disponible en `htk.eng.cam.ac.uk/`.

ISMIR, the international society of music information retrieval. Disponible en `http://www.ismir.net/`.

jMIR, software. Disponible en `https://github.com/DDMAL/jMIR`.

Ludwig, software. Disponible en `http://www.write-music.com/`.

MIREX, the Music Information Retrieval Evaluation eXchange. Disponible en `http://www.music-ir.org/?q=node/13`.

MSAF, software library. Disponible en `https://github.com/urinieto/msaf`.

Musescore, notation software. Disponible en `https://musescore.org/es`.

Musicpedia, music searcher engine. Disponible en `http://musicpedia.org/`.

Pandora, software. Disponible en `http://www.pandora.com/mir/`.

Peachnote, music searcher engine. Disponible en `http://www.peachnote.com/`.

Rastamat, MFCC extraction. Disponible en `http://labrosa.ee.columbia.edu/matlab/rastamat/`.

Rock Band 4, video game. Disponible en `http://www.rockband4.com/`.

ScoreCloud, automatic notation software. Disponible en `http://scorecloud.com/`.

Songs2See, music learning video game. Disponible en `http://songs2see.com/`.

Soundsnap, sound library. Disponible en `http://www.soundsnap.com/`.

The Beamz, therapy and rehab interactive music tools. Disponible en `http://www.thebeamz.com/`.

VirtualBand. Disponible en `http://medias.ircam.fr/xbec879_virtualband-a-mir-approach-to-interactive/`.

ALVES, D., PAULUS, J. y FONSECA, J. Drum transcription from multichannel recordings with non-negative matrix factorization. *Proceedings of 17th European Signal Processing Conference*, 2009.

BAUM, L. E. y PETRIE, T. Statistical inference for probabilistic functions of finite state markov chains. *Annals of Mathematics Statistics 37 (1966), no. 6, 1554–1563*, 1966.

BAUM, L. E., PETRIE, T., SOULES, G. y WEISS, N. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematics Statistics 41 (1970), no. 1, 164-171*, 1970.

BELLO, J. P. Drum sound analysis for the manipulation of rhythm in drum loops. *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, 2006.

BELLO, J. P., DAUDET, L., ABDALLAH, S. A., DUXBURY, C., DAVIES, M. y SANDLER, M. B. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 2005.

BHAR, R. y HAMORI, S. *Hidden Markov Models: Applications to Financial Economics*. Springer, 2004.

BHARATH, A. y MADHVANATH, S. Hidden markov models for online handwritten tamil word recognition. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2007.

BILMES, J. A. *Timing is of the essence: Perceptual and computational techniques for representing, learning and reproducting expressive timing in percussive rhythm*. Proyecto Fin de Carrera, Massachusetts Institute of Technology, 1993.

CHAFE, C., KASHIMA, K., MONT-REYNAUD, B. y SMITH, J. Techniques for note identification in polyphonic music. *Proceedings of the International Computer Music Conference*, 1985.

DEMPSTER, A. P., LAIRD, N. M. y RUBIN, D. B. Maximun likelihood for incomplete data via the em algorithm. *J. Royal Statistical Society, 39 (Series B), 1-38*, 1977.

DITTMAR, C. y GARTNER, D. Drumloop separation using adaptive spectrogram templates. *Proceedings of the 36th Jahrestagung fuer Akustik (DAGA)*, 2010.

DITTMAR, C. y GARTNER, D. Real-time transcription and separation of drum recordings based on nmf decomposition. *Proceedings of 17th International Conference on Digital Audio Effects (DAFx)*, 2014.

DIXON, S. Automatic extraction of tempo and beat from expressive permformances. *Journal of New Music Research*, 2001.

EMIYA, V., BADEAU, R. y BAVID, B. Automatic transcription on piano music based on hmm tracking of jointly-estimated pitches. *Signal Processing Conference, 2008 16th European*, 2008.

FITZGERALD, D. *Automatic drum transcription and source separation*. Tesis Doctoral, Dublin Institute of Technology, 2004.

FITZGERALD, D., LAWLOR, B. y COYLE, E. Drum transcription in the presence of pitched instruments using prior subspace analysis. *Proceedings of Irish Signals and Systems Conference*, 2003.

FITZGERALD, D. y PAULUS, J. *Unpitched percussion transcription*. Springer, 2006.

GILLET, O. y RICHARD, G. Automatic transcription of drum loops. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2004.

GILLET, O. y RICHARD, G. Automatic transcription of drum sequences using audiovisual features. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2005.

GILLET, O. y RICHARD, G. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

GOTO, M. y MURAOKA, Y. A beat tracking system for acoustic signals of music. *Proceedings of ACM Multimedia*, 1994a.

GOTO, M. y MURAOKA, Y. A sound source separation system for percussion instruments. *The Transactions of the Institute of Electronics, Information and Communication Engineers*, 1994b.

GOUYON, F. y HERRERA, P. Exploration of techniques for automatic labeling of audio drum tracks instruments. *Proceedings of MOSART Workshop on Current Research Directions in Computer Music*, 2001.

GOUYON, F., PACHET, F. y DELERUE, O. On the use of zero-crossing rate for an application of classfication of percussive sounds. *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*, 2000.

HELEN, M. y VIRTANEN, T. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machines. *Proceedings of 13th European Signal Processing Conference*, 2005.

HERRERA, P., YETERIAN, A. y GOUYON, F. Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. *Proceedings of 2nd Iternational Conference on Music and Artificial Intelligence*, 2002.

KARLOF, C. y WAGNER, D. *Hidden Markov model cryptanalysis*. Springer, 2003.

