PROGRAMA DE POSTGRADO
EN SISTEMAS Y REDES DE COMUNICACIONES

EXTRACTOR DE INFORMACIÓN RÍTMICA DE SEÑALES MUSICALES

Lino García Morales

Escuela Técnica Superior de Ingenieros de Telecomunicación
Universidad Politécnica de Madrid

Extractor de información rítmica de señales musicales

por

Lino García Morales

Se propone un sistema para el análisis y seguimiento del ritmo directamente de una señal acústica musical. La inducción del ritmo es la base para la sincronización de diversas aplicaciones multimedia como la edición de vídeo, audio y el control de iluminación.

El sistema segmenta el material acústico en eventos musicales mediante la detección de las transiciones bruscas de energía (discriminación de los componentes de mayor rapidez de incremento de potencia) respecto a una vecindad potencial en el plano tiempo-frecuencia y genera una lista de eventos con la información de los tiempos de ataque y la contribución energética relativa por componente.

La inducción del tempo estima, con cierta certidumbre, la próxima posición temporal correspondiente a un acento rítmico importante. El modelo de percepción del ritmo se basa en la descomposición del contexto o patrón temporal en curvas de experanza básica por cada intervalo implícito y su proyección hacia el futuro. Se propone una formalización matemática del modelo desde el paradigma teórico psicoperceptual. La esperanza resultante del patrón temporal complejo generado por la segmentación sirve para modelar tópicos tan diversos como la percepción categórica del ritmo, la inducción del reloj y el metro, la ritmicidad y la similitud de secuencias temporales.

La inducción del tempo directamente del material acústico permite modelar la expresividad temporal e inflexiones del tempo en favor del proceso de cuantización.

Agradecimientos

Son muchas las personas e instituciones sin la cual éste trabajo no hubiera sido posible:

Al profesor Javier Casajús, tutor de ésta tesis.

A la Agencia Española de Cooperación Internacional.

A la Escuela Técnica Superior de Ingenieros de Telecomunicaciones de la Universidad Politécnica de Madrid.

A FUNDESCO.

Al Instituto Superior de Arte de La Habana.

A Jose Antonio Alonso, Manuel Iglesias y Cipri Martín.

A mis amigos de Agustín Parejo.

A mi familia.

Índice

A Anita, Cipri y Jose

La música, ..., no es más que una palabra

John Cage

Capítulo 1 Introducción

Cuando alguien dice “tengo el tempo” probablemente se refiera a que puede seguir el ritmo con el pie o mover el cuerpo, o emprender cualquier tipo de acción al unísono con tal secuencia musical. Es decir, se da una sincronización que nos permite, en base a la experiencia acumulada, adelantarnos a los acontecimientos y predecir con cierta exactitud cuando ocurrirá esa próxima marca oculta que nos permite movernos al compás de la música. Este proceso de inducción del *pulso* o *acento* consiste en detectar la regularidad de algún énfasis espaciado en la sucesión de los sonidos y silencios. Esta tarea cognitiva relativamente simple es parte de un complejo sistema de procesamiento de información sólo parcialmente conocido donde, además del aparato auditivo, interviene el sistema nervioso. Normalmente el *ritmo* de una interpretación musical nos resulta obvio apenas a los pocos segundos de escucha pero esto no significa que escribir un programa computacional para detectarlo sea lo mismo de sencillo. A menudo olvidamos que la maquinaria mental que detecta el ritmo ha sido el producto de una larga evolución y es aparentemente harto sofisticada [Rosenthal,1992]. La emulación de esta sofisticación sobre una maquina no es una tarea nada trivial. Al proceso de inducción del pulso nos acercaremos en dos pasos jerárquicamente concatenados:

La *segmentación*[1] de la secuencia musical en eventos
La *inducción del pulso* a partir de los intervalos de tiempo entre eventos

Grande ha sido el esfuerzo de muchos investigadores en la búsqueda de un método de segmentación fiable y generalizable. Sin embargo la complejidad intrínseca al material musical sólo ha permitido soluciones satisfactorias de compromiso. Muchos de los investigadores han atacado el proceso de segmentación en el dominio del tiempo, asociándolo frecuentemente a

1 Un método de segmentación es aquel que permite descomponer la secuencia musical en elementos que correspondan a eventos musicales (o en el caso de la voz, fonéticos)

variaciones de la amplitud, pendiente del ataque, incremento brusco de la energía, etc., otros menos en el dominio de la frecuencia, soporte para la investigación melódica, armónica, etc.

Los *eventos* musicales[2] tienen asociado una serie de parámetros físicos del sonido como son duración, intensidad, altura, timbre, etc., cuyas permutaciones dan lugar al universo infinito de lo que el hombre conoce como música. La frecuencia de ciertas combinaciones de estos rasgos a diferentes niveles dan lugar a clasificaciones de estilos, análisis semiótico, etc., y es el soporte de un impresionante despliegue de esfuerzos de análisis que intenta desde los más diversos puntos de vista la comprensión del fenómeno de la creación musical ya sea para su sistematización cognoscitiva como para su incorporación en las nuevas tecnologías.

El tema de la localización y seguimiento del pulso ha generado un sin número de trabajos y artículos desde un amplio espectro de formalismos computacionales: sistemas basados en reglas, métodos de optimización, búsqueda, teoría de control, sistemas distribuidos, redes neuronales, modelos estadísticos, etc.

Muchas son las aplicaciones tecnológicas que encuentra el seguimiento o localización del pulso y pueden ser subdivididas en dos grandes grupos: los sistemas de edición y los sistemas de interpretación-en vivo [Rosenthal,1992].

Los sistemas de edición incluyen los editores de partituras en computadoras, los sistemas de edición de vídeo utilizados para sincronizar pistas de audio musical con pistas de vídeo y los sistemas de producción de grabaciones con diversas pistas de audio.

La desviación deliberada de la metricalidad, como el rubato, es utilizada para enfatizar la estructura musical. Añadidos a estos efectos, colectivamente conocidos como *expresividad temporal*, existen otra serie de efectos involuntarios, como errores temporales aleatorios provocados por los límites en la precisión del sistema motor [Shafer,1981] y errores en el proceso mental de mantenimiento del tiempo [Vorberg y Hambuch,1978]. Estos efectos son muy pequeños, del orden de 10-100 mseg. Para apreciar la mayoría de los estilos musicales, es necesario separar las componentes de las dos escalas del

2 El evento se corresponde con la idea de lo que acostumbramos llamar nota musical.

tiempo musical: continua y discreta[3]. Este proceso se conoce con el nombre de *cuantización*, aunque el término es generalmente empleado sólo para reflejar la extracción del pentagrama métrico de la secuencia musical[4]. La percepción de los intervalos de tiempo sobre la escala de tiempo discreta es también denominada *percepción categórica*. Las curvas de expectación complejas generadas por el método de inducción del pulso expuesto se pueden utilizar para modelar la percepción categórica del ritmo y en consecuencia para el proceso de cuantización.

La cuantización tiene aplicaciones técnicas tales como la categorización en la transcripción automática de la música directa, la composición en tiempo real y en componentes interactivos donde la computadora improvisa o interactúa con un ejecutante en vivo[5].

La posibilidad de análisis inteligente del ritmo en los editores de partituras permitiría a los compositores la introducción y experimentación directa de sus ideas musicales al pentagrama a través de un instrumento MIDI en lugar del tiempo que emplearían en escribirlas.

En los sistemas de edición de vídeo la localización y seguimiento del pulso facilitaría la sincronización de las pistas visuales con las pistas de audio. Esto beneficia la edición de videos promocionales donde normalmente los movimientos visuales están en sincronía con los pulsos musicales. Con un sistema de seguimiento del pulso es fácil crear gráficos por computadora en tiempo real sincronizados con la música. Por ejemplo un bailarín virtual sobre la pantalla gráfica de la computadora puede bailar al compás de la música. Los movimientos, pasos y la posición del bailarín pueden cambiar en función de los pulsos. Esto facilitaría de manera notable la producción de los dibujos animados por computadora.

El proceso de producción de grabaciones de audio podría ser sensiblemente afectado con la disponibilidad de los sistemas de seguimiento del ritmo.

---

3 Los intervalos de tiempo discreto de la estructura métrica y las escalas de tiempo continuo de los cambios de tempo y expresividad temporal.

4 La notación pentagramada simbólica establece duraciones fijas de tiempo entre las distintas figuras métricas en función del tempo. Es el ejecutante en su interpretación de la obra el encargado de impregnarle *expresividad* o inflexiones al tiempo. El proceso inverso o *nivel de interpretación* consiste en la asignación de niveles rítmicos a la secuencia musical ejecutada.

5 En algunas de estas aplicaciones la computadora actúa como un interprete autónomo, en otras como un instrumento musical inteligente.

Actualmente, cada pista de audio debe estar sincronizada con una pista maestro (denominada en la música popular “pista clic”) o de sincronía. Con la posibilidad de comprensión del ritmo, simplemente habría que decirle al sistema que sincronice dos pistas de audio grabadas independientemente, permitiendo una potente flexibilidad a los ingenieros de grabación, intérpretes y productores. En los sistemas de edición de audio o los sistemas de grabación digital directa a disco duro la localización de los pulsos permite además el ordenamiento indexado del material musical. Los usuarios de estos sistemas pueden tratar las señales acústicas como un conjunto de pulsos[6] en lugar de tratar directamente con las formas de ondas acústicas en bruto.

La posibilidad de comprensión del ritmo por las computadoras debe añadir una dimensión completamente nueva en la participación inteligente de las computadoras en las interpretaciones en vivo. La mayoría de las computadoras actualmente participan en las interpretaciones en directo como una grabadora de cinta inteligente, produciendo secuencias establecidas de eventos musicales en respuesta a sugestiones establecidas. Los músicos en un conjunto, sin embargo, están inmersos en una intensa comunicación duración la interpretación, y están continuamente ajustando y corrigiendo su sentido de convenio sobre donde ocurrirán los acentos y otros eventos importantes. Parte de esta comunicación es visual, como cuando los ejecutantes de una orquesta siguen al director, o cuando los integrantes de un conjunto de cámara se comunican por sutiles gestos, pero parte de esta comunicación es también aural. Las computadoras ahora mismo están incapacitadas de cualquier comunicación aural que demande la comprensión de los pulsos.

Otro tipo de aplicación de los sistemas de seguimiento del pulso en las intervenciones en vivo es el control de la iluminación en sincronía con el material musical. Por ejemplo, la variación de las diferentes propiedades de la luz tales como el color, intensidad, dirección y efectos en sincronía con la música.

Existen diferencias en los distintos sistemas en cuanto al tipo de demanda respecto a la localización y seguimiento del pulso. Para un sistema de ejecución es crítico que el localizador del ritmo produzca una respuesta rápida y asincrónica; por supuesto, tales sistemas no pueden esperar hasta que la

6 O a otros niveles de abstracción musical.

pieza termine antes de decidir su ritmo. Para otros, el sistema de ejecución puede requerir una representación musical menos completa que un sistema de trascripción, y producir una representación sólo de los niveles rítmicos más importantes ignorando el resto. Estos niveles de demandas pueden establecer etapas en el acercamiento al proceso de comprensión automática del ritmo[7].

Los sistemas de localización y seguimiento del pulso tienen también aplicaciones teóricas de incalculable valor, pues constituyen diferentes estadios hacia el desarrollo de un modelo genérico de percepción del ritmo. Para la construcción de un modelo es necesario descomponer el problema en un conjunto de pequeños problemas que impulsan el desarrollo de nuevas tecnologías y nuevas formas de pensamiento. Procesos tales como la generación de eventos fantasmas puede que no tengan contraparte en el cerebro. Sin embargo el cerebro, de alguna manera, tiene que resolver problemas como éste.

"El camino hacia un modelo creíble de percepción del ritmo puede parecer desalentadoramente largo e incierto. Sin embargo existen razones para esperar que a lo largo del camino, resolveremos más problemas que los planteados originalmente. Y esto es porque existen otros tipos de percepción que son notablemente similares a la percepción del ritmo. La percepción del ritmo es la mejor idea de una parte del conjunto de herramientas que tenemos los humanos (y, en diferente grado en la mayoría de los animales) para apreciar ciertos tipos de regularidades en el mundo. No todas estas herramientas son aurales; de hecho, algunos de nuestros más importantes *detectores rítmicos* son visuales" [Rosenthal,1992].

7 Aunque existen aplicaciones en tiempo real, son aún muy limitadas y requieren de un despliegue tecnológico impresionante. El sistema de seguimiento del pulso para señales acústicas musicales de Goto y Muraoka [Goto y Muraoka,1994] por ejemplo utiliza una computadora paralelo, la Fujitsu AP1000 de 64 celdas o unidades de procesamiento elemental. Cada celda consiste de una SPARC con FPU a 25 MHz, 16 Mbytes DRAM y 128 Kbytes de memoria *cache* mapa-directo.

Capítulo 2 El tempo

Introducción

Independientemente de la definición que consideremos más apropiada para el ritmo, éste parece estar inevitablemente ligado a algún tipo de percepción precisa de los eventos en el tiempo [Schloss,1985]. Povel y Essens teorizaron acerca de la ritmicidad como el resultado de sintonizar un reloj interno con la llegada de los eventos musicales[1] [Povel,1984] y su modelo, implementado en un programa de computadora, calcula el reloj mejor inducido para los datos de entrada [Povel y Essens,1985]. La inducción del pulso[2] es un proceso rápido[3] que ocurre apenas iniciada la escucha. Una vez inducido el pulso, se establece una estructura mental persistente que orienta la percepción del próximo material que llega [Desain,1994]. Este proceso facilita la percepción de la síncopa[4]. Sin embargo, la percepción del pulso debe ser capaz de adaptarse continuamente con suficiente sensibilidad a los cambios naturales del *tempo* de la música real. Esta dualidad, donde el modelo necesita inferir un pulso desde marcas imprecisas y a la vez permitir a partir de la guía perceptual que establece el pulso inducido organizar el material que llega, es muy difícil de modelar [Desain,1994]. Existen muchas teorías incompatibles acerca de la percepción temporal y la memoria, que explican bien un conjunto de fenómenos pero fallan al predecir otros.

1 Este proceso forma parte del proceso normal de escucha de la música en el cual el oyente ajusta continuamente su reloj interno (trama métrica) a las irregularidades temporales locales y a las variaciones de tiempo.

2 *Pulso* es la palabra utilizada con más frecuencia como equivalente castellano de la palabra inglesa *beat*, quizá porque su significado induce la idea de periodicidad (pulso cardíaco, etc.). aunque en determinado contexto la palabra *acento* quizá resulte más apropiada. En cualquier caso serán utilizadas ambas palabras para significar lo mismo.

3 Bastan sólo unas pocas notas para sentir una fuerte sensación de inducción del pulso.

4 “Escuchar” un *acento* donde hay silencio. La posición de los pulsos no tiene porque corresponder directamente con un sonido real, ni algún sonido específico tiene necesariamente que indicar directamente la posición de los pulsos o acentos.

Sea cual fuera el punto de partida en el propósito de comprensión del ritmo, la granularidad o precisión requerida está determinada por los niveles perceptuales de respuesta humanos investigados a partir de experimentos psicoacústicos.

Discriminación temporal

Muchos investigadores han elaborado sus hipótesis de discriminación temporal a partir de tres paradigmas fundamentales: *duración*, *intermitencia* y *regularidad*. La discriminación de la duración se basa en la estimación de un único intervalo de tiempo. La discriminación de la intermitencia se refiere a la estimación de la frecuencia del pulso. La discriminación de la regularidad evalúa la uniformidad de repetición de los pulsos.

En el contexto musical, donde nos interesa evaluar con precisión la regularidad del patrón de acentos o ritmo, estas pruebas son de gran interés. No menos importante resulta el límite de la discriminación temporal que determina la distancia temporal mínima o crítica que somos capaces de percibir entre dos eventos consecutivos muy próximos.

Los estudios psicoperceptuales demuestran que la discriminación temporal cumple con la ley de Weber[5] sólo en determinados intervalos o rangos de tiempo y fuertemente dependiente de la naturaleza del estímulo con que se mide. En general, aunque los rangos difieren entre los distintos investigadores, la ley de Weber falla para duraciones muy cortas y muy largas. En el rango de las notas musicales típicas[6] sólo se cumple alguna versión modificada de la ley de Weber. Lunney plantea que los límites de la discriminabilidad son impuestos biológicamente [Lunney,1974].

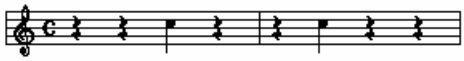
Ira Hirsh [Hirsh,1959] detectó, utilizando estímulos sintéticos, que es posible separar perceptualmente dos sonidos separados entre sí 2 mseg. Sin

5 La ley de Weber es una ley psicofísica general que plantea que, la discriminabilidad perceptual de un sujeto con respecto a un atributo físico, es proporcional a su magnitud, de manera tal que, $k = \frac{\Delta x}{x}$, donde x es el atributo medido y Δx es el cambio perceptual mínimo detectado. La magnitud k es denominada razón de Weber y es adimensional.

6 El “rango musical” se corresponde al intervalo de 100 mseg a 2 seg aproximadamente. La ley de Weber no se comporta igual en todos los modelos. Muchos investigadores reportan ratios de Weber ligeramente generalizados, otros, ratios de versiones de la ley de Weber e incluso ratios de Weber para distintos subintervalos del rango musical.

embargo, Patterson y Green reportan, para el mismo experimento, discriminación temporal de intervalos más pequeños [Patterson y Green,1970]. La diferencia entre ambas pruebas fue sólo la duración de los estímulos, que en el caso de Patterson y Green eran más cortos. El hecho demuestra que el límite de la discriminación temporal entre eventos depende en gran medida de la duración de los estímulos utilizados.

Se ha comprobado que las personas tenemos la mayor agudeza temporal en el intervalo de 500 a 800 mseg, lo que corresponde para ♩ = 60 a valores de notas desde 𝅘𝅥𝅯 hasta 𝅗𝅥. . En el rango musical normal parece suficiente una discriminabilidad temporal de 5 mseg para la determinación de los tiempos de ataque sin perder la información de tiempo esencial.

Para precisar acerca del límite de discriminabilidad temporal, se diseño un experimento con dos tipos de estímulos diferentes. La prueba consistió en generar una señal en banda base de 8 símbolos [00100100] y velocidad 8 bits por segundo, lo que equivale, haciendo corresponder cada símbolo con un evento musical[7]  , con ♩ = 480, muestreada a 4 KHz y modulada, primero con ruido blanco gaussiano y posteriormente con una señal sinusoidal obteniendo una señal modulada en amplitud de doble banda lateral con portadora suprimida. La señal de prueba o patrón de medida de la discriminabilidad límite se obtiene de sumar la señal modulada con una versión desplazada en el tiempo de ella misma. La longitud de esta distancia temporal determina el grado de solapamiento. Para valores mayores que la duración de un símbolo (evento musical simulado) se obtienen eventos separados un intervalo de tiempo igual a la diferencia del tiempo de desplazamiento con el tiempo de duración del símbolo (intervalo de tiempo *t* representado en la Fig. 2.1). La distancia temporal entre eventos se obtiene ajustando este grado de solapamiento.

7 La representación musical dada representa sólo los valores temporales (las figuras) y no frecuenciales. La frecuencia de la portadora generada es de 1000 Hz y no tiene correspondencia diferecta con ningún valor de altura de los instrumentos occidentales. El DO5, la altura más próxima, tiene 1046,5 Hz.

La Fig. 2.1 representa el patrón generado para la evaluación de la discriminabilidad límite en el caso de la modulación con portadora sinusoidal[8].

Con la señal en banda base modulada por el ruido blanco gaussiano normalizado (media nula y desviación típica unidad) se obtuvo un límite de discriminabilidad temporal alrededor de los 5 mseg, mientras que con la señal de doble banda lateral con portadora suprimida se obtuvieron valores en torno a 0.25 mseg[9].

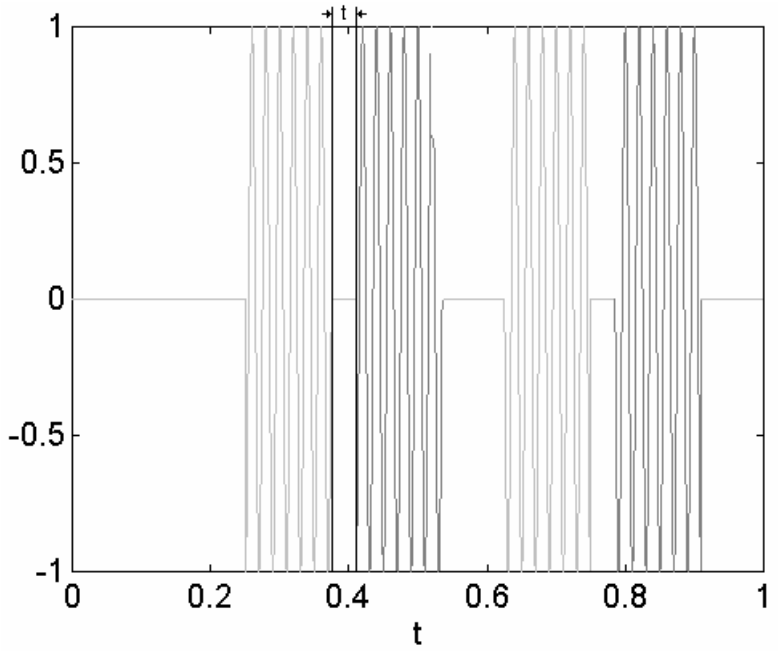


Fig. 2.1 Patrón para la evaluación de la discriminabilidad límite.

Esto significa que el método de segmentación empleado debería ser capaz de discriminar eventos separados entre sí intervalos del orden de 0.25 milisegundos aunque resulte una distancia entre eventos muy poco común para la música más frecuente. Estos resultados de ninguna manera son

8 Aunque las pruebas se realizaron con un tono de 1000 Hz, la frecuencia portadora de la señal representada es de 50 Hz a los efectos de una mejor visualización. A la frecuencia de muestreo la señal de 1 KHz aparece como una gran mancha. También en la Fig. 2.1 se ha exagerado en el solapamiento. La distancia marcada por el intervalo *t* en la parte superior del esquema es de 75 mseg. 5 mseg apenas sería apreciable en la representación. Por último la diferencia de tonos de grises resalta la descomposición de la señal de prueba en las dos componentes que la forman (la señal modulada en amplitud y una versión desplazada de ella misma).