KLAPURI, A. Sound onset detection by applying psychoacoustic knowledge. *Proceedings of IEEE International conference on Acoustics, Speech and Signal Processing*, 1999.

KLAPURI, A. *Signal processing methods for the automatic transcription of music*. Tesis Doctoral, Tampere University of Technology, 2004.

KROGH, A., BROWN, M., MIAN, I. S., SJOLANDER, K. y HAUSSLER, D. Hidden markov models in computational biology: applications to protein modelling. *Journal of Molecular biology, 235, 1501-1531*, 1994.

KUNDU, A., HE, Y. y BAHL, P. Recognition of handwritten word: first and second order hidden markov model based approach. *Computer Vision and Pattern Recognition, 1988. Proceedings CVPR '88, Computer Society*, 1988.

KUPIEC, J. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language, Volume 6, Issue 3, 225-242*, 1992.

LOGAN, B. Mel frequency cepstral coefficients for music modeling. *Proceedings of 1st International Symposium on Music Information Retrieval*, 2000.

MAHER, R. C. *An approach for the separation of voices in composite music signals*. Tesis Doctoral, University of Illinois, 1989.

MOORER, J. A. *On the segmentation and analysis of continuous musical sound by digital computer*. Tesis Doctoral, Stanford University, 1977.

OCH, F. J. y NEY, H. A systematic comparison of various statistical alignment model. *Computational Linguistics, 30, 417-449*, 2003.

PAPANIKAS, G. *Real-time automatic transcription of drums music tracks on an FPGA platform*. Proyecto Fin de Carrera, Technical University of Denmark, 2012.

PAULUS, J. *Signal processing methods for drum transcription and music structure analysys*. Tesis Doctoral, Tampere University of Technology, 2009.

PAULUS, J. y KLAPURI, A. Model-based event labeling in the transcription of percussive audio signals. *Proceedings of 6th International Conference on Digital Audio Effects (DAFx)*, 2003.

PAULUS, J. y KLAPURI, A. Drum sound detection in polyphonic music with hidden markov models. *EURASIP Journal on Audio, Speech and Music Processing*, 2009.

PAULUS, J. y VIRTANEN, T. Drum transcription with non-negative spectrogram factorisation. *Proceedings of 13th European Signal Processing Conference*, 2005.

PISZCZALSKI, M. B. *A computational model of music transcription*. Tesis Doctoral, University of Michigan, 1986.

PLUMBEY, M. D. y ABDALLAH, S. A. Automatic music transcription and audio source separation. *Cybernetics and Systems 33*, 2002.

RABINER, L. B. y JUANG, B. H. *Fundamentals of speech recognition*. Prentice Hall, 1993.

RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE, pages 257-286*, 1989.

RICARD, J. An implementation of multi-band onset detection. *First Annual Music Information Retrieval Evaluation eXchange. MIREX*, 2005.

ROWE, R. *Interactive music systems*. MIT Press, 1993.

RUSSELL, S. y NORVIG, P. *Artificial Intelligence: A modern Approach*. Pearson New International Edition, 2014.

SANDVOLD, V., GOUYON, F. y HERRERA, P. Percussion classification in polyphonic audio recordings using localized sound models. *Proceedings of 5th International Conference on Music Information Retrieval*, 2004.

SCHLOSS, W. A. *On the automatic transcription of percussive music - from acoustic signal to high-level analysis*. Tesis Doctoral, Stanford University, 1985.

SCOREWRITER. Scorewriter — Wikipedia, the free encyclopedia. Disponible en `https://en.wikipedia.org/wiki/Scorewriter`. [Online; accedido por ultima vez 25-Marzo-2016].

SHAO, X., XU, C. y KANKANHALLI, M. S. Unsupervised classification of music genre using hidden markov model. *Proceedings of IEEE International Conference on Multimedia and Expo*, 2004.

SILLANPAA, J., KLAPURI, A., SEPPANEN, J. y VIRTANEN, T. Recognition of acoustic noise mixtures by combined bottom-up and top-down processing. *Proceedings of European Signal Processing Conference*, 2000.

SIMSEKLI, U., JYLHA, A., ERKUT, C. y CEMGIL, A. Real-time recognition of percussive sounds by a model-based method. *EURASIP Journal on Advances in Signal Processing*, 2010.

SPICH, A., ZANONI, M., SARTI, A. y TUBARO, S. Drum music transcription using prior subspace analysis and pattern recognition. *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)*, 2010.

VAN STEELANT, D., TANGHE, K., DEGROEVE, S., BAETS, B. D., LEMAN, M. y MARTENS, J. P. Classification of percussive sounds using support vector machines. *Proceedings of The Annual Machine Learning Conference of Belgium and The Netherlands*, 2004.

TANGHE, K., DEGROEVE, S., y BAETS, B. D. An algorithm for detecting and labelling drum events in polyphonic music. *Proceedings of First Annual Music Information Retrieval Evaluation eXchange*, 2005.

THE DRAGON SYSTEM, A. O. Maximun likelihood for incomplete data via the em algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing 23: 24-29*, 1975.

VITERBI, A. J. Error codes for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory, 13(2), 260-269*, 1967.

WILSON, A. D. y BOBICK, A. F. Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999.

WU, C.-W. y LERCH, A. Drum transcription using partially fixed non-negative matrix factorization with template adaptation. *Proceedings of the 16th International Conference on Music Information Retrieval (ISMIR)*, 2015.

YOSHII, K., GOTO, M. y OKUNO, H. G. Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression. *IEEE Transactions on Audio, Speech and Language Processing*, 2007.