9 Si observamos detalladamente el intervalo *t* que simula la distancia entre eventos alrededor del nivel 0 veremos que los eventos terminan con discontinuidades de fase que probablemente determinan la percepción de intervalos de tiempo tan pequeños. El tiempo de discriminabilidad límite percibido debe ser mucho menor que los que se puedan obtener con casos reales.

conclusivos, dada las condiciones en los que fueron obtenidos, pero si una posible guía para el diseño del segmentador. Los límites de discriminación temporal obtenidos probablemente sean mucho más críticos con respecto a la mayor parte de las composiciones musicales, por las condiciones sintéticas en los que han sido obtenidos donde los cambios abruptos de fase pueden incidir sustancialmente en la percepción. Un cambio de fase de 180º puede inducir la percepción de un tiempo de ataque fantasma. Si el método de segmentación es capaz de resolver éstas diferencias de tiempo, con seguridad podrá operar correctamente en la mayoría de los casos reales. Obsérvese que para ♩. = 480, el valor de una semifusa corresponde a 7.81 mseg.

Tiempo de ataque perceptual

El tiempo de ataque perceptual o TAP[10] es el intervalo que transcurre entre el instante en que se inicia la perturbación de las moléculas del aire y el instante en que percibimos el sonido[11]. Existe una demora inevitable hasta que el sonido es registrado como un nuevo evento después de ser físicamente evidente. Esta demora variable podría, si no se tiene en cuenta, introducir irregularidades rítmicas durante la síntesis de una secuencia musical.

Muchos han sido los esfuerzos en el intento de modelar el TAP. Tiempo de máxima amplitud, umbral de amplitud absoluta, umbral de amplitud relativa, umbral de integración y diversos métodos de umbral de pendiente son sólo algunos ejemplos de estos intentos, pero ninguno parece satisfacer plenamente todos los casos de prueba. No obstante, parecen ajustarse mejor los métodos basados en la pendiente de la envolvente, sobre todo, en el tratamiento de la música percutida. Gordon, por ejemplo, propone el cálculo de la pendiente sobre un ajuste regresivo lineal de la envolvente de amplitud [Gordon,1984].

El TAP en el caso de los sonidos de los instrumentos de percusión suele ser muy pequeño porque las pendientes de ataque suelen ser muy abruptas. En general afecta poco a los instrumentos de ataque rápido introduciendo un desplazamiento constante y no altera la duración de los intervalos entre ataques. Para instrumentos de ataque lento el efecto del TAP dependerá del algoritmo de segmentación empleado, siendo más inmunes aquellos cuya

10 PAT (Perceptual Attack Time).

11 El cual no tiene que coincidir necesariamente con el tiempo en que el sonido es percibido como un evento rítmico.

detección de los eventos esté relacionada con variaciones rápidas de energía y menos aquellos con gran sensibilidad a pequeñas variaciones de la amplitud[12].

En nuestro caso el método de segmentación propuesto trabaja con la velocidad de incremento de la potencia y por lo tanto no es un problema serio a tener en cuenta. Su efecto será un desplazamiento constante respecto a cada tiempo de ataque físico, de manera tal que el intervalo de tiempo entre eventos permanecerá muy estable.

El tiempo en el contexto musical

Algunos investigadores, en lugar de experimentar con clics y mediciones de intervalos aislados, tratan con la percepción musical del tiempo en un contexto musical real, es decir, desde datos de tiempo provenientes de interpretaciones musicales directamente. Para ello, en ausencia de un extractor de datos de tiempo robusto, se han valido de materiales experimentales fuera del contexto natural[13]. Henderson por ejemplo, en 1936 llega a la creencia de un “tempo local” analizando la sección coral del Nocturno No. 6 de Chopin, en el sentido que, tomando en cuenta las fluctuaciones del tempo, las relaciones intencionales entre notas adyacentes se mantienen. Henderson encontró que cuando existe un patrón de duración (como la segunda *negra* de un patrón más corto con respecto al primero), esta relación tiende a mantenerse invariable en el contexto de accelerando, ritardando, crescendo o decrescendo y también que los acordes[14] son a menudo ejecutados como arpegios con distintos

12 Como es el caso de los métodos basados en umbrales de amplitud.

13 Algunos investigadores para realizar un análisis objetivo de la música trataron directamente con datos musicales obtenidos de interpretaciones reales y no de situaciones artificiales. Para ello adaptaron a los instrumentos musicales convencionales objetos de estudio aditamentos y dispositivos que permitieran la graficación objetiva de los eventos musicales en el tiempo desde interpretaciones reales. Carl Seashore, por ejemplo, en 1932, hizo grandes contribuciones al estudio del tiempo en la música de piano, a partir de los datos obtenidos de su “cámara del piano", lo que a la vez podrían ser considerados los primeros pasos hacia la trascripción automática. La “cámara del piano” proporcionaba una grabación fotográfica de los tiempos de ataque, duración, e intensidad relativa de cada nota ejecutada al piano. La cámara era un verdadero “Cubo Goldberg" que consistía en bandas de madera de balsa pegadas al extremo de cada martillo del piano y una complicada secuencia de eventos que grababan la duración y la velocidad de cada nota pulsada sobre una película en movimiento. La grabación fotográfica era luego transcrita a un “pentagrama de patrones musicales” con el gráfico de barras convencional superpuesto en una representación granular del pentagrama musical. [Seashore,1932] Hoy día, cualquier sintetizador convencional ofrece una representación parametrizada de todos los eventos en el tiempo a través del estándar MIDI.

14 Otra área de gran interés es la sincronización de las notas en un acorde.

grados de variación (.01 a .04 segundos para un pianista, .02 a .2 para otro) [Henderson,1936].

En el contexto musical, existe una enorme diferencia entre la desviación aleatoria (como una distribución Gaussiana alrededor de la duración intencional), y una desviación *sistemática* relacionada con al contexto musical [Schloss,1985].

En 1980, estudios realizados por Bengtsson y Gabrielsson reportan que la desviación desde la notación es tan alta como de 15 a 20% para *blancas* y *negras* o de un 20 a 40% para las *corcheas* y *semicorcheas* [Gabrielsson,1974]. Estas grandes variaciones son obviamente perceptibles (para los límites de discriminación temporal analizados) y a pesar de ser tan grandes aparentemente no destruyen la *estructura* del ritmo. Por el contrario, determinan el carácter del *flujo* del ritmo. En otras palabras, estas desviaciones, si son errores aleatorios, simplemente deberían *destruir* la intención rítmica, sin embargo, debido a la colocación de las desviaciones, se escuchan como un efecto de *embellecimiento* del carácter rítmico. Existe una amplia zona de tolerancia alrededor de cada valor y cualquier desviación, por larga que fuera, que caiga dentro de esta zona no cambia la relación temporal-estructural percibida, pero afectan el carácter de movimiento percibido, a veces de una manera sutil [Gabrielsson,1983].

Bengtsson y Gabrielsson hacen una distinción entre el tiempo entre ataques y el intervalo de tiempo desde que una nota termina y comienza el ataque de la próxima. Ellos plantean que aunque es importante detectar este último (normalmente se corresponde con un intervalo de silencio), es el tiempo entre ataques quien caracteriza el ritmo, mientras que el intervalo de silencio caracteriza la *articulación* del ritmo.

Saul Sternberg en 1982 reporta que, en general, los músicos, ante intervalos temporales muy pequeños, le asignan valores demasiado grandes (sobreestimación), y en ambas: producción e imitación de intervalos temporales, producen intervalos demasiado grandes (sobreproducción) [Sternberg,1982].

Percepción del ritmo

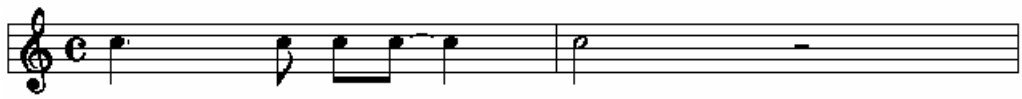
Riemann plantea, acerca de la escencia del ritmo: “Así como la esencia del elemento melódico-armónico es el cambio de altura, la escencia del elemento métrico-rítmico son los cambios de energía, de la *intensidad de los tonos* (dinámica) y de la *rapidez de sucesión de los tonos* (agógica, tempo)” [Riemann,1884].

Cooper y Meyer asumen que debe exitir algún tipo de diferenciación entre los *tonos* que permita distinguir entre pulsos acentuados y no-acentuados. Cooper y Meyer, en lugar de en *términos de duración*, definen el ritmo en *términos de acento*: “El ritmo puede ser definido como la manera en que uno o más pulsos sin acentuar se agrupan alrededor de uno acentuado”. El metro, a su vez, “es la medición del número de pulsos entre acentos repetidos más o menos regularmente”. A pesar de utilizar el acento como primitiva de sus planteamientos no plantean ninguna definición al respecto [Cooper y Meyer,1960].

Yeston, presenta una teoría considerablemente mas consistente, basada sobre dos posibles métodos de análisis: ritmo-a-altura y altura-a-ritmo, lo cual significa “evaluar la altura en términos de su colocación acentual (ritmo-a-altura), mientras, a la vez, se posicione el esquema acentual sobre la base del valor de altura (altura-a-ritmo)”. Yeston intenta desacoplar, o separar el análisis rítmico del análisis de altura [Yeston,1976].

Yeston describe cinco criterios mediante los cuales aislar el ritmo en sub-patrones. Estos sub-patrones son muy importantes para encontrar el estrato rítmico en una pieza. Los criterios son:

Punto de ataque Derivado de la partitura, se refiere a la distancia (duración) entre ataques, en términos de la unidad local más pequeña. Por ejemplo,



tiene la secuencia 3 1 1 3 4.

Timbre Los sub-patrones rítmicos se distinguen por cambios en la instrumentación o por desplazamientos tímbricos, como en un desplazamiento de registro abrupto de un solo instrumento.

Dinámica Ya sean los acentos notados, o el cambio equivalente a nivel dinámico, puede determinar un sub-patrón rítmico sin interpretar[15].

Densidad Se refiere a los cambios en ya sea la "cantidad" de sonido o el número de voces simultáneas en la textura general.

Patrón recursivo Búsqueda de los patrones de unidades repetidas, ya sea en figuras de duración repetidas, o contornos de altura, o combinaciones del mismo.

Inicialmente se observa la música como una "estructura sin interpretar" porque aún no se ha establecido ningún agrupamiento interno. Los sub-patrones rítmicos identificados por los cinco criterios son utilizados para crear el primer estrato, desde el cual se pueda abstraer los eventos estructurales. En cada nivel, los eventos estructurales están dados por las duraciones combinadas de los eventos desde los cuales son abstraídos. En cada nivel, la nueva representación es considerada como una nueva estructura sin interpretar, y se repite el proceso.

Yeston ve la interacción de estos niveles como el factor determinante en la determinación del metro. De hecho dice: "el metro nunca aparece sobre cualquier estrato simple, sino que surge de la interacción de dos estratos, uno de los cuales debería ser siempre de un nivel intermedio".

Muchos han sido los intentos de modelar el ritmo y muchos los puntos de vistas. El conexionismo es otro paradigma que puede resultar atractivo en la búsqueda de una base teórica común del ritmo, pero la mayoría de estos modelos fallan composicionalmente. Esto significa que tal modelo como un todo monolítico podría funcionar bien, pero no es posible descomponer su comportamiento complejo en partes pequeñas significativas. Chandrasekaran en 1990 argumenta que la composicionalidad es una condición para un modelado cognitivo satisfactorio, aún en el paradigma conexionista [Chandrasekaran,1990].

15 Sub-patrón al que aún no se le han asignado los pulsos fuertes y débiles (acentos).

Desain, en 1990, describe el comportamiento de un modelo subsimbólico (conexionista) de cuantización temporal de manera tal que puede ser comparado con cualquier modelo simbólico incompatible desde el paradigma de la inteligencia artificial tradicional y concluye con la abstracción del comportamiento del cuantizador en la forma de “esperanza de eventos” con un patrón temporal como contexto a priori. La esperanza resultante de los patrones temporales complejos puede servir para modelar diversos tópicos como la percepción categórica del ritmo, inducción de reloj y metro, ritmicidad, y la similaridad de secuencias temporales [Desain,1990].

La expectación admite descomposición, lo que permite basar una teoría de percepción de estímulos complejos en un modelo simple por la percepción de sus componentes constituyentes. El concepto de expectación parece explicar la dependencia de la percepción de la estructura rítmica sobre el tempo global, la influencia del contexto sobre la percepción categórica, y otros fenómenos complejos. Desain [Desain,1992] propone su uso como una base común para teorizar acerca de la percepción temporal y la memoria.

El modelo de inducción del pulso propuesto en este trabajo se basa en la teoría de las curvas de expectación descrita por Desain y será explicado detalladamente en el Capítulo 4: Inducción del Pulso, conjuntamente con una propuesta de fundamentación matemática del modelo.

Capítulo 3 Segmentación

Introducción

El análisis de bajo nivel propone un método de segmentación capaz de romper una secuencia musical en piezas o elementos que correspondan a *eventos*[1] musicales como paso previo a la inducción del pulso. Nuestro análisis se limita a caracterizar los eventos u objetos musicales sólo con aquellas propiedades que resultan de interés para el propósito de inducción del pulso tales como la duración y la velocidad de incremento de la potencia pero un análisis extensivo de estas propiedades podría servir de base a aplicaciones de alto nivel como son la generación de notación musical occidental, seguimiento de la altura, análisis sintagmático, etc., o a una generalización del método de inducción del pulso fuera del alcance de este trabajo.

La segmentación debe ser capaz de generar una *lista de eventos* que permita un posterior análisis de alto nivel. Esta lista de eventos es una matriz con tantas filas como eventos tenga la secuencia y tantas columnas como propiedades se requieran para satisfacer el objetivo propuesto. El tiempo puede ser representado, ya sea *relativo*: por intervalos entre eventos consecutivos o *absoluto*: por marcas de ataque donde ocurren fuertes incrementos de energía.

Aunque con un formato diferente esta lista de eventos la generan o aceptan la mayoría de los dispositivos musicales controlados a través del estándar MIDI[2]. La reproducción de una pieza musical a menudo está formada por una secuencia de estructuras del tipo [altura, intensidad] que almacenan las computadoras en archivos de formato estandarizados donde inevitablemente

1 Un *evento* es una estructura que tiene tiempo, altura, duración e intensidad y se corresponde a la idea usual de nota musical. Caracteriza los rasgos físicos del sonido.

2 MIDI (Musical Instruments Digital Interface) Interfaz Digital de Instrumentos Digitales. El MIDI es el estándar por excelencia de interconexión de los instrumentos digitales. Define las características eléctricas y estructuras de datos para el entendimiento armónico entre dispositivos digitales musicales de diversos fabricantes, funciones, arquitectura, etc.

se requiere la adición de un nuevo campo a cada una de estas estructuras: el tiempo. Sólo así, es posible reproducir una obra idénticamente una y mil veces. Tales archivos u organización de estas órdenes digitales son una especie de lista de eventos que los dispositivos digitales entienden y proveen apoyándose en el estándar. Cuando se trabaja directamente con material acústico bruto, tales como interpretaciones reales o grabaciones esta información no es tan evidente. A pesar de lo sencillo que parece distinguir un instrumento en una orquesta, reconocer un patrón rítmico, etc., cuando intentamos automatizar el proceso nos encontramos con numerosas dificultades.

Una *ejecución*[3] es una lista de eventos. La diferencia en el tiempo entre ataques de eventos sucesivos es denominada *intervalo-entre-ataques,* o IEA. Las ejecuciones pueden ser creadas grabando los eventos producidos por un intérprete ejecutando sobre un instrumento MIDI, o producidos directamente de la trascripción de un pentagrama con un tempo conveniente (e.g., el cuarto-de-nota igual a 1 seg). Para diferenciar estas dos formas de producción del material musical se puede hablar de ejecuciones *flexibles* e *inflexibles* [Rosenthal,1992]. Las ejecuciones inflexibles son secuencias de tiempo obtenidas directamente del pentagrama. En una ejecución flexible los tiempos de ataque varían desde sus valores inflexibles alguna cantidad, normalmente pequeña en comparación con el típico IEA. La flexibilidad de una ejecución es una acción voluntaria del ejecutante, que varía el tiempo desde los valores pentagramados a fin de producir algún *efecto* musical[4].

Acercamiento a bajo nivel

Las marcas de tiempo en que ocurren los eventos, llamadas con frecuencia tiempos de ataque[5], son muy difíciles de detectar con absoluta precisión. La energía de los sonidos naturales normalmente varía de forma lenta siguiendo

3 Los términos *ejecución* e *interpretación* serán utilizados indistintamente para referenciar la hecho de la generación conciente del material musical. Ya sea un violinista rasgando con el arco las cuerdas del violín, como el de una computadora secuenciando una lista de notas previamente introducida.

4 Esto es debido a las limitaciones del control motor del ejecutante y a las limitaciones del instrumento.

5 *Onsets* en la literatura anglosajona.

una evolución en el tiempo conocida también como envolvente[6]. La duración prolongada del tiempo de desvanecimiento puede provocar solapamiento del evento que se extingue con su vecino consecutivo. Esta interferencia entre las *colas* de la envolvente de eventos consecutivos *indefine* o *emborrona* sus límites de duración respectivos. La amplitud, en este sentido, es una característica *débil* del material musical, pero sin duda un parámetro importante a considerar en cualquier estudio cuidadoso del tiempo en la música [Schloss,1985]. El tiempo de ataque perceptual está probablemente muy relacionado con la pendiente de la amplitud [Gordon,1984].

Los métodos de detección o localización de los eventos se han desarrollado desde dos puntos de vista diferentes: en el dominio del *tiempo* y en el dominio de la *frecuencia*. Los primeros normalmente en el contexto de análisis de señales cuasiperiódicas[7] con espectro inarmónico[8] como es el caso de las sonidos que producen los instrumentos de percusión [Schloss,1985]. Los segundos normalmente en el contexto de aplicaciones de seguimiento de la altura y de notación.

Análisis localizado de la energía

Existe un conjunto de señales (e.g., música percutida) bien caracterizadas en términos de energía, con ataques breves muy pronunciados y desvanecimientos exponenciales decrecientes. En este contexto, debido a la corta fase de relajación, los solapamientos entre eventos son poco frecuentes, lo que facilita la tarea de segmentación. El algoritmo propuesto trabaja en el dominio del tiempo y se basa en el análisis localizado de la energía:

$$e[n] = \sum_{m=-\infty}^{\infty} x[n]^2 v[n-m] \qquad (3.1)$$

6 La envolvente de amplitud se suele modelar fragmentando la duración total del evento en cuatro componentes de tiempo llamados por la literatura curvas ADSR (Atack/Decay/Sustain/Release) (tiempos de subida/bajada/estabilización/relajación). Según este modelo el ataque o subida es el tiempo que tarda el sonido en alcanzar su máximo nivel; una vez alcanzado el pico de máxima amplitud el sonido disminuye su intensidad hasta cierto nivel donde permanece estable durante un período hasta que finalmente comienza a disminuir hasta extinguirse completamente. El modelo ADSR más simplificado considera variaciones lineales de la amplitud en cada intervalo, otros más complejos introducen segmentación con diferentes pendientes y formas de ondas no lineales como son exponenciales o logarítmicas. Algunos sonidos naturales suelen tener envolventes caprichosas mucho más complejas de modelar.

7 La periodicidad en los sonidos percutidos normalmente es un parámetro débil.

8 Los componentes armónicos no son múltiplos de la frecuencia fundamental.

La energía e_n es la salida de un filtro cuya respuesta al impulso es v_n. El ancho de banda del filtro determina el ancho de banda de e_n y el factor de diezmado que podemos aplicar a e_n sin perder ninguna información. La energía e_n es una versión muestreada de la envolvente a partir de donde se pueden extraer los tiempos de ataque de los eventos con muchas menos muestras.

La parte superior de la Fig. 3.1 corresponde a la representación de un fragmento de señal de un bajo sintetizado, donde se pueden observar cinco eventos. La parte inferior muestra la energía localizada correspondiente al fragmento.

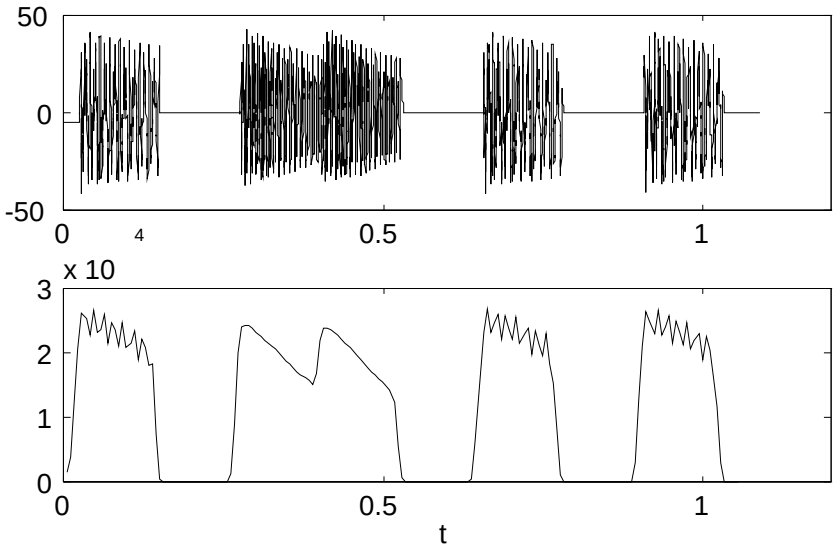


Fig. 3.1 Representación de un fragmento de señal y su energía.

La secuencia e_n representada fue obtenida con una ventana hanning de longitud 30 milisegundos y 80 porcentaje de solapamiento entre ventanas sucesivas. La Fig. 3.1 sugiere una cierta periodicidad visual y permite apreciar como los picos de amplitud de la energía mantienen cierta alineación temporal con los ataques de los eventos. Seleccionando los picos de energía (máximos y mínimos) de e_n obtenemos la secuencia p_n

$$p_n = \left\{ e_{n[k]} \,\middle|\, e_{n[k]} \in P \right\} \tag{3.2}$$

$$e_{n[k]} \in P, si \; \left|e_{n[k-1]}\right| < \left|e_{n[k]}\right| > \left|e_{n[k+1]}\right| \qquad \text{para} \quad k = 2,\ldots,N-1$$

Donde N es el número de muestras de la secuencia p_n.

Para mejorar la representación de los disparos o ataques, se aplica a p_n un operador diferenciador de primer orden

$$q_n = p_n - p_{n-1} \tag{3.3}$$

La aproximación de la primera derivada se comporta como un detector de bordes y produce una señal de salida con las transiciones abruptas acentuadas. Por último, para eliminar ataques espurios se aplica a q_n un operador de extracción de máximos locales deslizante

$$b_n = max\{q_{n-k}, \ldots, q_n, \ldots, q_{n+k}\} \tag{3.4}$$

b_n es el valor máximo sobre una ventana de longitud $L = 2k+1$ muestras. La longitud L de esta ventana deberá ser menor que la mínima distancia entre dos ataques consecutivos, pero es deseable sea lo más ancha posible

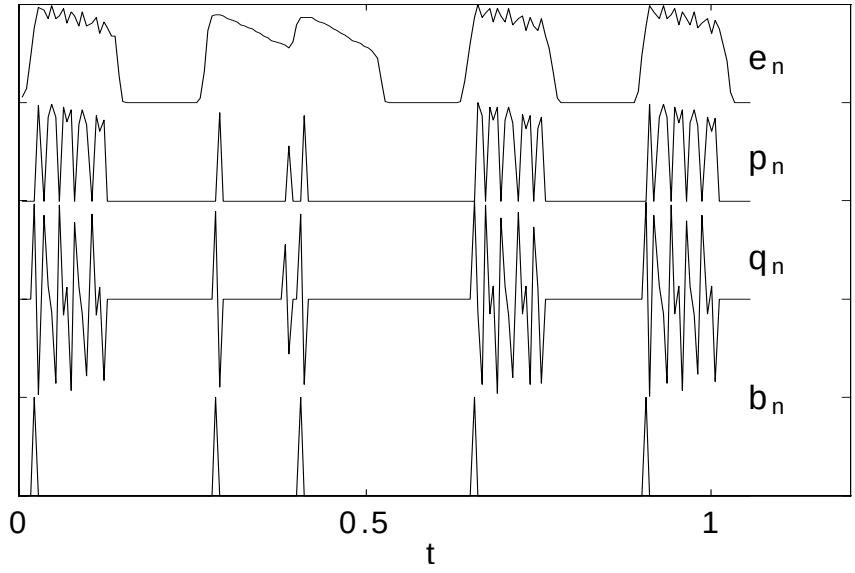


Fig. 3.2 Ejemplo de detección de ataques.

Como se aprecia en la Fig. 3.2, la secuencia b_n toma valor sólo en aquellas muestras que se corresponden al instante de ataque de un evento. Este algoritmo, aunque sencillo y fácil de implementar requiere de un conocimiento *a priori* del material musical para el ajuste de la resolución frecuencial del filtro cuya respuesta al impulso es v_n, y de la longitud de la ventana de extracción de

máximos locales, lo que de cierta manera se corresponde a una estimación del tiempo entre eventos.

Parámetros de la segmentación

El empleo de este algoritmo probablemente requiera de una fase de sintonización o ajuste con algún fragmento significativo[9] del material acústico, lo que en ausencia de algún tipo de mecanismo adaptativo, prácticamente invalida su uso en aplicaciones de tiempo real. Los valores de estos parámetros dependerán en gran medida del material sonoro a analizar y deberán ser ajustados cuidadosamente.

Trama La longitud de la ventana de análisis determina el intervalo en el que se mide la energía. Disminuyendo la longitud de la *trama* se aumenta la resolución temporal. En los casos de prueba se consideraron valores en torno a los 30 mseg.

Solapamiento El porcentaje de solapamiento entre *tramas* permite ajustar el refinamiento temporal de la representación. Con mayor solapamiento se obtiene un mayor número de muestras. Su valor debe ajustarse al mínimo que permita una representación clara y precisa de los eventos. En los casos de prueba se utilizaron valores mayores al 50%.

Frecuencia de muestreo Frecuencia de conversión de la señal analógica a digital. Depende del proceso de adquisición de las muestras y debe garantizar la ausencia de solapes espectrales.

TramaFiltro Longitud de la ventana para el filtrado de los máximos locales. Se establece una región prohibida donde no debe ocurrir otro evento. Su valor debe disminuir según aumente el *tempo* del material musical. En los casos de prueba se utilizaron valores en torno a los 5 mseg. El algoritmo tiene como valor implícito la sexta parte de la *trama.*

9 Entiéndase por *significativos* aquellos fragmentos donde se produzcan las menores distancias temporales entre eventos o probablemente solapamientos entre las colas de las envolventes de dos eventos consecutivos.

Solapamiento entre eventos

Otro elemento a considerar son aquellas situaciones de solapamiento entre envolventes de eventos consecutivos. Los eventos de señal percutiva de espectros inarmónicos a pesar de las normalmente bruscas caídas exponenciales de sus evolventes cuando están muy cercanos pueden provocar enmascaramiento de aquellos eventos de menor energía pero en presencia de señales de espectros armónicos los efectos suelen ser de mayor envergadura sobre todo porque la representación temporal es una mala representación de las características espectrales de la señal. En tales casos el ajuste de los parámetros del algoritmo se torna mas crítico exigiendo cada vez más el conocimiento *a priori* del material musical y nos induce a pensar que considerando las variaciones energéticas en el dominio transformado frecuencial podríamos capturar esta información prácticamente invisible en el dominio temporal. Seleccionemos el siguiente fragmento de partitura para ilustrar el método.

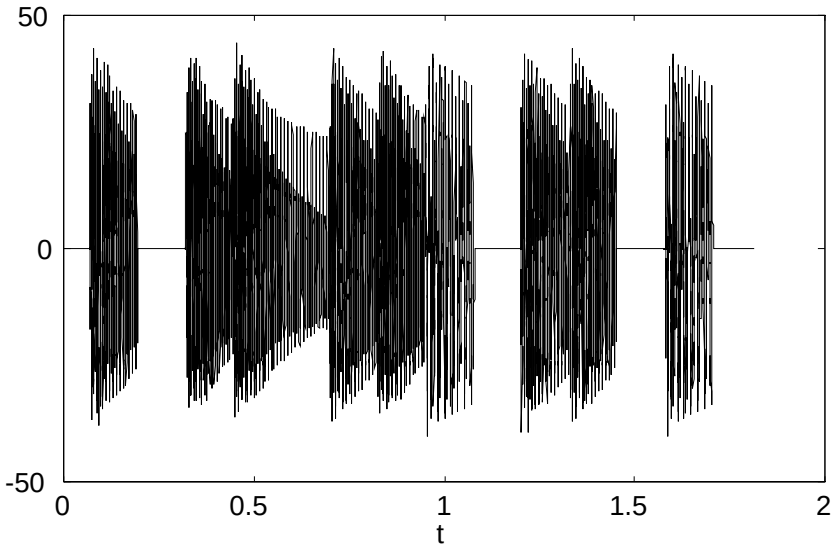


Fig. 3.3 Fragmento de secuencia.

La Fig. 3.3 muestra un caso típico de solapamiento en un fragmento de señal de bajo sintetizado correspondiente a la notación dada arriba. Visualmente parece haber nueve eventos, en lugar de los diez que existen realmente. Ocurre, que además del solapamiento entre las colas de las envolventes, el evento oculto es de más baja frecuencia y energía que sus vecinos y queda completamente enmascarado en la zona de mayor concentración de eventos.

Aplicando el algoritmo propuesto al fragmento donde se encuentra el evento enmascarado con *Trama* 30 mseg, *Solapamiento* 80% y *TramaFiltro* 6 mseg se obtiene la segmentación que muestra la Fig. 3.4. Véase que a pesar de extraer el evento oculto no logra una buena estimación de su posición temporal.

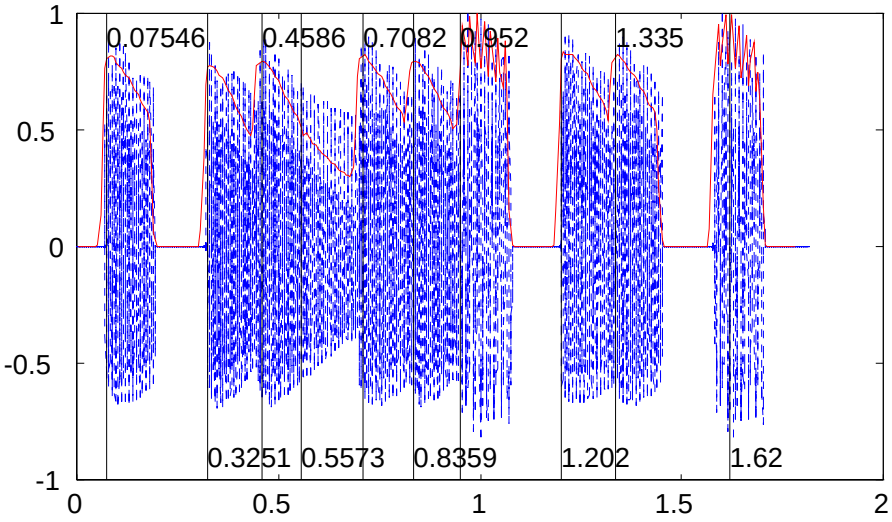


Fig. 3.4 Resultado de aplicar el algoritmo de análisis localizado de energía a un fragmento de señal de bajo sintetizado.

El solapamiento entre eventos de espectros armónicos suele ser más complejo de tratar cuanto mayor sea la dinámica del material musical. En tal situación, cuanto mas fuerte sea el entorno energético menor será la posibilidad de discriminación de aquellos eventos de energía relativamente menor.

Filtrado pre-segmentación

El algoritmo de análisis localizado de la energía podría fallar o producir sesgo en la segmentación en aquellos fragmentos o pasajes donde las envolventes de eventos adyacentes se solapen substancialmente como se muestra en la Fig. 3.4. Sin embargo, la discontinuidad de fase o la brusca interrupción de la oscilación en curso[10] en el dominio temporal, debido al ataque del nuevo evento, provoca una sensible variación en el dominio espectral, con gran

10 En los instrumentos percutidos corresponde a la reinicialización de la membrana ante el nuevo golpe en un instante aleatorio. La alta concentración de energía del ataque provoca una acuciante dispersión o ensanchamiento en el dominio frecuencial.

contenido de altas frecuencias. Este hecho nos permite recuperar los ataques filtrando previamente la señal por un filtro paso alto.

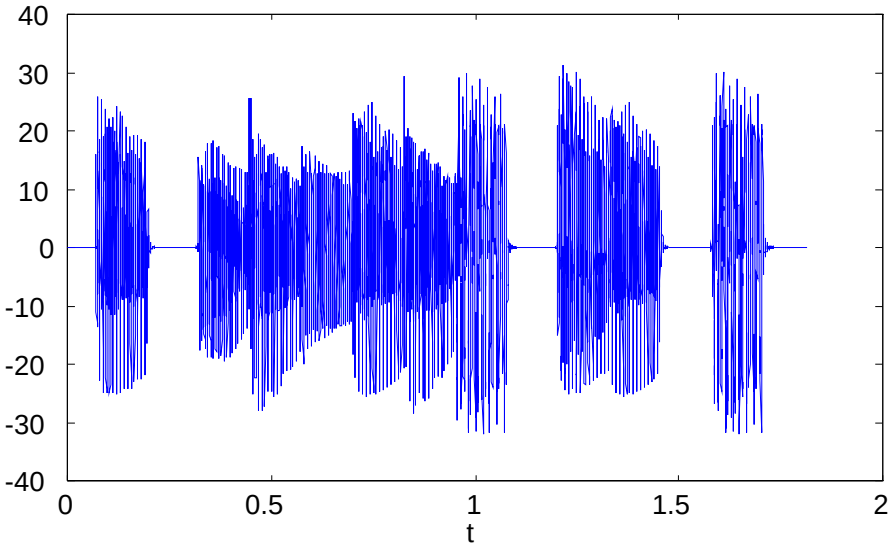


Fig. 3.5 Fragmento de señal filtrada paso alto con un filtro Chebyshev tipo II de orden 4 a 2.5 KHz.

La Fig. 3.5 muestra el resultado de filtrar paso alto la secuencia de la Fig. 3.3. Como se puede apreciar el filtro logra extraer el evento oculto. En la Fig. 3.6 se muestra el resultado de aplicar el algoritmo al fragmento de señal previamente filtrada paso alto. Esta vez, los ataques obtenidos, además de ser la cantidad correcta están mejor alineados temporalmente.

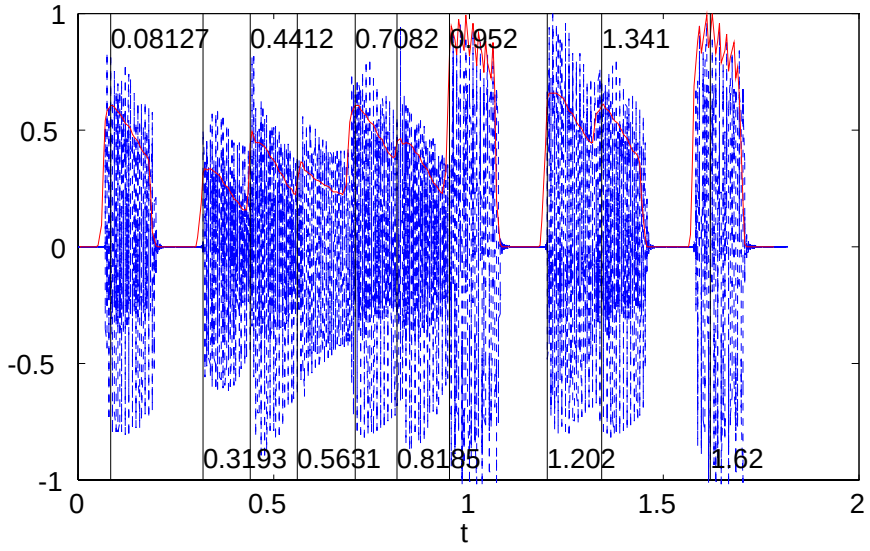


Fig. 3.6 Segmentación de la señal previamente filtrada paso alto.

Aumentando la frecuencia de corte del filtro se puede llegar a obtener una señal representativa únicamente de las grandes variaciones o discontinuidades temporales a partir de la cual se puedan detectar fácilmente los ataques[11]. Aunque no debemos olvidar que la presencia de ruido de alta frecuencia podría dar lugar a falsas interpretaciones en la segmentación de los eventos. Elemento que debe tenerse muy en cuenta cuando se manipulan señales obtenidas del mundo real.

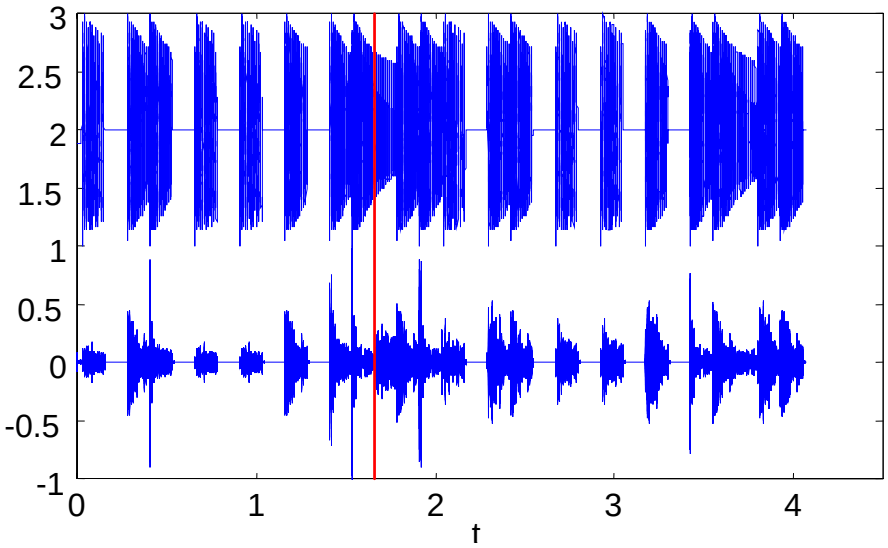


Fig. 3.7 Efecto del filtrado paso alto. En la parte superior se muestra un fragmento de bajo sintetizado y en la parte inferior la señal obtenida a la salida de un filtro paso alto elíptico (6 polos, 6 ceros) a 5KHz. La marca señaliza la posición temporal del evento oculto.

El sistema auditivo es muy sensible a las discontinuidades de fase. Si a una señal se le añade ruido de bajo nivel, pero se retiene la relación de fase inicial, nuestro oído escuchará los *clics* debidos al ruido, pero no percibirá un nuevo ataque. Sin embargo, si en lugar de añadir ruido se provoca artificialmente una discontinuidad de fase de 180º entonces nuestro oído probablemente percibirá un nuevo ataque [Schloss,1985]. En la Fig. 3.7 se observa el efecto de realce de los ataques al prefiltrar paso alto la secuencia musical sintética a segmentar.

11 El aumento de la frecuencia de corte del filtro aumenta la legibilidad y aislamiento entre eventos.

Discriminabilidad límite

El límite de discriminabilidad entre dos eventos consecutivos muy cercanos en el tiempo ha sido estudiado por diversos investigadores y todos parecen estar de acuerdo que alrededor de 5 mseg es suficiente para determinar los tiempos entre eventos capturando la información de tiempo esencial (intencional) para la mayoría de los casos.

Para verificar la discriminabilidad del algoritmo se generó una señal sintética descrita en el apartado *Discriminación temporal* del Capítulo 2. La señal base está compuesta de dos eventos y su solapamiento con una versión de ella misma desplazada determinado el intervalo de tiempo entre eventos. La señal base está formada por 8 símbolos[12] a un *tempo* de 8 bits por segundo muestreados a 4 KHz y modulados con una portadora sinusoidal de frecuencia 1 KHz. La señal base está compuesta por los símbolos [0 0 1 0 0 1 0 0][13] cada uno de duración 125 ms. Los símbolos con valor 1 corresponden a un evento.

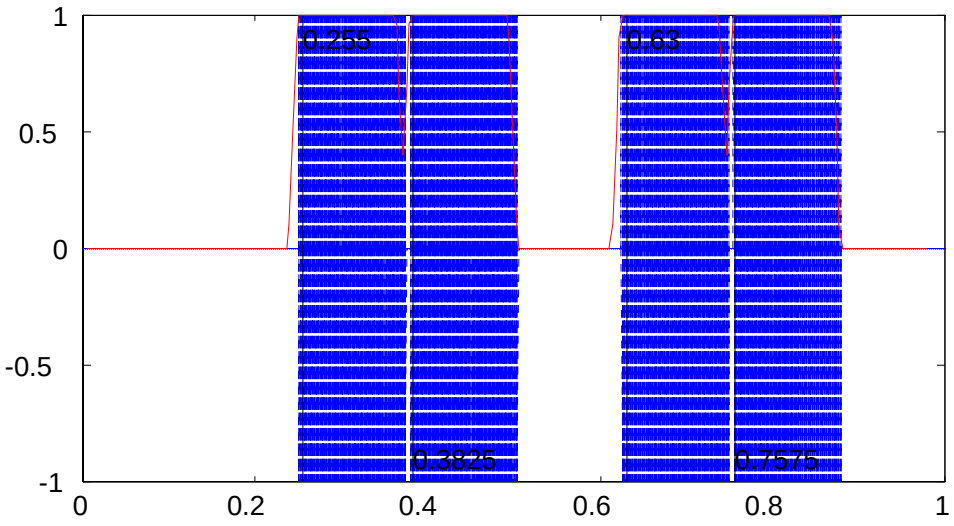


Fig. 3.8 Resultado de aplicar el algoritmo de análisis localizado de energía a una señal con eventos separados entre sí 5 mseg.

La Fig. 3.8 muestra el resultado de aplicar el algoritmo de análisis localizado de energía a una señal con intervalo de tiempo entre eventos de 5 ms[14]. Para este caso se utilizaron los parámetros *Trama* 20 mseg, *Solapamiento* 80% y

12 Cada bit representa a un símbolo. La duración de cada símbolo es de 125 ms (el inverso de la frecuencia de símbolos 1/8).

13 Silencio de 250 ms, tono de 1 Khz de 125 ms,...

14 Para obtener una distancia entre eventos de 5 mseg se requiere un desplazamiento de la señal base de 130 mseg porque la duración de cada símbolo es de 125 mseg.

TramaFiltro 35 mseg. Como se aprecia en la figura si restamos los tiempos de ataques correspondientes a dos evento consecutivos obtendremos un intervalo entre eventos de 127.5 ms con un error de 2.5 ms respecto a la distancia real.

Análisis en el dominio frecuencial

Las señales musicales pertenecen a la clase de señales no-estacionarias[15] cuyas propiedades varían con el tiempo, y pueden ser modeladas por la suma de componentes sinusoidales cuyas amplitudes, frecuencias o fases varían con el tiempo. La estimación de una sola DFT no es suficiente para describir tales señales, y como resultado se llega al concepto de la *Transformada de Fourier dependiente del tiempo*, también denominada como *Transformada de Fourier localizada[16].*

La transformada de Fourier localizada de una señal $x[n]$ se define como

$$X[n,\Omega) = \sum_{m=-\infty}^{\infty} x[n+m]v[m]e^{-j\Omega m} \tag{3.5}$$

donde $v[m]$ es una secuencia ventana. La Ec. (3.5) puede ser interpretada como la transformada de Fourier de $x[n+m]$ vista a través de la ventana $v[m]$. Se introduce un parámetro de *frecuencia local* (local en el tiempo) de manera tal que la transformada de Fourier *local* ve la señal a través de una ventana sobre la cual es aproximadamente estacionaria. Esta representación $X[n,\Omega)$ es similar a la notación utilizada en el pentagrama musical, que también representa “frecuencias” a través del tiempo[17], aunque una representación más real se correspondería con la Ec. (3.7) dado que las notas musicales habituales

15 La noción de estacionariedad es formalizada en la literatura de procesado de señales estadístico.

16 La discusión de la transformada de Fourier dependiente del tiempo se puede encontrar en una variedad de referencias incluyendo Allen y Rabiner (1977), Rabiner y Schafer (1978), Crochiere y Rabiner (1983), Nawab y Quatieri (1988) y Oppenheim y Schafer (1989).

17 Esta analogía tal vez llevó a Charles Seeger en 1951 al convencimiento de que una representación gráfica de la música era la respuesta para su estudio objetivo. “Tengo el presentimiento que antes de que pasen cien años nuestra actual notación se verá más como un método gráfico que simbólico de escritura” [Seeger,1951]. Seis años más tarde Seeger moderó un poco su posición mencionando que la notación tradicional complementa la representación gráfica. Y la cuestión es que la representación gráfica, aún siendo una grabación objetiva de los eventos en el tiempo, no representa las abstracciones *musicales* tan importantes perceptualmente.

se corresponden a sistemas de escalas discretas, es decir, sólo un subconjunto de todas las Ω[18] frecuencias posibles.

El conjunto explícito de $X[n,\Omega)$ sólo es posible en un conjunto finito de valores de Ω, correspondientes al muestreo de la transformada de Fourier dependiente del tiempo en el dominio de la variable frecuencia. Al igual que las señales de longitud finita se pueden representar exactamente muestreando la transformada de Fourier de tiempo discreto, las señales de longitud indeterminada se pueden representar mediante muestras de la transformada de Fourier dependiente del tiempo si la ventana $v[m]$ de la Ec. (3.5) tiene longitud finita. Si suponemos que la ventana tiene longitud L y que las muestras comienzan en $m = 0$[19],

$$v[m] = 0 \qquad \text{fuera del intervalo } 0 \le m \le L\text{-}1 \tag{3.6}$$

si muestreamos $X[n,\Omega)$ en K frecuencias igualmente espaciadas $\Omega_k = \frac{2\pi}{K}k$, con $K \not< L$, aún podemos recuperar la secuencia original desde la transformada de Fourier dependiente del tiempo muestreada. Específicamente, si definimos $X[n,k]$ como

$$X[n,k] = X\left[n, \frac{2\pi}{K}k\right] = \sum_{m=0}^{L-1} x[n+m]v[m]e^{-j\frac{2\pi}{K}km} \qquad ,0 \le k \le K-1 \tag{3.7}$$

Luego $X[n,k]$ es la DFT de la secuencia enventanada $x[n+m]v[m]$. La Ec. (3.7) corresponde al muestreo de la Ec. (3.5) en Ω. El análisis sugiere que se

18 Se denota la variable de frecuencia de la transformada de Fourier dependiente del tiempo por Ω para mantener una distinción con la variable de frecuencia de la transformada de Fourier de tiempo-discreto convencional, denotada por *w*. Se utiliza la notación corchete-paréntesis mezclada para destacar que n es una variable discreta y Ω una variable continua.

19 La ventana descrita en (3.6) es no causal. Se podría haber utilizado una ventana causal con $v[m] \neq 0$ para $-(L-1) \le m \le 0$ ó una ventana simétrica tal que $v[m] = v[-m]$ para $|m| \le (L-1)/2$,con L entero impar. El uso de una ventana causal es simplemente más conveniente porque ilustra de manera muy natural la interpretación de la transformada de Fourier dependiente del tiempo muestreada como la DFT de la secuencia enventanada que comienza en la muestra *n*.

puede reconstruir $x[n]$ si $X[n,\Omega)$ ó $X[n,k]$ está muestreada en la dimensión del tiempo también. Específicamente, utilizando la ecuación[20]

$$x[n+m] = \frac{1}{Kv[m]} \sum_{k=0}^{K-1} X[n,k] e^{j\frac{2\pi}{K}km} \qquad ,0 \le m \le L-1 \tag{3.8}$$

se puede reconstruir la señal en el intervalo $n_0 \le n \le n_0 + L - 1$ desde $X[n_0, k]$, y luego reconstruir la señal en el intervalo $n_0 + L \le n \le n_0 + 2L - 1$ desde $X[n_0 + L, k]$, etc. Por lo tanto $x[n]$ puede ser reconstruida exactamente desde la transformada de Fourier dependiente del tiempo muestreada en ambas dimensiones: la frecuencia y el tiempo. En general, dada la ventana especificada en (3.6) podemos definir la transformada de Fourier dependiente del tiempo como

$$X[rR,k] = X\left[rR, \frac{2\pi}{K}k\right] = \sum_{m=0}^{L-1} x[rR+m]v[m] e^{-j\frac{2\pi}{K}km} \tag{3.9}$$

donde r y k son enteros tal que $-\infty < r < \infty$ y $0 \le k \le K-1$. Para mayor simplificación se define

$$X_r[k] = X[rR,k] = X[rR,\Omega_k) \qquad -\infty < r < \infty,\ 0 \le k \le K-1 \tag{3.10}$$

donde $\Omega_k = \frac{2\pi}{K}k$. Esta notación denota explícitamente que la transformada de Fourier dependiente del tiempo muestreada es simplemente una secuencia de DFTs de K puntos de los segmentos de señal enventanados.

$$x_r[m] = x[rR+m]v[m] \qquad -\infty < r < \infty,\ 0 \le m \le L-1 \tag{3.11}$$

La Ec. (3.9) involucra los siguientes parámetros enteros: $\boldsymbol{L}$ es la longitud de la ventana de análisis, $\boldsymbol{K}$ es el número de muestras en la dimensión de frecuencia o longitud de la DFT y $\boldsymbol{R}$ es el intervalo de muestreo en la dimensión del tiempo.

---

20 Como se asumió una ventana $v[m] \neq 0$ para $0 \le m \le L-1$, los valores de la secuencia $x[n]$ pueden ser recuperados en el intervalo desde n hasta $(n+L-1)$ utilizando la Ec. (3.8).

No todas las posibilidades de estos parámetros permiten la reconstrucción exacta de la señal. La opción $L \leq K$ garantiza que podemos reconstruir los segmentos enventanados $x_r[m]$ desde los bloques transformados $X_r[k]$. Si $R < L$, los segmentos se solapan, pero si $R > L$, algunas de las muestras de la señal no serán utilizadas y por lo tanto no podrán ser reconstruidas desde $X_r[k]$. Por lo tanto, en general, los tres parámetros de muestreo deben satisfacer la relación $K \geq L \geq R$.

Tanto el análisis como la síntesis de Fourier dependen críticamente de la ventana $v[m]$ seleccionada. El propósito primario de la ventana en la transformada de Fourier dependiente del tiempo es limitar la extensión de la secuencia a transformar de manera tal que las características espectrales sean razonablemente estacionarias sobre la duración de la ventana. Mientras más rápido cambien las características de la señal, más estrecha deberá de ser la ventana. Pero mientras más corta sea la duración de la ventana, menor será la resolución frecuencial.

Consecuentemente la selección de la ventana establece un compromiso entre la resolución frecuencial y la resolución temporal. Un punto de vista alternativo del mismo proceso se basa en la interpretación de un banco de filtros. A una frecuencia dada Ω_k, la cantidad filtrada de señal "de todos los tiempos" con un filtro paso-banda que tiene como respuesta al impulso la función ventana modulada a dicha frecuencia. Dada una función ventana $v[n]$ y su transformada de Fourier $V(\Omega_k)$, el ancho de banda del filtro $\Delta\Omega_k$ se define como

$$\Delta\Omega_k{}^2 = \frac{\sum \Omega_k^2 \left|V(\Omega k)\right|^2}{\sum \left|V(\Omega_k)\right|^2} \tag{3.12}$$

donde el denominador es la energía de $v[n]$. Dos sinusoides serán discriminadas sólo si su separación es mayor que $\Delta\Omega_k$. Por lo tanto la resolución en frecuencia de la transformada de Fourier localizada de análisis está dada por $\Delta\Omega_k$. De manera similar la apertura en el tiempo está dada por

$$\Delta n^2 = \frac{\sum n^2 \left|v[n]\right|^2}{\sum \left|v[n]\right|^2} \tag{3.13}$$

donde el denominador es nuevamente la energía de $v[n]$. Dos pulsos en el tiempo podrán ser discriminados sólo si están separados un intervalo mayor que Δn. La resolución en tiempo y frecuencia no puede ser arbitrariamente pequeña, porque su producto tiene cota inferior

$$\text{Producto (tiempo - ancho de banda)} = \Delta n \Delta \Omega_k \geq \frac{1}{4\pi} \tag{3.14}$$

Este límite se conoce como principio de incertidumbre o desigualdad de Heisenberg. Esto significa que uno sólo puede negociar la resolución en tiempo para la resolución en frecuencia, o *vice versa.*

Lo más importante es que, una vez seleccionada una ventana para la transformada localizada de Fourier, la resolución en frecuencia dada por (3.12) y (3.13) es *fija* sobre el plano tiempo-frecuencia entero[21].

Densidad espectral de potencia

La señal musical se modela mejor como señal aleatoria porque el proceso que la genera es demasiado complejo para un modelo determinístico razonable. Las señales musicales son cuasi-estacionarias[22] y muchas de sus características suelen ser bien representadas por sus valores promedios como el valor medio (nivel de directa), varianza (potencia promedio), función de autocorrelación, o densidad espectral de potencia.

La densidad espectral de potencia representa el comportamiento en frecuencia de una secuencia aleatoria estacionaria. Se define por

$$S_{xx}(\Omega) = F\{R_{xx}[m]\} = \sum_{m=-\infty}^{\infty} R_{xx}[m] e^{-j\Omega m} \quad ,|\Omega| \leq \pi \tag{3.15}$$

siendo $R_{xx}[m]$ la secuencia de autocorrelación definida por

$$R_{xx}[m] = E\{x_{n+m} x_n^*\} = \lim_{M\to\infty} \frac{1}{M} \sum_{n=0}^{M-1} x[n+m] x^*[n] \tag{3.16}$$

21 Se usa la misma ventana para todas las frecuencias.

22 La definición de cuasi-estacionariedad es formalizada en la literatura de procesado de señales estadístico.

$S_{xx}(\Omega)$ mide la distribución de potencias de un registro de longitud finita M de una realización x_n en el espectro normalizado (-π,π).

$$X_n = X_0, X_1, X_2, \ldots, X_{M-1}$$

El *periodograma* es definido como el módulo al cuadrado de la transformada de Fourier y tiene la ventaja de proporcionar una mayor eficiencia computacional al estar basado en la DFT.

$$S_{xx}(\Omega) = \frac{1}{LU}\left|X(\Omega)\right|^2 \qquad ,|\Omega| \le \pi \tag{3.17}$$

siendo

$$X(\Omega) = \sum_{m=0}^{L-1} x[n+m]v[m]e^{-j\Omega m} \tag{3.18}$$

Una versión modificada del periodograma permite reducir la dispersión espectral[23] y obtener la distribución de la energía de la señal en el plano tiempo-frecuencia. El método de estimación espectral mediante promediado de estimaciones de Welch combina los algoritmos computacionales de la DFT con el uso de una ventana de datos $v[n]$.

En el método de promediado de estimaciones de Welch la secuencia de datos $x[n],\ 0 \le n \le M-1,$ es dividida en N segmentos de longitud L muestras (con una ventana de longitud L aplicada a cada segmento) definidos en la Ec. (3.11).

Si $R < L$ los segmentos se solapan, y si $R = L$ los segmentos son contiguos. M es la longitud de los datos disponibles. El número total de segmentos depende de R, L y M. Específicamente existirán N segmentos completos de longitud L. Donde N es el entero más grande para el cual $(N-1)R + (L-1) \le M-1$. El periodograma para el segmento *r*-ésimo es

$$S_r(\Omega) = \frac{1}{LU}\left|X_r(\Omega)\right|^2 \tag{3.19}$$

23 La dispersión espectral o *leakage* consiste en la aparición de frecuencias espúreas en forma de lóbulos laterales y se debe a las discontinuidades en el extremos del intervalo. La dispersión espectral se puede mejorar enventanando los datos porque reduce las discontinuidades de los bordes pero ello empeora la resolución espectral porque reduce la longitud efectiva del registro.

donde U [24] es una constante necesaria para eliminar el sesgo en la estimación espectral. La escala puede ser ajustada seleccionando la constante de normalización U de manera tal que

$$\frac{1}{2\pi LU}\int_{-\pi}^{\pi}\left|V(\Omega)\right|^2 d\Omega = \frac{1}{LU}\sum_{n=0}^{L-1}\left(v[n]\right)^2 = 1, \tag{3.20}$$

ó

$$U = \frac{1}{L}\sum_{n=0}^{L-1}\left(v[n]\right)^2 \tag{3.21}$$

El método de Welch consiste en promediar los N periodogramas estimados $S_r(\Omega)$

$$S(\Omega) = \frac{1}{N}\sum_{r=0}^{N-1} S_r(\Omega) \tag{3.22}$$

La varianza de la suma de las N variables aleatorias independientes distribuidas idénticamente es $1/N$ veces la varianza de cada variable aleatoria individual [Papoulis,1984]. Por lo tanto la varianza del periodograma promedio para una ventana rectangular sin solapamiento es

$$\mathrm{var}\left[S(\Omega)\right] \approx \frac{1}{N} S^2(\Omega) \tag{3.23}$$

La varianza de $S(\Omega)$ es inversamente proporcional al número de periodogramas promediados, y según N incrementa la varianza tiende a cero. Sin embargo, para la longitud de datos total M, el número total de segmentos (asumiendo $L = R$) es M/L; por lo tanto, si L incrementa, N decrementa y de acuerdo a la Ec. (3.23) la varianza de $S(\Omega)$ incrementará. Por lo tanto, como es típico en los problemas de estimación estadística, para longitud de datos fija existe un compromiso entre el sesgo y la varianza. Sin embargo, cuando la longitud de los datos M incrementa, se puede incrementar ambos L y N, de manera tal que cuando M tiende a ∞, el sesgo y la varianza de $S(\Omega)$ tienden a cero. Consecuentemente, el promediado de las estimaciones provee una

---

24 $U = 1$ para una ventana rectangular. Para otras ventanas, si $v[n]$ está normalizado a valor máximo 1, $0 < U < 1$. El periodograma modificado, si está adecuadamente normalizado, es asintóticamente insesgado: e.g., el sesgo tiende a cero según aumenta la longitud de la ventana.

estimación de la densidad espectral de potencia $S(\Omega)$ asintóticamente insesgada y consistente.

Si en lugar de ventanas rectangulares se utilizan ventanas de forma diferente la varianza del periodograma promedio se comporta mejor que como en la Ec. (3.23). Welch también consideró el solapamiento entre ventanas y demostró que si el solapamiento es la mitad de la longitud de la ventana, la varianza se reduce casi por un factor de 2 debido al doblado del número de secciones. Por debajo de esta cantidad el aumento de la longitud de solapamiento no contribuye a reducir la varianza porque los segmentos son cada vez menos independientes según aumenta el solapamiento.

El periodograma promedio puede ser evaluado explícitamente sólo sobre un conjunto discreto de frecuencias. Debido a la disponibilidad de la FFT[25] para calcular la DFT, una opción particularmente conveniente y ampliamente usada es la selección de las frecuencias $\Omega_k = 2\pi k/K$ para un adecuado valor de K. De la Ec. (3.22) podemos ver que si sustituimos la DFT de $\chi_r[n]$ por la transformada de Fourier de $\chi_r[n]$ en la Ec. (3.19), obtendremos muestras de $S(\Omega)$ en las frecuencias $\Omega_k = \frac{2\pi}{K}k \;, k = 0,1,\ldots,K-1.$ Específicamente, con $\chi_r[k]$ denotando la DFT de $\chi_r[n]$,

$$S_r[k] = S_r(\Omega_k) = \frac{1}{LU}\left|X_r[k]\right|^2 \tag{3.24a}$$

$$S[k] = S(\Omega_k) = \frac{1}{N}\sum_{r=0}^{N-1} S_r[k] \tag{3.24b}$$

Si denotamos $S_r(2\pi k/K)$ como la secuencia $S_r[k]$ y $S(2\pi k/K)$ como la secuencia $S[k]$. De acuerdo a las Ecs. (3.24), el promediado de las estimaciones del espectro de potencia es calculado en K frecuencias igualmente espaciadas promediando las DFTs de los segmentos de datos enventanados con el factor de normalización LU. Este método de estimación del espectro de potencia provee un estructura muy conveniente dentro de la cual existe un compromiso entre la resolución y la varianza de la estimación espectral. Es particularmente simple y eficiente de implementar utilizando la FFT y garantiza que la estimación del espectro es siempre no negativa.

25 Algoritmo de la transformada rápida de Fourier.

Por último, podemos expresar la densidad espectral de potencia en términos de dos variables discretas que representan el plano tiempo-frecuencia muestreado, a partir del cual se puede reconstruir exactamente la señal

$$S[n,k] = S[rR,k] = S[rR,\Omega_k] = \frac{1}{N}\sum_{r=0}^{N-1}\frac{1}{L}\left|\sum_{m=0}^{L-1} x[rR+m]v[m]e^{-j\frac{2\pi}{K}km}\right|^2 \tag{3.25}$$

donde $\Omega_k = \frac{2\pi}{K}k \quad , k = 0,1,2,\ldots,K-1$. Se supone la constante de normalización $U = 1$. La densidad espectral de potencia $S[n,k]$ se puede representar como una matriz $\boldsymbol{S}$ de K filas y N columnas correspondientes al eje de frecuencias y al eje de tiempos respectivamente.

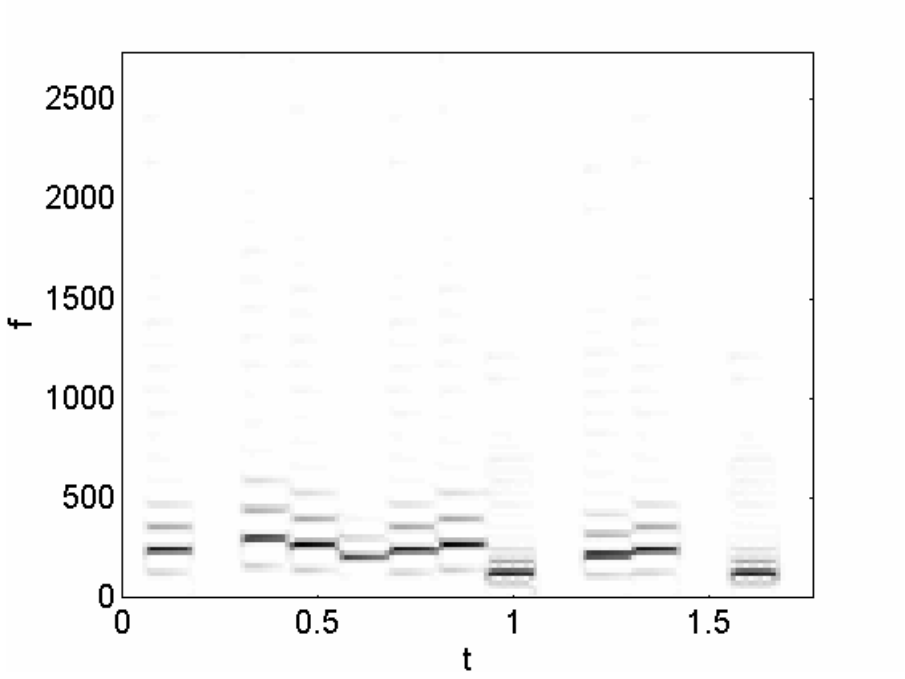


Fig. 3.9 Espectrograma de un fragmento de señal de bajo sintetizado.

La Fig. 3.9 muestra el espectrograma[26] del fragmento de señal sintetizado de la Fig. 3.3. La señal tiene una longitud de 1001 muestras y fue muestreada a 5512.5 Hz. La densidad espectral de potencia se obtuvo con una ventana hanning de 46.45 mseg (L=256), resolución espectral de 21.53 Hz (K=256) y un solapamiento del 75% (256*0.75=192 muestras). El eje de tiempos tiene una resolución de 11.61 mseg (R=256-192=64 muestras). El eje horizontal

26 El espectograma se obtiene de diezmar la variable temporal en la Ec. (3.5), es decir, calcular el espectro local desplazando la ventana R muestras cada vez. Es la representación de la secuencia de los periodogramas obtenidos.

representa el tiempo en unidades de segundos. El eje vertical es la frecuencia en el espectrograma y la escala de color representa las amplitudes.

En la Fig. 3.9 es posible reconocer visualmente los 10 eventos musicales que forman la secuencia, inclusive el evento oculto. Obsérvese que las zonas de mayor energía están concentradas en los valores de frecuencias correspondientes del eje vertical y alineadas en el eje horizontal con los tiempos ataque de los eventos correspondientes.

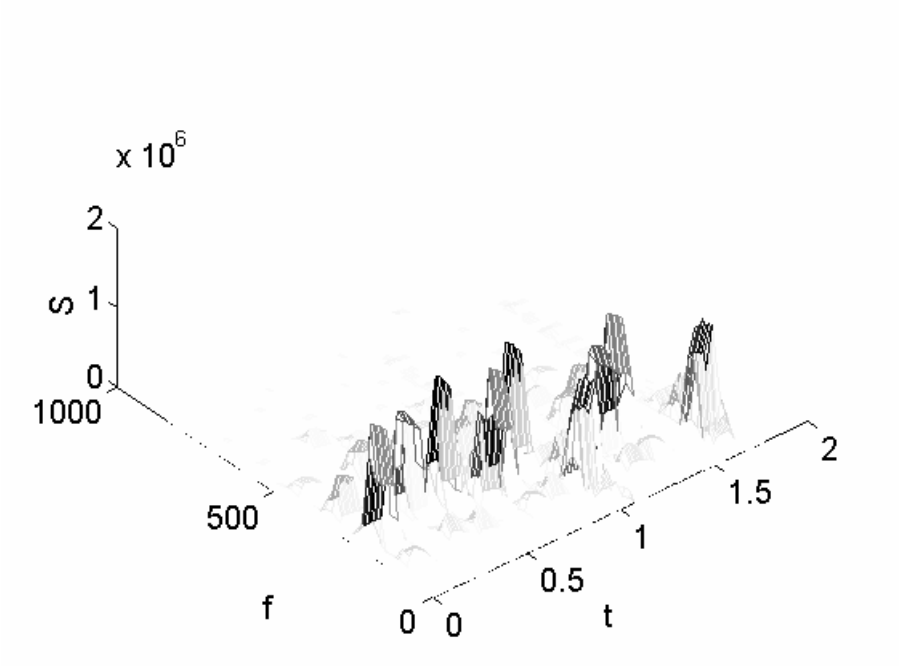


Fig. 3.10 Representación alternativa del fragmento de señal de bajo sintetizado.

La Fig. 3.10 muestra una representación tridimensional alternativa del mismo fragmento donde se puede apreciar claramente la distribución espectral de los eventos en el plano tiempo-frecuencia. Esta señal tiene mayor contenido espectral en la bajas frecuencias (de hecho corresponde a la señal de un bajo sintetizado). Este hecho sugiere un análisis del contenido energético de la señal sólo en aquellas zonas frecuenciales “activas”, aunque esto supone una estimación *a priori* del contenido espectral de la señal. Véase que la Fig. 3.10 tiene el eje de frecuencias truncado y sin embargo, como cubre el rango de mayor contenido espectral contiene toda la información necesaria para determinar los instantes donde ocurran transiciones bruscas de energía.

La Fig. 3.10 sugiere que los eventos están relacionados con transiciones bruscas de energía o velocidad de incremento de la potencia. Masataka Goto y Yoichi Muraoka en 1994 propusieron un método de localización de eventos o segmentación basado en este principio [Goto y Muraoka,1994]. Su propuesta de localización temporal de los pulsos en tiempo real, se basa en un complejo sistema de procesamiento paralelo que examina, a partir de la señal acústica real, múltiples hipótesis de posiciones del pulso simultáneamente, calcula la confiabilidad de cada tiempo de ataque y descarta aquellas de menor valor. A partir de los tiempos de ataque más "confiables", un conjunto de "agentes programáticos" predicen la localización del próximo pulso de acuerdo a sus propias estrategias. La posición del próximo pulso se determina sobre la base del agente que aporte mayor confiabilidad.

La extracción de los componentes o eventos consiste en evaluar a partir de la densidad espectral de potencia discreta bidimensional $S[n,k]$ aquellos componentes de mayor velocidad de incremento de potencia.

$$p[n,k] \begin{cases} S[n,k] > pp \\ np > pp \end{cases} \tag{3.26}$$

donde $S[n,k]$ es la potencia del espectro de la frecuencia discreta k en el tiempo discreto n, y pp y np vienen dados por

$$pp = max\{S[n-1,k], S[n-1,k\pm1], S[n-2,k]\} \tag{3.27}$$

$$np = min\{S[n+1,k], S[n+1,k\pm1]\} \tag{3.28}$$

Estas condiciones extraen las componentes de frecuencia cuya potencia ha sido incrementada. El grado de ataque $d[n,k]$ está dado por

$$d[n,k] = S[n,k] - pp + max\{0, S[n+1,k] - S[n,k]\} \tag{3.29}$$

La Fig. 3.11 ilustra el algoritmo de selección de los $p[n,k]$ componentes de mayor rapidez de incremento de potencia. Obsérvese que toma en cuenta, no sólo el incremento de potencia respecto al instante anterior, sino que también una vecindad de valores pasados y "futuros" y el par de frecuencias adyacentes próximas. Este algoritmo logra extraer los componentes de mayor rapidez de

cambio de potencia (relacionados perceptualmente con la aparición de los eventos) relativos en una vecindad del plano tiempo-frecuencia.

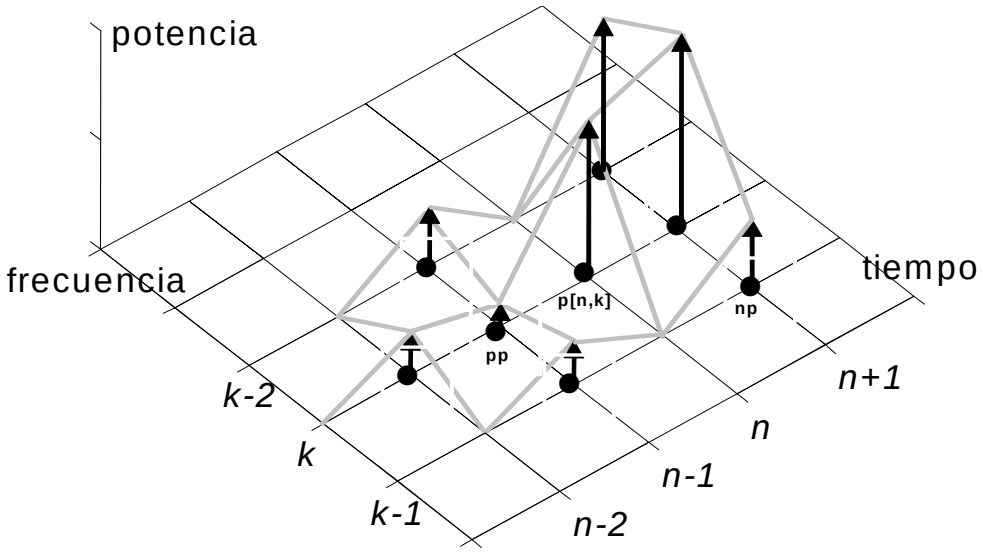


Fig. 3.11 Extracción de los componentes de ataque.

Para encontrar el tiempo de ataque o marca del evento primero es necesario sumar todas las contribuciones de máximo incremento de potencia en el eje de frecuencia, a partir de los grados de ataque obtenidos en la Ec. (3.29). De esta operación se obtiene una secuencia unidimensional de longitud N que representa las variaciones potenciales de mayor velocidad a lo largo del tiempo definida por

$$D_n = D[n] = \sum_k d[n,k] \tag{3.30}$$

Los tiempos de ataque se corresponden con los tiempos de los máximos de la secuencia D_n. Antes de obtener los tiempos correspondientes a estos picos, se puede suavizar D_n mediante un kernel de convolución. Este proceso, aunque disminuye la varianza de D_n disminuye también la resolución temporal por lo que debe aplicarse con mucho cuidado. La longitud del kernel de convolución determina la *sensibilidad.* Si el kernel es lo suficientemente grande como para emborronar significativamente las crestas de potencia, podemos perder algún tiempo de ataque o introducir cierto sesgo. Por el contrario mientras menor sea la longitud del kernel de convolución mayor será la sensibilidad. Otro parámetro que puede orientar la búsqueda de los eventos es el *rango de frecuencia* en la sumatoria de D_n en la Ec. (3.30). Limitando este

rango es posible encontrar tiempos de ataque en rangos de frecuencia diferentes[27].

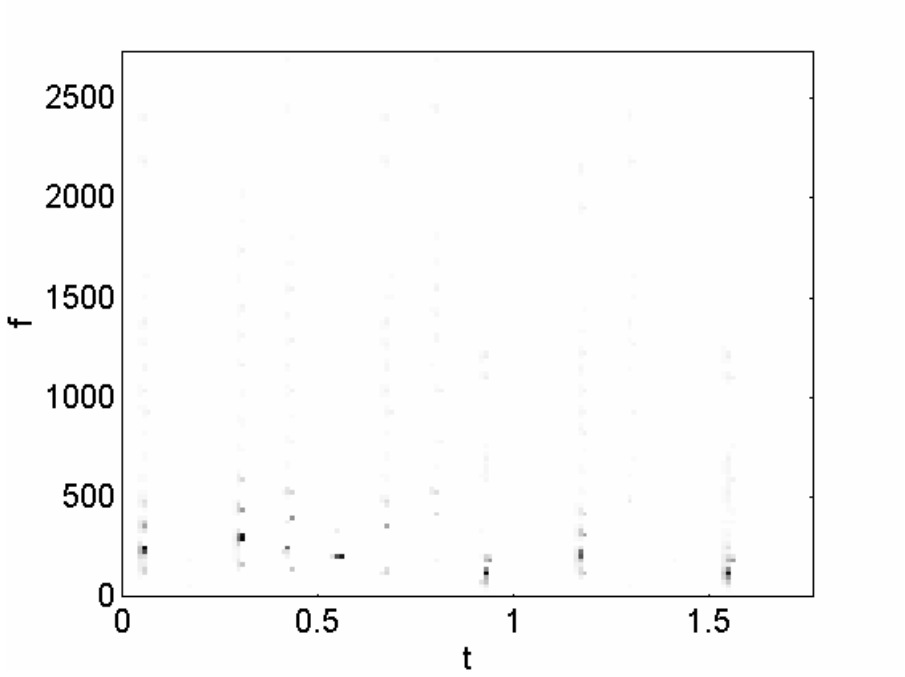


Fig. 3.12 Componentes de mayor velocidad de incremento de potencia extraídos del espectrograma de la Fig. 3.9.

La Fig. 3.12 muestra la distribución de los componentes de mayor velocidad de incremento de potencia extraídos del espectrograma correspondiente al fragmento de la Fig. 3.3. La alineación de los componentes representados en la Fig. 3.12 con los eventos de la secuencia de análisis (Fig. 3.3) evidencia la relación velocidad de cambio de la energía-percepción del evento, aunque sólo unos pocos de estos componentes corresponderán a eventos como veremos posteriormente.

27 Estos dos parámetros: *sensibilidad* y *rango de frecuencia* son utilizados por Goto y Muraoka para evaluar múltiples hipótesis. El sistema ideado por ellos cuenta con un conjunto de agentes *buscadores* de eventos en paralelo, cada uno de los cuales evalúa con sensibilidad y rango de frecuencia diferentes (15 pares de agentes donde la sensibilidad varía de 11 a 15 muestras y el rango de frecuencias de 0-11kHz hasta 6.5-11kHz). La confiabilidad o peso del tiempo de ataque es obtenida por la relación de cada valor pico con el valor de pico máximo reciente. El sistema, para la localización de los eventos utiliza los tiempos correspondientes a los picos de mayor peso o confiabilidad. El sistema propuesto por estos investigadores está concebido para operar en tiempo real y con material musical que contiene diversos instrumentos (canciones populares muestreadas desde discos compactos) y asume una signatura de tiempo de 4/4 (signatura más frecuente en el repertorio que ellos consideran). Para inferir el tipo de acento asumen también que el bombo de la batería suena normalmente en los tiempos fuertes y el high en los tiempos débiles. El ajuste del rango de frecuencia les permite, a partir de los componentes de mayor rapidez de cambio de potencia detectar la existencia de ambos elementos percutidos.

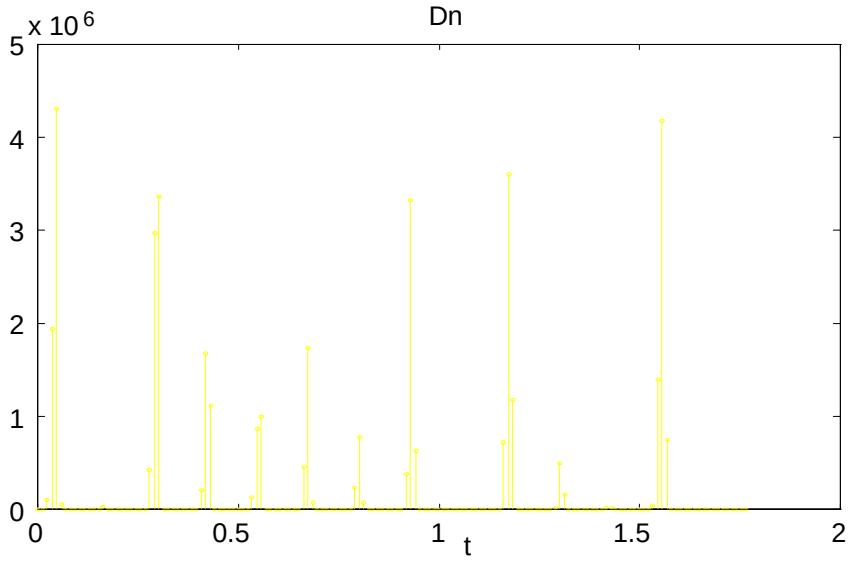


Fig. 3.13 Componentes versus tiempo.

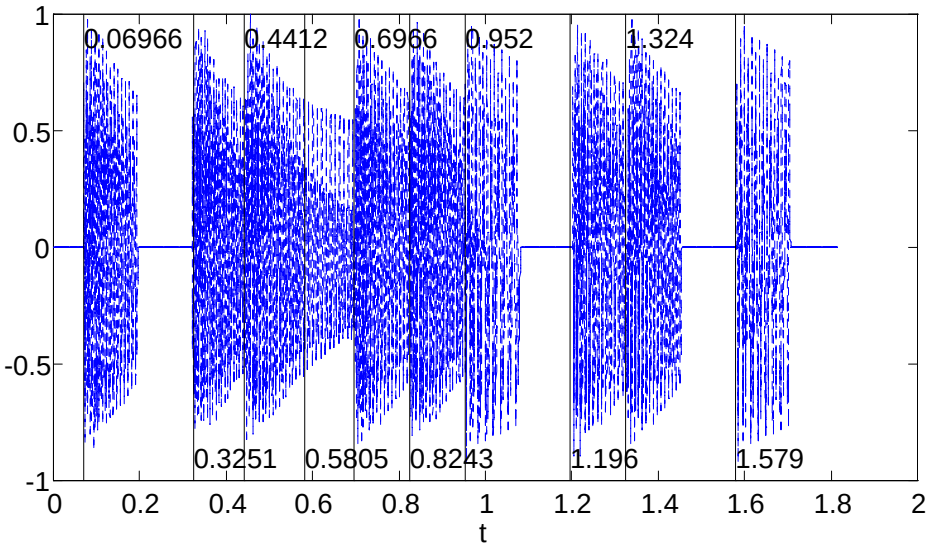


Fig. 3.14 Tiempos de ataque obtenidos para el fragmento de la Fig. 3.3 de longitud 10001 muestras con ventana de análisis hanning de 256 muestras (21.53 Hz), solapamiento de 192 muestras (75%), resolución temporal de 64 muestras (11.61 mseg), máxima sensibilidad (sin aplicar a D_n algún kernel de convolución) y con rango frecuencial máximo (17.22 Hz).

La distribución de los componentes unidimensionales prominentes en la Fig. 3.13 se corresponde con el inicio de cada evento. Si obtenemos los máximos o picos de esta localización de concentración energética por encima de un valor umbral podríamos obtener un tiempo de ataque por cada evento.

La extracción de los máximos de la secuencia D_n no es suficiente para determinar los tiempos de ataque representados en la Fig. 3.14 superpuestos al fragmento de señal analizado, porque junto a estos aparecerán otros máximos muy cercanos a cero[28] (de muy baja potencia[29]). Para ello es posible establecer un umbral de discriminación, preferiblemente sobre la secuencia D_n normalizada[30].

$$D_n = \frac{D_n}{max\{D_n\}} > \xi \tag{3.31}$$

Si observamos el comportamiento estadístico de la secuencia D_n veremos que tiene una alta concentración de valores muy cerca de cero correspondiente a los intervalos de ausencia de ataques[31], es decir una media muy cerca de cero.

Es conveniente normalizar la secuencia D_n para eliminar la dependencia del algoritmo con la señal de entrada. La elección del valor de umbral de discriminación ξ deberá eliminar la mayor parte del ruido en torno a la media con valores en el orden de 0.03 y 0.05. Para los experimentos realizados un valor $\xi = 0.05$ resultó apropiado en todos los casos. El umbral de discriminación se puede interpretar como el porciento de velocidad de incremento de potencia para el cual no consideramos un componente evento.

Esta primera barrera o umbral aún no resulta suficiente porque a pesar de eliminar gran cantidad de señal inútil deja algunos componentes de mayor nivel que no se corresponden con ataques. Un procedimiento más apropiado resulta de utilizar como valor de umbral ξ la media de la desviación absoluta.

Sea $D_n = D_0, D_1, \ldots, D_{N-1}$

$$\xi = \frac{1}{N}\sum_{n=0}^{N-1}\left|D_n - \bar{D}\right| \tag{3.32}$$

28 El método de Welch para el cálculo del espectograma garantiza valores de potencia no negativos.

29 Máximos locales de muy bajo nivel. Estos valores no son nada representativos de las características energéticas de la señal. Aparecen como ruido de fondo.

30 La discriminación *relativa* de los componentes exige normalización. En caso contrario habría que tener cierto conocimiento *a priori* del contenido espectral de estos componentes.

31 Estos intervalos de tiempo no se corresponden exáctamente con silencios musicales, sino sólo con el resto del tiempo donde no ocurren ataques, los cuales son mucho más frecuentes que los instantes de tiempo donde hay ataque.

$$\bar{D} = \frac{1}{N}\sum_{n=0}^{N-1} D_n \tag{3.33}$$

Este umbral aún no resultaría suficiente pero si mas apropiado que un valor de umbral absoluto porque depende en sí de la propia señal. Para completar la depuración de los componentes de ataque es necesario someter a la secuencia D_n a un filtro deslizante de componentes importantes similar al de la Ec. (3.4).

Este filtro "dejará pasar" aquellos componentes cuya relación de potencia con el máximo local de la ventana sobrepase determinado porciento. Es decir, se compara cada muestra D_n contra el máximo local de la ventana de longitud $L=2k+1$ MD y se eligen sólo aquellos componentes cuya relación con el máximo local sobrepase determinado umbral U. La longitud de la ventana debe ser preferiblemente mayor que la mitad del mayor nivel de intervalos entre ataques. Para todas las pruebas resultó suficiente una ventana de 200 mseg y un 10% de proporción[32].

$$MD = max\{D_{n-k},\ldots,D_{n-1},D_n,D_{n+1},\ldots,D_{n+k}\} \tag{3.34}$$

$$B_n = \begin{cases} D_n & \text{Si } \dfrac{D_n}{MD} > U \\ 0 & \text{de otra manera} \end{cases} \tag{3.35}$$

Otro estadístico que podría ser apropiado es la desviación típica pero tiene el inconveniente que al ser mayor que la media de la desviación absoluta podría eliminar los componentes de la señal de más bajo nivel[33]. Este procedimiento

32 Por debajo de esta proporción se considera ruido. Si existe algún componente de ataque con un valor inferior al umbral elegido y se escoge un valor relativamente bajo (como puede ser un 10%) será excluido del conjunto de ataques válidos B_n. Este error no es apreciable en el resultado final para el proceso subsiguiente de inducción del pulso porque como se verá éste proceso tiene en cuenta los pesos de cada ataque y los ataques de bajo nivel no aportan sensiblemente al proceso de inducción del pulso. En el conjunto de pruebas realizadas el algoritmo de segmentación eliminó sólo 2 componentes de ataque en un total de 100 ataques. Los resultados de la inducción con y sin estos componentes fueron muy similares, en cualquier caso el tiempo predecido de mayor probabilidad del próximo pulso fue en ambos casos el mismo.

33 En este caso, el algoritmo eliminó 10 ataques, pero los resultados finales en la inducción del pulso fueron los mismos. Este procedimiento tiene la ventaja de que no es necesario aplicar el filtro de máximos deslizante por lo que puede ser más eficiente desde el punto de vista computacional.

tiene la ventaja que resulta suficiente para eliminar todo el ruido y no requiere de un filtrado posterior de los componentes importantes.

Una vez obtenidos los B_n componentes importantes de ataque de la secuencia musical se determinan los instantes de tiempos correspondientes y se genera la lista de eventos. La lista de eventos contiene aquellos ataques o incrementos bruscos energéticos del material musical en dos campos: tiempo y peso relativo del componente correspondiente a una descripción precisa de los eventos musicales. Arrivar a la lista de eventos es una parte importante del trabajo pero no el resultado final. Desde éste nivel de análisis existen muchos derroteros[34], uno de los cuales será nuestro objetivo: La inducción del pulso.

0.06966	1.0000
0.32510	0.7826
0.4412	0.3887
0.5805	0.2305
0.6966	0.4035
0.8243	0.1804
0.9520	0.7703
1.1960	0.8360
1.3240	0.1160
1.5790	0.9677

Tabla. 3.1 Lista de eventos[35]

34 Hacia la abstracción de los números (pentagrama), o en la dirección análisis/síntesis, o en el control de otros instrumentos en relaciones arbitrariamente complejas en relación con la música original, etc. Éstas direcciones están interrelacionadas.

35 También conocida en la literatura por *notelist* es un listado donde cada fila se corresponde a un evento y las columnas representan diferentes propiedades inherentes a cada evento. A menudo una de las columnas representa el tiempo. El tiempo puede estar dado en intervalos entre dos eventos consecutivos o marcas como en nuestro caso (primera columna). La segunda columna indica el peso de cada evento (en términos porcentuales tanto por uno).

Acercamiento al tempo

La Tabla. 3.1 suele servir la información base para un análisis de alto nivel del material musical y sería el punto de partida a la inducción del pulso si se hubiera partido de un teclado MIDI convencional en lugar de una señal acústica. Otra de las columnas de esta lista de eventos podría ser a la altura[36]. Aunque tal propiedad no es un requerimiento para nuestro análisis de inducción del pulso no debería faltar en otros tipos de análisis como puede ser la notación occidental a partir de una secuencia musical real. Nótese que el pentagrama es una abstracción de estos números.

La columna de tiempo representa las marcas de los eventos. Si alimentáramos un diferenciador de primer orden con esta secuencia obtendríamos otra secuencia que representaría los intervalos transcurridos entre dos eventos consecutivos o *tiempos entre-ataques.* El tiempo entre-ataques es el que verdaderamente describe un ritmo dado y no la duración de los eventos. El tiempo entre ataques es una magnitud relativa, de contexto, a diferencia de la duración de los eventos, que es una magnitud absoluta. Si un evento termina antes de que comience un nuevo ataque habrá de transcurrir un período de duración parcial del evento y otro de silencio hasta la próxima marca de ataque. Ambos, la duración parcial del evento y el silencio siguiente, no son perceptualmente sobresalientes en comparación con la duración completa implicada por el intervalo de tiempo entre-ataques. De hecho se puede decir que este último define el ritmo, mientras que los anteriores la *articulación* del ritmo [Schloss,1985]. Por esta razón se enfatiza en la detección de los ataques y no en la detección de los silencios entre eventos o sus duraciones parciales.

La Tabla. 3.2 nos permite comparar ambos métodos de extracción de los tiempos de ataques: en el dominio temporal y en el dominio frecuencial. La primera columna se corresponde a los valores de tiempo extraídos por el primer método descrito que opera sobre la energía en el dominio temporal. La segunda columna se corresponde a los valores comentados anteriormente extraídos desde la densidad espectral de potencia en el dominio frecuencial.

36 El *pitch* o altura del evento está asociado normalmente a un conjunto discreto de valores de frecuencia. La estimación de la altura es objetivo de otros estudios y está muy poco relacionado con la estimación del ritmo donde son más importantes los intervalos o distancias temporales entre eventos.

La tercera columna se corresponde a ataques obtenidos manualmente y aunque son valores de escasa fiabilidad, dada la forma en que han sido obtenidos, ilustra los rangos alrededor de los valores de tiempos estimados. La última columna se corresponde a la diferencia entre las dos primeras o error relativo entre ambas.

0.0755	0.0697	1.5838	0.0001
0.3251	0.3251	0.3198	0.0000
0.4586	0.4412	0.4467	0.0002
0.5573	0.5805	0.5787	-0.0003
0.7082	0.6966	0.6980	0.0001
0.8359	0.8243	0.8325	0.0001
0.9520	0.9520	0.9594	0.0000
1.2020	1.1960	1.2030	0.0001
1.3350	1.3240	1.3299	0.0001
1.6200	1.5790	0.0685	0.0005

La tercera columna podría representar valores de tiempo exactos, extraídos o calculados desde el pentagrama conociendo que

si el tempo ♩ = 120, 𝅘𝅥𝅯 = 0.125

Entonces esta columna tendría mayor sentido

Tabla. 3.2 Análisis comparativo de los tiempos obtenidos por ambos procedimientos: en el dominio temporal y en el dominio frecuencial. Ataques obtenidos manualmente y diferencia entre los dos primeros resultados.

Llama la atención el orden del promedio de error del orden de fracción de milisegundo; intervalo perceptualmente inapreciable como ya hemos comentado en la segunda sección de este trabajo. No obstante la relativa carga computacional de este último método resulta conveniente su uso en presencia de señales coloreadas con solapamientos entre los eventos.

Otro análisis de tiempo se obtendría sobre la base de los intervalos de tiempo entre-ataques (IEA[37]) en lugar de los tiempos de ataque directamente. El fragmento de secuencia de análisis representado en la Fig. 3.3 a pesar de simular una señal acústica real, es obtenido desde un pentagrama con tiempos de ataque muy precisos derivados del metro en lugar de las inflexiones voluntarias que producen en el tempo los instrumentistas reales y se conoce *a priori* que sólo existen intervalos de determinada duración básica (e.g., semicorcheas). Este conocimiento nos permitirá comprobar si efectivamente los IEA obtenidos se corresponden a los de la señal de análisis.

37 IEA Intervalo de tiempo Entre Ataques. Los IEA no son salida directa de los métodos comentados pero pueden ser obtenidos aplicando un simple diferenciador de primer orden de la secuencia de tiempos de ataque: $t_{ea}[n] = t_a[n] - t_a[n-1]$, siendo t_a los tiempos de ataque y t_{ea} los tiempos entre-ataques.

Si observamos el histograma de los IEA generados por ambos métodos en la Fig. 3.15 notaremos que existe cierta concentración de puntos alrededor de dos valores 0.25 y 0.125, como era de esperar a partir del comentario anterior. Esta figura ilustra claramente el sesgo que produce la estimación del mayor subintervalo el método de análisis localizado de la energía que se muestra en la ventana superior. Obsérvese también que el método de análisis basado en la velocidad de incremento de la potencia tiene menor dispersión alrededor de los subintervalos múltiples del subintervalo básico (en este caso binario).

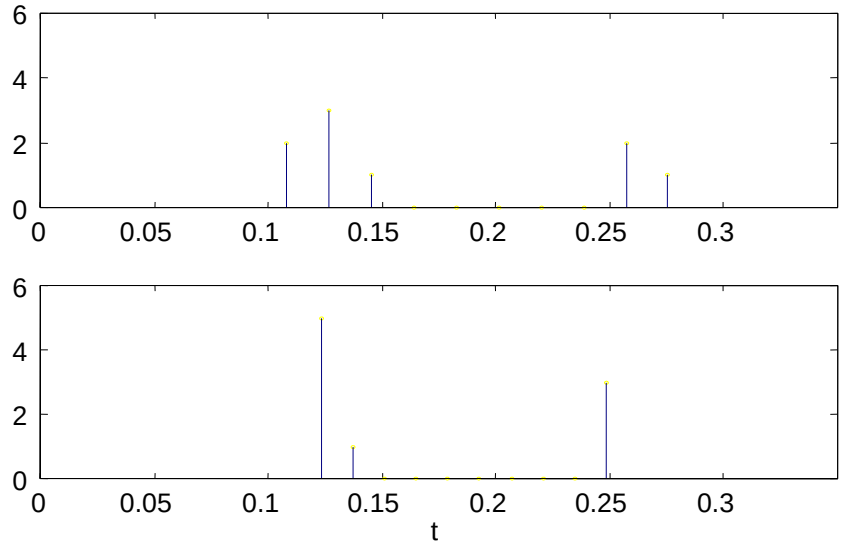


Fig. 3.15 Histogramas de los tiempos entre-ataques (IEA) obtenidos por ambos procedimientos: dominio temporal en el diagrama superior y dominio frecuencial en el diagrama inferior.

La reflexión anterior a partir del examen de la lista de eventos nos lleva a encontrar las duraciones o intervalos repetidos. En una secuencia real es de esperar mayor desviación alrededor de los valores de tiempo múltiplos del IEA básico[38].

Si deseamos obtener los subintervalos de tiempo de duración más frecuente a partir de los tiempos de ataques de la lista de eventos podemos seguir el siguiente procedimiento. Primero, convertimos los tiempos de ataques en intervalos entre-ataques aproximando a la primera derivada.

$$t_{ea}[n] = t_a[n] - t_a[n-1] \tag{3.36}$$

38 Debe existir cierta tolerancia para aceptar las fluctuaciones intencionales alrededor de los múltiplos del subintervalo de tiempo entre-ataques básico.

luego calculamos el histograma[39]

$$\begin{aligned} &min\{t_{ea}[n]\} \le m \le max\{t_{ea}[n]\} \\ &Hist[m] = 0, \quad \forall m \\ &Hist[m] = Hist[m] + 1 \end{aligned} \tag{3.37}$$

y por último, buscamos los máximos locales o picos predominantes a partir del histograma o de una versión del histograma suavizada[40]. Los picos se corresponden con aquellos valores donde ocurre un cambio de pendiente.

$$p_n = H_n > H_{n\pm1} \tag{3.38}$$

Los IEA que se correspondan con los valores máximos del histograma de la secuencia p_n corresponden a los subintervalos métricos que Schloss denomina *duraciones importantes*. Goto y Muraoka inducen el intervalo entre pulsos (IEP) a partir del máximo de los IEA máximos, es decir, el intervalo más frecuente entre eventos o duración más importante.

Para nuestro caso de prueba se obtuvieron los siguientes resultados

IEA	0.1231	0.2485
IEP	0.1231	

El IEP estimado inspeccionando manualmente la señal fue de 0.1267[41]. Obsérvese que, aunque 0.2485 no es exactamente el doble de 0.1231 el error cometido es de aproximadamente 2 mseg. Este error se puede disminuir aumentando el intervalo de muestreo en la dimensión temporal y la resolución temporal, es decir, aumentando el solapamiento y disminuyendo la longitud de

39 El histograma cuenta la frecuencia de ocurrencias de cada intervalo de tiempo entre-ataques obtenido (en este caso). Representa la función densidad de probabilidad de los intervalos de tiempo entre-ataques.

40 En este punto el criterio a aplicar al histograma depende de lo que se pretenda. Goto y Muraoka utilizan la técnica del histograma para detectar, de los componentes $d[n,k]$ extraídos, cuales corresponden al bombo (BD) y cuales al high (SD) y lo afectan por el grado de los componentes. Schloss, en cambio, calcula el histograma a partir de los intervalos entre eventos. Primero, utiliza una escala logarítmica para las tales duraciones, porque son más importantes en el dominio rítmico los ratios que las diferencias. Segundo, afecta cada valor del histograma por su raíz cuadrada, para enfatizar las duraciones largas (no es deseable que las notas cortas dominen el histograma) y por último, suaviza los histogramas para que cada duración contribuya a la célula vecina, en una especie de convolución de los datos con una curva Gaussiana.

41 Probablemente correspondiente a 125 mseg teniendo en cuenta que el metro es derivado directamente del pentagrama. Como ♩ = 120 ≡ 1 seg, 𝅘𝅥𝅯 ≡ 0.125 seg.

la ventana de análisis entre ventanas contiguas. Aplicando el mismo procedimiento con una ventana de 5 ms vs 11.61 ms empleada anteriormente a una pieza de señal acústica de aproximadamente el doble de duración del fragmento de la Fig. 3.3 que además lo incluye se obtuvo

IEA 0.1244 0.2478
IEP 0.1244

El error obtenido se reduce a 1ms a costa de aumentar la carga computacional.

Para probar la discriminabilidad temporal del detector de eventos basado en el dominio frecuencial podemos aplicarlo al caso de prueba descrito en la sección *discriminabilidad límite*. La Tabla. 3.3 muestra los resultados obtenidos que en un gráfico similar al de la Fig. 3.8 estos valores se corresponderían con las marcas de tiempo.

0.246
0.376
0.621
0.751

Tabla. 3.3 Tiempos de ataque obtenidos en un caso de prueba de discriminabilidad límite.

Obsérvese que si restamos entre sí los tiempos entre dos ataques consecutivos obtenemos:

0.376 - 0.246 = 0.13 s
0.751 - 0.621 = 0.13 s

Es decir, los 130 ms de separación entre símbolos dos a dos[42] lo que demuestra la capacidad del procedimiento de discriminar eventos separados entre sí 5 ms. Si además restamos el intervalo entre los tiempos de ataque correspondientes a la señal base a partir de la cual se generó el caso de prueba:

0.621 - 0.246 = 0.375 s

42 El procedimiento de síntesis del caso de prueba se comenta detalladamente en la sección *discriminabilidad límite.*

0.751 - 0.376 = 0.375 s

lo cual indica que no se produce sesgo en las estimaciones. Obsérvese que los 375 ms son el resultado de sumar 125 ms de duración del evento (longitud del símbolo) con 250 ms correspondientes a dos intervalos de silencio. El algoritmo se aplicó con una ventana hanning de 40 muestras (100 Hz), un solapamiento del 90% (36 muestras), una resolución temporal de 4 muestras (1 ms) y un umbral del 30%. Obsérvese que para aumentar la resolución temporal es necesario disminuir la resolución frecuencial.

Consideraciones finales

La estimación de los IEA a partir de la localización de los tiempos de ataques de los eventos en el dominio espectral tiene la ventaja de que, independientemente de sus mejores prestaciones, al operar con el algoritmo de FFT no requiere de una alta carga computacional, lo que le permite implementación inmediata en tiempo real. Pero la transformada de Fourier tiene ciertas características muy criticadas por algunos autores para el tratamiento de señales no estacionarias con relación a sus propiedades de muestreo y resolución.

La transformada de Fourier localizada depende de la ventana de análisis seleccionada. Dada una versión enventanada alrededor del tiempo n, se calculan todas las “frecuencias” de la transformada de Fourier dependiente del tiempo. Lo mas importante es que, una vez seleccionada la ventana para obtener la transformación de Fourier dependiente del tiempo, queda *fija* la resolución tiempo-frecuencia dada por las Ecs (3.12) y (3.13). sobre todo el plano tiempo-frecuencia (porque se utiliza la misma ventana para todas las frecuencias). Por ejemplo, si una señal está compuesta de pequeñas discontinuidades asociadas con grandes componentes cuasi-estacionarios, entonces se puede analizar cada tipo de componente con buena resolución de tiempo o frecuencia, pero no de ambas.

Moorer, por ejemplo, no recomienda el empleo de la FFT porque dice, son falseadas por cualquier cambio debido ya sea a la amplitud o a la frecuencia, y por lo tanto ofrecen resultados engañosos en presencia de vibrato o reverberación de sala [Moorer,1975]. Tampoco recomienda dos de las técnicas de análisis de voz: el *cepstrum* (la DFT inversa del logaritmo del valor absoluto

de la DFT de la entrada) y el *predictor lineal*, por sus claros problemas al tratar la polifonía. Moorer, en cambio, en su disertación sobre el timbre musical llamada “Una Exploración del Timbre Musical” describe el filtro heterodino utilizado para producir los diagramas de John Grey [Grey,1975]. El filtro heterodino es básicamente una adaptación de la DFT; su problema es que es demasiado sensible a las variaciones de la frecuencia y de la amplitud (presente en la mayoría de las señales musicales), y si los componentes parciales caen fuera de un canal, se dificulta extraordinariamente la interpretación de los componentes. Así, el filtro heterodino descrito por Moorer tiene mayor utilidad en el análisis de tonos singulares aislados de diversos instrumentos musicales[43], pero no para música contínua o situaciones mas complejas.

De los métodos usados en la investigación de Moorer, el primero, aplicado a los datos es una técnica relacionada a la función de autocorrelación, que Moorer llama el *“peine óptimo”.* Por cada dato x_n en la forma de onda de M muestras, forma la suma

$$\sum_{i=0}^{M-1} \left| x_{n+i} - x_{n+i-m} \right| \tag{3.39}$$

y busca el mínimo sobre m en esta suma. El peine óptimo de Moorer es realmente un preprocesador utilizado para obtener la periodicidad básica, o *“armonía”* de los datos, principalmente para reducir la magnitud de procesamiento masivo. La idea es que, mediante este método, se puede hallar un mínimo común divisor de armónicos, que ayude a establecer posteriormente las frecuencias centrales paso-bandas. El objetivo de Moorer es la notación musical y simplifica el análisis del tempo asumiéndolo constante.

Piszczalski y Galler, sin embargo, basan sus trabajos sobre transcripción automática en FFTs sucesivas de la señal (transformada de Fourier localizada) [Piszczalski y Galler,1979]. Su acercamiento es básicamente pragmático y tiene poco que ver con teorías perceptuales, sin embargo, introducen cierta *inteligencia* al considerar la combinación de los armónicos para producir un

43 James Moorer es un pionero en el intento de transcripción polifónico, a pesar de restringir la entrada a sólo dos voces simultáneas y prohibir unísonos, octavas, docenas y algunos otros intervalos porque los armónicos de dos instrumentos que producen estos intervalos se solapan, complicando enormemente la desambigüedad de estos casos.

tono contínuo; por ejemplo, si la fundamental tiene un descenso en amplitud pero el segundo armónico es contínuo, entonces no debe haber segmentación en el punto donde la fundamental cae[44].

La notación se realiza con la información de la altura y el tiempo, sin embargo, ellos simplifican el tempo como una constante y evitan considerar sus fluctuaciones (una situación muy improbable en una interpretación real).

Stautner, intentó desarrollar e implementar un método de análisis más cercano al proceso auditivo natural y crea gráficos 3D cuidadosamente sintonizados con el logaritmo del valor absoluto de la transformada de Fourier localizada a la que denomina *transformada auditiva* [Stautner,1983]. Estos gráficos, visualmente informativos muestran de que modo el sonido puede ser resintetizado a partir del gráfico, pero para ello debe extenderse el análisis en un procedimiento muy similar al *cepstrum* aunque Stautner no le refiere como tal.

Una variante alternativa al empleo de las FFTs en el análisis de señales no-estacionarias o cuasi-estacionarias podría ser el empleo de la *Transformada de Ondículas* [Rioul y Vetterli,1991]. En contraste a la transformada de Fourier localizada, que utiliza una sola ventana de análisis, la transformada de ondículas utiliza ventanas cortas a altas frecuencias y ventanas largas a bajas frecuencias. La transformada de ondículas resuelve la limitación de la transformada de Fourier localizada porque permite variar la resolución Δt y Δf en el plano tiempo-frecuencia a fin de permitir un análisis multi-resolución. Intuitivamente, cuando se realiza el análisis como un banco de filtros, la resolución temporal debe incrementar con la frecuencia central de los filtros de análisis.

El banco de filtro de análisis está compuesto por filtros paso-banda con ancho de banda relativo constante (también llamado análisis “Q-constante”). Si imponemos que Δf sea proporcional a f ó

$$\frac{\Delta f}{f} = c \tag{3.40}$$

44 O sea, no habría un ataque nuevo en el punto donde la fundamental aumenta a su amplitud original; el programa debe mirar justo la fundamental cuando trata de hallar nuevos ataques (segmentación).

donde c es una constante, Δf y por lo tanto también Δt cambian con la frecuencia central del filtro de análisis. Por supuesto, se satisface la desigualdad de Heisenberg (3.14), pero ahora, la resolución temporal puede ser arbitrariamente buena a altas frecuencias, mientras que la resolución frecuencial puede ser arbitrariamente buena a bajas frecuencias. Este tipo de análisis trabaja mejor si la señal está compuesta de componentes de alta frecuencia de corta duración y componentes de baja frecuencia de larga duración, lo cual es caso muy frecuente en la práctica.

Este trabajo no pretende profundizar en cualquiera de estas alternativas, ni revisar el extenso volumen de investigaciones al respecto pero se comentan como formas alternativas al mismo propósito.

Capítulo 4 Inducción

Introducción

El proceso de inducción del pulso estima, con cierta certidumbre, el próximo tiempo de ataque correspondiente a un acento rítmico importante a partir del *contexto* disponible hasta ese momento. El contexto o *patrón temporal* es un subconjunto de la lista de eventos que agrupa los eventos disponibles en un intervalo dado. La inducción del pulso corresponde a un análisis de alto nivel. Este trabajo propone una formulación matemática de la teoría de descomposición[1] de la percepción del ritmo descrita por Desain en 1992 (ver [Desain,1992] para una descripción más detallada de la teoría) y su posible utilización como modelo de inducción acentual.

El atractivo de esta teoría consiste en la posibilidad de descomposición de la compleja conducta del modelo de inducción del pulso en partes significativas más pequeñas y para ello se apoya en la *esperanza de los eventos*. La distribución de la percepción del ritmo en componentes individuales, de cuya interacción resulta posible una predicción de comportamiento futuro, intenta modelar el fenómeno de la percepción del ritmo y defiende la composicionalidad como un requisito para un modelado cognoscitivo satisfactorio [Chandrasekaran,1990][2].

1 Desain, en su artículo “Una teoría (des)componible de percepción del ritmo” [Desain,1992], *descompone* éste término en dos acepciones probablemente para significar el principio y la reversibilidad de la construcción teórica.

2 El conexionismo puede ser un paradigma atractivo en la búsqueda de una base teórica común pero falla en cuanto a composicionabilidad. Esto significa que aunque el modelo como un todo monolítico pueda funcionar bien, es imposible descomponer su comportamiento complejo en pequeñas partes significativas. El propio Desain y Honing [Desain y Honing,1989] trabajaron sobre un modelo subsimbólico (conexionista) de cuantización temporal y un año después Desain publicó una disertación sobre los modelos simbólicos versus subsimbólicos en la percepción del ritmo que concluye con una abstracción del comportamiento del cuantizador en la forma de *esperanza de eventos* con un patrón temporal como contexto a priori [Desain,1990].

El concepto de esperanza parece explicar la dependencia de la percepción de la estructura rítmica sobre el tempo global, la influencia del contexto sobre la percepción categórica[3] y otros fenómenos complejos. La esperanza es descomponible y esta propiedad hace posible basar la teoría de percepción de estímulos complejos sobre un modelo simple de percepción de sus componentes constituyentes. Estas posibilidades animaron a Desain a proponer el método de la esperanza como base común para teorizar acerca de la percepción temporal y la memoria.

El modelo

El modelo consiste de una compleja red de desplazamiento de longitud N que recibe los eventos generados por el segmentador de bajo nivel, los procesa y produce a su salida un evento cuyo tiempo de ataque se corresponde con el tiempo de máxima esperanza de la llegada del próximo *evento importante* en el futuro.

La red está formada por N nodos interconectados secuencialmente que determinan el contexto o patrón temporal. Cada nodo almacena un evento (grado y tiempo de ataque, en nuestro caso). La llegada de cada nuevo evento provoca el desplazamiento de toda la red un lugar[4]. El modelo también cuenta con módulos que generan la esperanza básica de cada intervalo implícito que se forma en la red[5] determinado por el número de nodos de la red y dado por

$$M = \sum 1 + 2 + \ldots + N \tag{4.1}$$

3 Si consideramos el tiempo musical como el producto de dos escalas de tiempo: los intervalos de tiempo discreto de la estructura métrica y los cambios de tiempo continuos o expresivos del tiempo [Clarke,1987] podemos decir que la percepción de los intervalos de tiempo sobre una escala discreta es obligatoriamente un proceso automático. En contraste la percepción y reproducción de tiempo continuo en una interpretación musical está asociada con un comportamiento experto.

4 Los nodos son fijos y son los eventos quienes se mueven hacia el nodo vecino con la llegada de cada nuevo evento. El registro de desplazamiento funciona como sus homólogos electrónicos sólo que el reloj que fuerza el desplazamiento en este caso es irregular. Los desplazamientos no están determinados directamente por un reloj sino por los instantes donde se producen los ataques. La dirección de desplazamiento desde el punto de vista de la teoría no es importante, pero eso se analizará más adelante.

5 Un intervalo está compuesto por la diferencia entre dos tiempos de ataque. El número de intervalos implícitos en el contexto está determinado por todas las combinaciones posibles irrepetibles entre los tiempos de ataques dado por la Ec. (4.1) (Los pesos están normalizados).

La Fig. 4.1 muestra las curvas de esperanza básicas generadas por el patrón temporal [0 0.5 0.75]. Este contexto corresponde a la salida del nivel más bajo y representa tres tiempos de ataques[6] ó dos intervalos entre-ataques (N=2 [0/0.5 0.5/0.75]) y da lugar a tres posibles intervalos (M = 1+2 [0/0.5 0.5/0.75 0/0.75]).

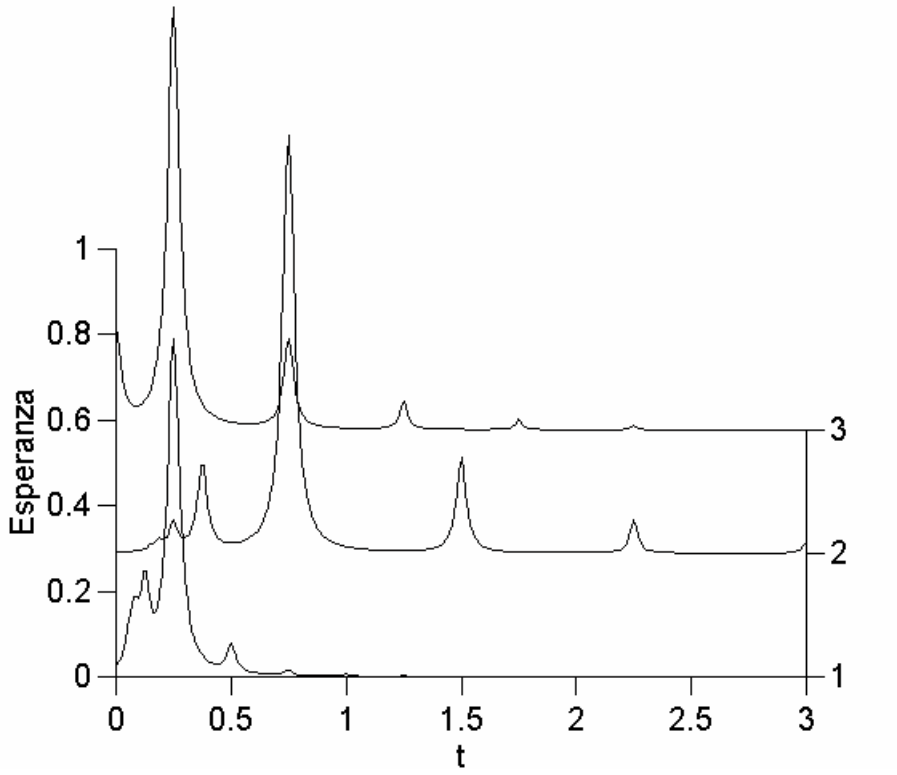


Fig. 4.1 Curvas de esperanza básicas generadas por el contexto [0 0.5 0.75].

En el ejemplo de la Fig. 4.1 se muestran las curvas de esperanza básicas generadas por cada intervalo implícito en el patrón. La curva inferior (con un 1 a la derecha[7]) corresponde al intervalo 0.5/0.75, la del medio (2) al intervalo 0/0.75 y la superior (3) al intervalo 0/0.5[8]. La generación de estas curvas y su formulación matemática se expondrá exhaustivamente más adelante pero obsérvese como los picos de las curvas se producen justo en aquellos instantes de tiempo correspondientes a divisores o múltiplos enteros de la longitud de cada intervalo de análisis.

6 En este caso todos los eventos tienen la misma contribución igual a 1.
7 El eje vertical de la derecha corresponde a todas las posibles combinaciones diferentes que forman los intervalos que conforman un contexto.
8 El eje de tiempos es relativo a la llegada del último evento (0.75). El instante 0 corresponde al instante 0.75.

Según el modelo y apoyándonos en el ejemplo hasta el instante 0.75 nuestra red tendría 3 casillas del registro ocupadas (deben de haberse producido 2 desplazamientos) y tres módulos de cálculo de esperanza básica activos. Existe aún otro elemento que debe sumar las aportaciones de cada intervalo para estimar el comportamiento en el futuro, el acumulador de esperanzas básicas. La arquitectura descrita sólo intenta descomponer funcionalmente el modelo en hipotéticas partes a fin de lograr una mejor comprensión. Efectivamente esta arquitectura es naturalmente distribuida pero el alcance de este trabajo no llega a una propuesta de implementación.

La esperanza

La esperanza de un evento está relacionada con la proyección de un componente de la lista de eventos hacia el futuro o hacia el pasado y se basa en el hecho de que estos cambios están completamente determinados por el contexto o patrón temporal acumulado[9].

La esperanza es una curva compleja que determina los instantes de tiempos que constituyan las *mejores continuaciones* del patrón temporal acumulado. A la diferencia entre los dos máximos de estos picos corresponderá el tiempo donde con mayor probabilidad se espera la llegada del próximo evento importante. La curva compleja de la esperanza es descomponible en las curvas de *esperanza básicas* aportadas por cada evento o visto de otra manera está formada por la suma de todas las aportaciones de esperanza básica de los distintos intervalos entre-ataques en la red. Esto efectivamente descompone ritmos de comportamiento complejo en un conjunto de componentes simples, uno para cada intervalo de tiempo implícito en el patrón.

9 La lista de eventos presentada contiene los ataques o componentes de todo el material musical con el propósito de ilustrar la generalización del método de segmentación propuesto pero en una implementación en tiempo real donde se pretenda inducir el pulso este análisis se deberá localizar y los tiempos de ataques no formarán una lista única general en un instante único sino que serán proporcionados al análisis de alto nivel según sean detectados de forma secuencial. El contexto está determinado por los intervalos de ataque disponibles en un instante dado. A pesar de la componente discreta del tempo musical, la componente continua provoca ciertos movimientos del tempo e inclusive de la estacionariedad de la propia ritmicidad. En la medida en que se disponga de un contexto o entorno abundante se podrá inducir el comportamiento futuro con mayor seguridad pero la acumulación excesiva puede conllevar a la deshomogeneización del contexto.

La esperanza básica es una función de dos parámetros: la longitud del intervalo de tiempo entre ataques y el tiempo transcurrido hasta el instante de llegada del evento. Es decir, un tiempo relativo que depende de la distancia temporal entre los distintos eventos y un tiempo absoluto cuyo origen se puede relativizar o bien respecto de un origen hipotético inicial[10] o bien respecto a cada nuevo evento que llegue[11]. Cada intervalo entre ataques genera una curva con máximos donde el tiempo transcurrido (o por transcurrir) es un divisor entero o un múltiplo entero de dicho intervalo. Estos tiempos se corresponden con las proporciones rítmicas relevantes. Un intervalo de A segundos de duración deberá generar picos de mayor esperanza en aquellos instantes de tiempo futuros Ar siendo r

$$r = \left\{ \frac{1}{Q}, \ldots, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, \ldots, Q \right\} \tag{4.2}$$

donde Q es el rango ó escala a tener en cuenta para la construcción de la curva de esperanza. r es el conjunto de todas las proporciones rítmicas relevantes. La estructura rítmica es un árbol binario (a veces ternario) [Longuet-Higgins,1976] cuyos terminales son una nota o un silencio. Tal árbol, refleja de forma clara, la estructura de la notación musical occidental. Esta idea básica del metro como gramática generativa demanda que el escucha cree su árbol mientras experimenta la música. El oyente debe identificar el metro que genera este ritmo, y representarlo posteriormente como una estructura de árbol que acomode todas las notas y silencios como símbolos terminales[12]. El rango Q nos servirá para establecer la granularidad de los terminales del árbol (ya sean binarios o ternarios) y paralelamente a ello el nivel rítmico[13].

La escala Q es el soporte de descomposición de un intervalo dado en las duraciones musicales más frecuentes o proporciones rítmicas relevantes. Supongamos que el intervalo A corresponde a la duración de un cuarto-de-nota

10 Por ejemplo desde el propio comienzo de la interpretación de la secuencia musical.

11 Dá igual decir que se espera un evento a los 6 segundos de iniciada la “escucha” o cuando expiren 2 segundos a partir de ese momento. Es sólo una cuestión de relativizar el origen de tiempos.

12 Categorización del ritmo percibido.

13 El nivel rítmico es una generalización de la idea del compás. Se puede pensar en una unidad de medida o compás como un pulso regular en la música, que ocurre en un período y desplazamiento particular; otros niveles rítmicos tienen períodos que son múltiplos o fracciones del nivel de la unidad de medida de referencia o compás, para que podamos hablar de nivel de mitad-de-compás, nivel de cuarto-de-nota, nivel de octavo-de-nota, nivel de doble-compás, etc.

o negra. $A/2$ se corresponde a una corchea, $2A$ a una blanca, $A/3$ a una corchea de tresillo, $3A$ a una blanca con puntillo, $A/4$ a una fusa, $4A$ a una redonda y así sucesivamente. Q determina el grano de la predicción, el nivel ó rango de medida de la duración musical a la que queremos ajustar la predicción y Ar el conjunto de todas las proporciones relevantes proyectadas sobre un tiempo T.

La Fig. 4.2 muestra la curva de esperanza básica para el intervalo de tiempo $[0\ A]$. Véase que la curva de esperanza tiene un pico máximo en A (r=1) y picos de nivel descendiente en cada Ar instante de tiempo de la escala de tiempo absoluta T.

La forma de la curva de la esperanza está determinada por la forma en que somos capaces de percibir proporciones de duración serie, con mayor proporción cuanto más difícil y menos esperada sea [Jones y Boltz,1989]. Esto también es cierto para las proporciones complejas en términos de sus divisores primos.

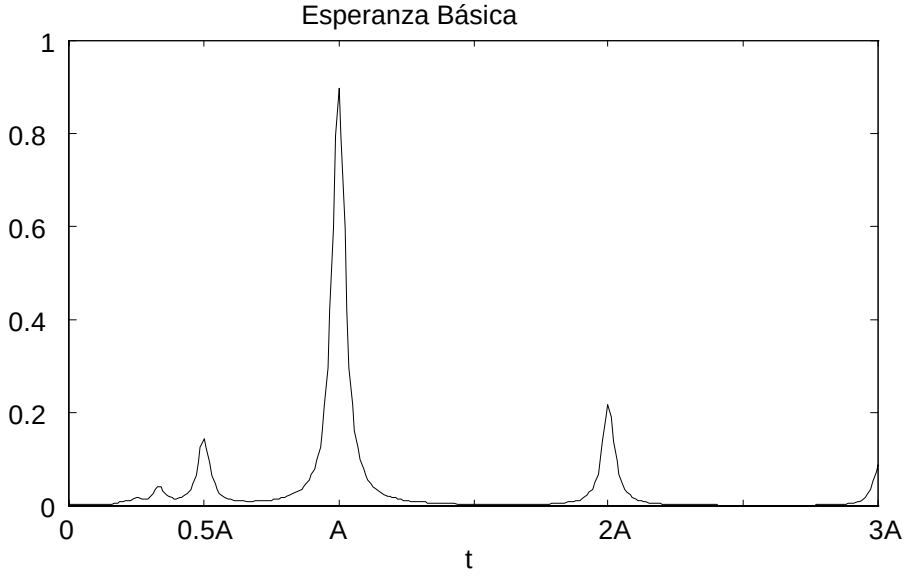


Fig. 4.2 Curva de esperanza básica generada por el intervalo $[0\ A]$.

Existen numerosas investigaciones y enfoques sobre los mecanismos de activación perceptuales respecto al ritmo y el tempo. La esperanza básica se aproximará a tal comportamiento en la medida en que simule con mayor naturalidad las hipótesis comprobadas y aquellas otras de valor heurístico: El juicio perceptual es dependiente de la proporción[14] de los intervalos

14 r es el conjunto de todas las proporciones involucradas en el cálculo de la esperanza y su alcance está determinado por Q.

involucrados y de su duración de tiempo absoluta [Sternberg,Knoll y Zukofsky,1982]. Los intervalos de tiempo muy cortos y muy largos son más difíciles de percibir con precisión. El máximo de sensibilidad ocurre alrededor de los 600 mseg y constituye el rango de *tempo preferido* [Fraisse,1982]. La longitud total del intervalo (par de tiempos de ataque) será utilizada como un segundo determinante de la esperanza para modelar esta dependencia sobre la escala de tiempo absoluta.

La esperanza básica descompone cada intervalo del contexto en r curvas cuyo valor máximo está centrado en la propia longitud del intervalo (r=1) y siguen una distribución de picos con máximos locales en las proporciones rítmicas relevantes cuya amplitud y ancho depende de la relación intervalo de análisis-tempo preferido[15], de las proporciones relevantes del intervalo de análisis con el tiempo absoluto Ar[16] y de la distribución que resulta de asignar mayor peso a las divisiones para intervalos de tiempo altos y a las subdivisiones para intervalos de tiempo bajos.

La distribución de los picos correspondientes a las proporciones rítmicas relevantes esta determinada por una función f simétrica respecto a r con máximo en $Ar{=}A$ y decreciente con determinada pendiente según $Ar{\to}A/Q$ y $Ar \to AQ$.

$$\begin{aligned} f(A) &= 1 \qquad , r = 1 \;\; (Ar = A) \\ f\left(\frac{A}{Q}\right) &= f(AQ) \quad , \forall r \neq 1 \;\; (Ar \neq A) \end{aligned} \tag{4.3}$$

$$r = \left\{ |R| + 1 \right\}^{\operatorname{sgn}(R)} \tag{4.4}$$

$$R = \left\{ -Q, -Q+1, -Q+2, \ldots, 0, .., Q-2, Q-1, Q \right\} \tag{4.5}$$

La definición de la Ec. (4.3) sugiere el empleo de una función que produzca un máximo centrado en A y valores decrecientes simétricos según r se aleje de 1 ($r{\to}|Q|$). La Ec. (4.4) que descompone los valores de r en función de Q

15 La relación del intervalo de análisis y el tempo preferido establecen una medida de sensibilidad.

16 La amplitud decrece según r se aleje de 1 con determinada pendiente de caída que caracteriza el grado de importancia concedido a los múltiplos y divisores enteros del intervalo de análisis.

induce a una escala simétrica respecto al origen que nos permita evaluar por igual múltiplos y divisores enteros[17].

El método para el cálculo de la esperanza básica también se puede analizar como una descomposición de cada intervalo implícito del contexto en un conjunto de r deltas, la convolución de cada una de estas deltas con una función con forma de campana[18] y la suma de todas las convoluciones. La amplitud de las deltas y el ancho de la función campana están determinados por funciones que tratan de simular el comportamiento perceptual humano.

$$E_B(A,B) = \sum_{R=-Q}^{Q} \Psi(\alpha, \beta, A, B, r, t) \tag{4.6}$$

Los valores del índice R del sumatorio están en función de Q y definidos por la Ec. (4.5). Ψ es una función compleja que depende de los siguientes parámetros y funciones: A es el intervalo de análisis, B es el peso asociado al ataque que determina el intervalo, r determina las proporciones rítmicas relevantes y está definido por la Ec. (4.4), t es la coordenada temporal y determina la longitud y resolución del intervalo de predicción, α y β son funciones que dependen de las proporciones rítmicas relevantes, la sensibilidad y de la longitud del intervalo de análisis A y determinan la altura y el ancho de la campana respectivamente. La sensibilidad es una constante que depende de la relación entre la longitud del intervalo de análisis A y el tempo preferido.

$$\Psi(\alpha, \beta, A, B, r, t) = \alpha B \frac{\beta}{\left|t - Ar\right|^2 + \beta}\Bigg|_{t=t_i}^{t_f} \tag{4.7}$$

t_i y t_f representan el intervalo de tiempo de predicción del comportamiento (tiempo inicial y final respectivamente). Se representan como parámetros de la Ec. (4.7) para el cálculo de los *r*ésimos componentes de la esperanza básica porque una vez definida la longitud del intervalo de predicción esta es constante para la generación de las curvas de esperanza en todo el contexto.

17 Donde un número y su inverso produzcan los mismos resultados (0.5 y 2, 0.33 y 3, 0.25 y 4, etc.).

18 Desain propone en su teoría el empleo de una función Gaussiana.

Esta función es una campana centrada en la *r*ésima proporción rítmica relevante referida al tempo absoluto, cuya altura y ancho dependen de varias funciones descritas a continuación. El trabajar con intervalos y subintervalos lleva implícito el tratamiento con múltiplos y divisores enteros cuyas aportaciones dado el caso deben ser consideradas por igual y conviene el empleo de ciertas funciones normalizadoras al respecto.

A continuación se define una de estas funciones basada en la propiedad de los logaritmos neperianos: $\log x = -\log 1/x$. La función f está definida por la Ec. (4.8) y la Fig. 4.3 muestra una representación gráfica de la misma

$$f(x) = \frac{a}{\log^2 x + a} \tag{4.8}$$

Véase que la representación de la Fig. 4.3 tiene un máximo en $r=1$ y cumple las condiciones de la Ec. (4.3) para $A=1$. El coeficiente a permite regular la pendiente de caída, si a disminuye la pendiente se hace más abrupta (el pico se estrecha), si a aumenta la pendiente disminuye (el pico se ensancha).

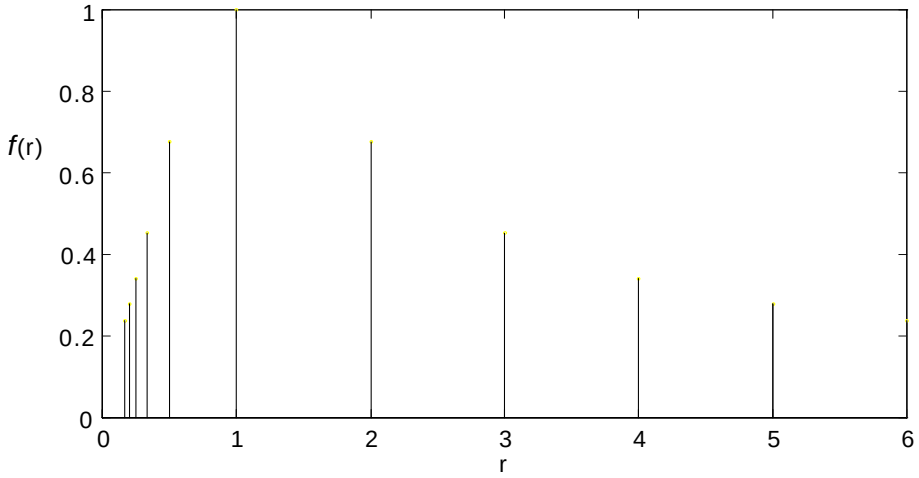


Fig. 4.3 Curva de distribución de pesos $f(r)$ para $Q=5$ y $a=1$.

La función α de la Ec. (4.7) promedia el resultado de un par de funciones, la primera corresponde a evaluar la Ec. (4.8) en r y permite una distribución de los *r*ésimos pesos en orden decreciente según $r \rightarrow |Q|$ (se aleje de 1).

$$\gamma(r) = \frac{0.1}{\log^2 r + 0.1} \tag{4.9}$$

La función definida por la Ec. (4.9) simula el comportamiento perceptual del aumento de la dificultad de percepción precisa de intervalos muy grandes y muy cortos por la distribución que consigue de los pesos sobre cada proporción rítmica relevante en la forma representada en la Fig. 4.3.

Existe un parámetro de sensibilidad que evalúa la relación del intervalo de análisis A con el tempo preferido T_p y que afecta toda la distribución de alturas de cada esperanza básica beneficiando la de aquellos intervalos próximos a T_p.

$$\zeta = \frac{1}{\log^2\left(\frac{A}{T_p}\right)+1} \tag{4.10}$$

Otro comportamiento perceptual que se debe simular es aquel que diferencia la distribución de los pesos sobre los múltiplos ($r \rightarrow Q$) y divisores ($r \rightarrow$-Q) del intervalo respecto a la magnitud de su longitud. Para intervalos de tiempo grandes los picos son relativamente altos para las divisiones y bajos para las subdivisiones y viceversa. Aunque las referencias consultadas no especifican como considerar esta apreciación de la longitud de un intervalo, podría pensarse en tomar como referencia el tempo preferido que distribuye la sensibilidad con máximo alrededor de los 600 mseg. Para definir esta nueva distribución además de la sensibilidad será necesario evaluar la relación del intervalo de análisis respecto al tempo preferido, si está por encima o por debajo de éste.

$$\lambda = \operatorname{sgn}\left\{-\log\frac{T_p}{A}\right\}\zeta \tag{4.11}$$

El valor del parámetro λ se corresponde al de la sensibilidad ζ pero su signo indica si $A<T_p$ (negativo) o $A>T_p$ (positivo). La distribución definida por la Ec. (4.12) en función de la sensibilidad, magnitud del intervalo y las proporciones rítmicas relevantes satisface este comportamiento.

$$\rho(\lambda, R) = \begin{cases} 1 & , R = 0 \\ \left| \dfrac{1}{e^{\lambda R} - e^{-\frac{R}{\lambda}}} \right| & , R \neq 0 \end{cases} \tag{4.12}$$

Por último se define la función α como el promedio de las distribuciones definidas en las Ecs. (4.9) y (4.12).

$$\alpha = \frac{\gamma + \rho}{2} \tag{4.13}$$

La función definida por la Ec. (4.13) permite una distribución para la altura de las campanas centradas en las proporciones rítmicas relevantes[19] aproximada al comportamiento perceptual descrito en las referencias consultadas y a pesar los resultados obtenidos los parámetros son susceptibles a variaciones a fin de una mejor sintonización. La esperanza básica, aunque no está definida en términos probabilísticos, produce como resultado el instante de tiempo donde se supone tenga mayor posibilidad de llegar el próximo evento relevante o pulso. El ancho del componente de pico máximo debe ser por lo tanto mayor que para el resto de las campanas componentes de la curva compleja de esperanza básica. Desain plantea que el área del pico más alto de la esperanza[20] es interpretada como la ventana en la cual se espera el próximo pulso [Desain,1994].

$$\beta(r,\zeta) = \sigma \frac{\zeta}{\log^2 r + \zeta} \tag{4.14}$$

La constante σ es un parámetro de ajuste global independiente.

Véase en la Ec. (4.7) para el cálculo de la esperanza básica, la sensibilidad de distinción entre eventos de diferentes pesos asociados B. Esto permite otorgar más importancia en el aporte al comportamiento futuro del tempo a los ataques de mayor contribución y modela una estructura más persistente. Esta característica de discriminación por contribución introduce factores perceptuales estudiados por Riemann: “Así como la esencia del elemento melódico-armónico es el cambio de altura, la esencia del elemento rítmico-métrico es el cambio de la energía viva, de la intensidad de los tonos (dinámica) por una parte, y la rapidez de la sucesión de los tonos (agógica, tempo) por otra” [Riemann,1884].

Aunque el material para el cálculo de la esperanza básica es muy simple (tan sólo un intervalo implícito del contexto, ó la diferencia entre dos tiempos de

19 Descomposición de la curva esperanza básica.
20 Sumatoria de las esperanzas básicas aportadas por cada intervalo implícito en el contexto.

ataque), la recolección de datos empíricos sobre la esperanza percibida puede ser sin embargo mucho más complicada. Carolyn Drake propuso una medida de precisión en la tarea de ajuste que utilizó en su trabajo sobre el acento [Drake,Botte y Gérard,1989]. También se debe considerar los juicios de discriminación usados por Palmer y Krumhansl de criterio de bondad y ajuste y confusión de la memoria [Palmer y Krumhansl,1990]. Se puede deducir una medida más indirecta calculando las probabilidades de los diferentes intervalos (pares de ataques) de tiempo A, en el cuerpo de las piezas musicales. Esto no es igual que la aproximación por conteo de la frecuencia utilizada por Palmer y Krumhansl, porque este último hace uso del conocimiento a priori del metro.

La Fig. 4.4 representa la curva de esperanza compleja proyectada por el contexto [0 0.5 0.75 1] y su descomposición en las esperanzas básicas de los M intervalos implícitos en el patrón. El intervalo de predicción ha sido escogido entre 0 y 2 segundos y la curva de esperanza compleja generada por el contexto proyectada sobre el eje de tiempos. Como se puede observar el máximo de la predicción cae sobre el instante 0.5 después de la llegada del último bit (instante 1), es decir, en el instante 1.5 respecto al tempo absoluto, aunque una medición más precisa se obtiene de la diferencia entre los dos picos de máximo nivel de la curva de esperanza compleja (en este caso se obtiene el mismo resultado).

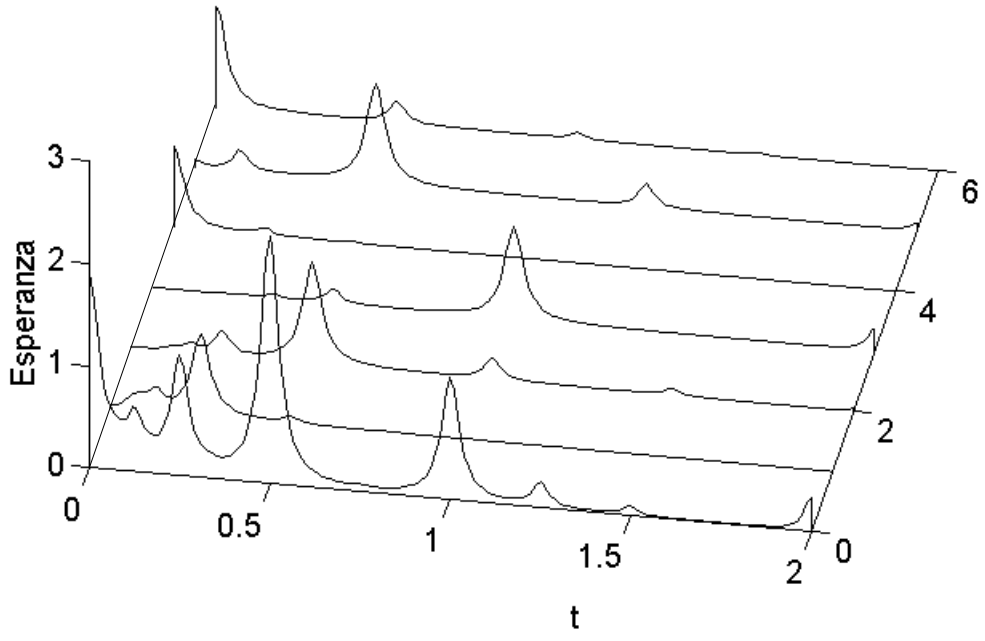


Fig. 4.4 Esperanza compleja proyectada por el patrón [0 0.5 0.75 1].

La Fig. 4.4 ilustra el concepto de descomposición, condición para un modelado cognoscitivo satisfactorio. En este punto solo resta añadir la selección de los intervalos implícitos en el contexto a la formulación matemática del modelo. Anteriormente se ha hecho referencia a estos intervalos como todas las posibles combinaciones diferentes de los tiempos de ataque que conforman el contexto y obsérvese que la diferencia está dada por la selección de los pares de tiempos de ataque y no por la longitud de los intervalos formados[21].

La Fig. 4.5 ilustra la descomposición del contexto o patrón temporal en sus intervalos implícitos. Véase que para el contexto formado por el conjunto $\{A1, A2, A3, A4\}$ de N intervalos existen M intervalos implícitos definidos por la Ec. (4.1). La línea vertical discontinua indica el instante inicial del intervalo de predicción. De igual forma se habrían podido obtener los aportes de esperanza básica de los intervalos implícitos del contexto en orden inverso comenzando a partir del último intervalo del contexto hacia el primero. El resultado sería el mismo. Véase también que aunque podemos calcular la curva de esperanza básica un intervalo de longitud tan larga como se quiera sólo es necesario evaluarla en el intervalo de tiempo absoluto correspondiente al tiempo de predicción (a partir de la línea vertical punteada).

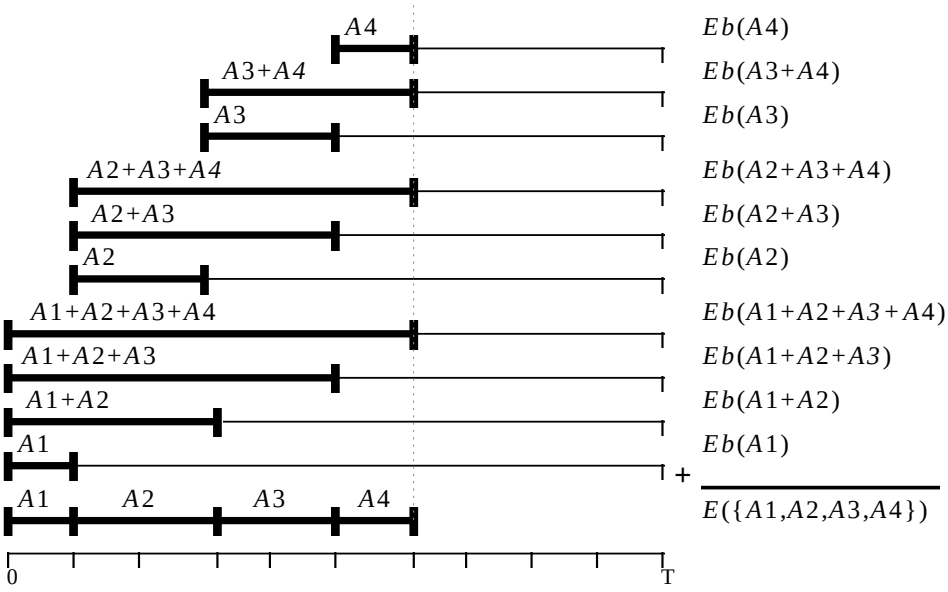


Fig. 4.5 Intervalos de tiempo utilizados en el cálculo de la esperanza básica.

21 Es muy probable que en determinado contexto existan varios intervalos de igual longitud (aproximadamente).

Dado un vector **A** que contiene los intervalos de tiempo básicos $A_i\left(1 \le i \le N\right)$ la esperanza de los intervalos desde p hasta q en el intervalo (t_i , t_f) se define como

$$E(\mathbf{A}) = \sum_{k=p}^{q} \sum_{j=k}^{q} E_b\left(\sum_{i=k}^{j} A_i, B_j \right) \tag{4.15}$$

El intervalo de predicción definido por los tiempos inicial y final t_i y t_f no aparece explícitamente en la formulación para mayor claridad aunque sí en la definición de la función Ψ que calcula los r componentes de la esperanza básica definida en la Ec. (4.7). En la representación dada por la Fig. 4.5 el tiempo inicial del intervalo de predicción t_i se corresponde a la línea vertical punteada ($t_i = \sum_{i=p}^{q} A_i$ respecto al tiempo absoluto) y $t_f > t_i$ a T. Véase que la línea punteada que indica el inicio del tiempo de predicción se corresponde al 0 del eje de tiempos de la Fig. 4.4 y T=2.

Como la esperanza básica $E_b(A,B)$ será más pequeña cuanto A sea más grande, existe un límite natural de longitud del contexto que contribuirá a la esperanza total $E(\mathbf{A})$, y el vector **A** puede funcionar como un constructor de memoria localizado.

El concepto de esperanza está estrechamente relacionado a las ideas de Jones y puede funcionar como una formalización de estas “Son como trayectorias psicológicas atencionales que guían rítmicamente nuestras energías atencionales a lo largo de trayectorias ideales. La atención es captada desde algún evento de referencia en un instante de tiempo hacia un evento objetivo lanzado un tiempo más tarde. Esta aproximación demuestra que la propia atención es dinámica, muchas acciones jerarquizadas basadas sobre ritmos internos concatenados. Nosotros mismos estamos continuamente lanzándonos hacia adelante anticipando rítmicamente eventos futuros que pueden ocurrir dentro de intervalos de tiempo pequeños y grandes. Estos derroteros forman los patrones de tiempo y espacio mental y pueden establecernos ese sentido de continuidad y conexión que acompaña la comprensión” [Jones,1981].

La esperanza evoluciona durante la presentación del patrón temporal y se prolonga desde el final del patrón hacia el futuro, como se muestra en la Fig.

4.4. Este tipo de curva muestra como falla un patrón temporal al realizar una alta esperanza (una síncopa) y permite ver como un nuevo evento refuerza el patrón de esperanza futura existente o introduce nuevos elementos en él.

La llegada de un nuevo evento contribuye en una extensión limitada a las protuberancias del estímulo percibido. Esto se puede demostrar construyendo los nuevos intervalos que introduce y proyectando sus esperanzas básicas hacia el pasado. Esto produce como resultado la contribución o reforzamiento a cada componente del patrón del nuevo evento. Los patrones temporales de mejor comportamiento, con alto soporte de eventos posteriores, serán recordados mejor. Ambos, esperanza y recuerdo son actividades ligadas a la dimensión del tiempo, caras opuestas de la misma moneda [Jones,1981]. El recuerdo es un proceso atencional dinámico que se desdobla en tiempo negativo. Este modelo predice, por lo tanto, como un evento puede facilitar o inhibir la memoria del pasado. Un ejemplo de este fenómeno es el “cierre” de un patrón temporal, que termina con un evento en un lugar altamente esperado: una posición métrica importante.

“El concepto de esperanza no tiene dirección en el tiempo. Está determinado completamente por un intervalo de tiempo, pero si los tiempos de ataque que marcan el intervalo fueron presentados en el pasado o deben de ocurrir en el futuro, es irrelevante a esta teoría” [Desain,1992]. Se puede, por lo tanto, hablar de la esperanza de un evento en futuro generado por dos tiempos de ataque en el pasado, el reforzamiento de un evento en el pasado por dos tiempos de ataque que ocurrirán en el futuro, o incluso la esperanza de un evento en un cierto tiempo entre dos instantes de tiempo. Todas estas ideas, a este nivel, son equivalentes y producen los mismos valores numéricos si la distancia entre los tiempos de ataques es la misma en los tres casos.

La Fig. 4.6 muestra las curvas de esperanza medidas con patrones temporales grandes introducidos como contexto a la red. Todos tienen la misma característica global: metro 2/4, y producen curvas muy similares. Véase los picos prominentes a mitad de barra[22] y en los límites de barra y picos de menos nivel en las subdivisiones de estos intervalos. Como se observa ciertos patrones de entrada producen picos pronunciados en las curvas de esperanza en las fronteras métricas importantes.

22 En los tiempos 2 y 4 contando en corcheas.

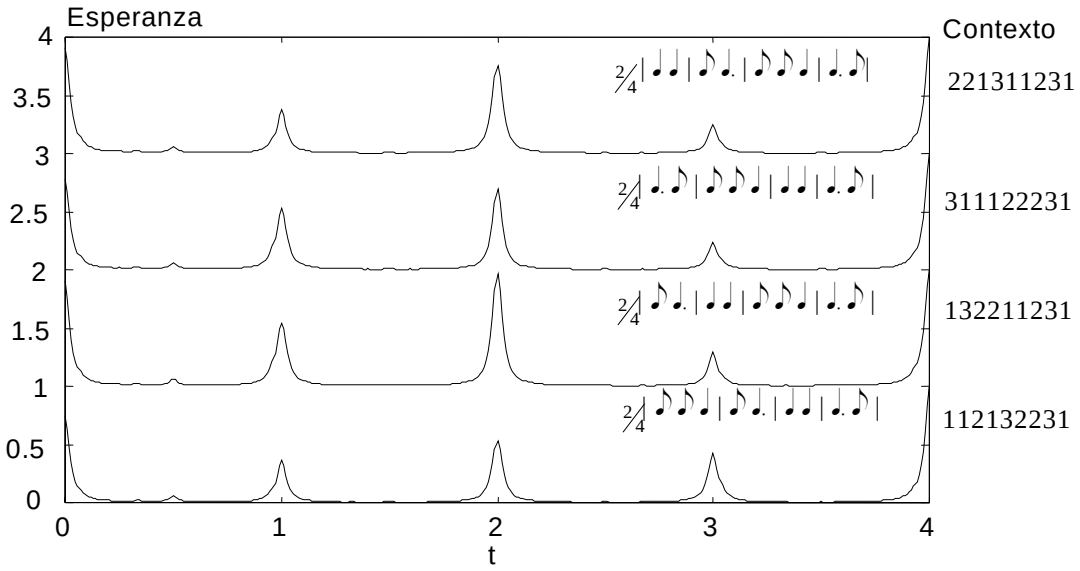


Fig. 4.6 Esperanza generada por diferentes patrones.

Otro aspecto atractivo del uso de las curvas de esperanza para un modelo de inducción del pulso es su naturaleza intrínseca incremental, la facilidad para tratar con ataques sobre la línea de tiempo continua, y su calidad para representar la inducción de pulsos ambiguos y la degradación de pulsos poco claros ó débiles.

La mayoría de los modelos de inducción del pulso tienden a interpretar la ausencia de una nota en la posición esperada del pulso[23] como indicación de error del pulso inducido y de la necesidad de actualización (e.g., [Longuet-Higgins y Lee,1982]; [Lee, 1985]) - ello minimiza, por consiguiente, la ocurrencia de la síncopa.

Sin embargo, es muy posible que después de la inducción de un pulso estable, por un período de tiempo relativamente grande, el material rítmico no contenga notas en las posiciones del pulso, y ocurran síncopas. Por lo tanto, un modelo sofisticado tiene que predecir como y cuando una síncopa puede ser escuchada como tal, en lugar de promover a un cambio en el pulso percibido.

En este modelo la síncopa es procesada realimentando los picos altos de la curva de esperanza en el contexto, como si el acento realmente haya sido escuchado. En otras palabras, si el evento no llega en un instante de alta expectación, se añade al patrón temporal un evento virtual o *pulso perdido*.

23 “Pulso perdido”.

Da la impresión que en los patrones fuertemente sincopados el silencio de cada pulso perdido casi puede ser oído. Aún los pulsos perdidos (e.g., ausentes en el patrón estímulo) influyen en la interpretación de los datos que llegan. Al igual que una nota y un pulso, tienen un peso asociado, el pulso perdido tiene asociado su propio peso. Esto regula la sensibilidad del modelo a la síncopa y controla su influencia sobre el nuevo material que llega.

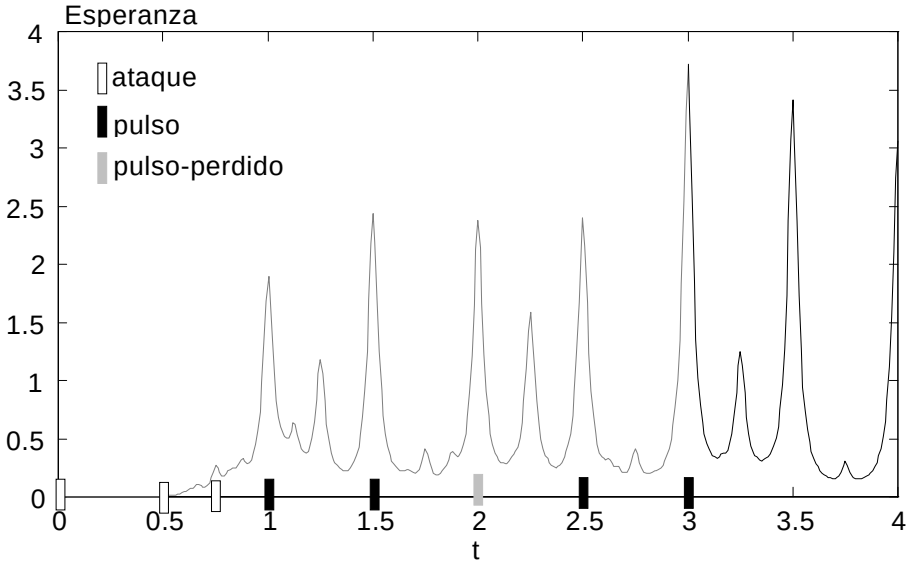


Fig. 4.7 Identificación de los pulsos y pulsos perdidos para el patrón 211242.

Cuando se presenta un patrón rítmico abstracto con diferente tempi, el pulso percibido puede desplazar niveles diferentes de la jerarquía métrica. En patrones ambiguos (para los cuales existen múltiples interpretaciones métricas mutuamente incompatibles) la alternativa del tempo global puede influir en la preferencia de uno u otro. El modelo de inducción del pulso puede asignar diferentes estructuras del pulso al mismo ritmo a diferente tempi debido a la dependencia de los intervalos de tiempo respecto al tempo global. Por lo tanto, el modelo puede seguir los cambios en el tempo global, como el ritardando en la Fig. 4.8b en comparación con la versión de tempo constante en la Fig. 4.8a.

La mayoría de los modelos existentes de inducción del pulso trabajan sobre duraciones extraídas del pentagrama (como se utilizó para la representación de la Fig. 4.8) y no de eventos extraídos directamente del material acústico[24]. Sin embargo la operación sobre datos reales permite modelar la expresividad

24 Basados en la observación de que el mayor volumen de los problemas asociados al modelado de inducción del pulso puede ser estudiado a partir de datos sintéticos (abstractos).

temporal y los cambios de tempo, a menudo considerados como ruido o fluctuaciones del tiempo[25]. Este modelo utiliza el patrón temporal extraído de la secuencia musical a través del método de segmentación descrito en el Capítulo 3, (con expresividad temporal).

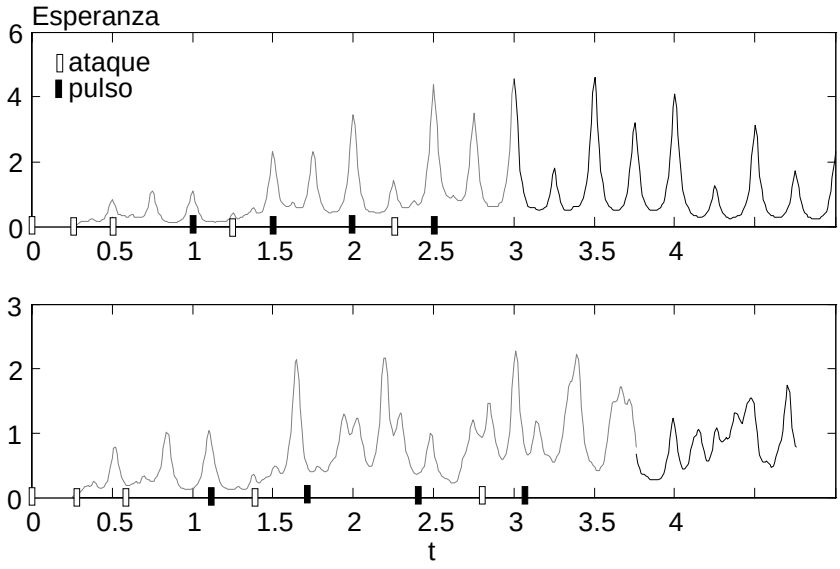


Fig. 4.8 Identificación de los pulsos dado un patrón en duraciones [0.25 0.25 0.5 0.25 0.25 0.5 0.25 0.25 0.5] (a), y el mismo patrón ejecutado con un ritardando [0.257 0.295 0.549 0.268 0.328 0.656 0.327 0.355 0.727] (b).

Sin embargo, en las interpretaciones en vivo a menudo, metro y pulso están comunicados por el tiempo [Sloboda,1985]. Por lo tanto los modelos de inducción que toman los datos directamente del material acústico como entrada pueden trabajar realmente mejor si utilizan la información de expresividad temporal presente en lugar de intentar librarse de ella.

En este modelo, la expresividad temporal ayuda al método de encontrar el pulso porque los tiempos consistentes agudizan los perfiles de las curvas en las fronteras estructurales enfatizadas en la ejecución y desecha los detalles irrelevantes de las regularidades no-intencionales. Por lo tanto, si se opera con la expresividad del tiempo los pulsos podrían ser interpretados correctamente incluso en aquellos patrones temporales ambiguos con respecto a la estructura métrica.

25 Inclusive en aquellas investigaciones que tratan con datos reales procesan la información de expresividad temporal por algún tipo de método de cuantización [Desain y Honing,1991].

Sloboda argumenta la existencia de la percepción categórica en el ritmo destacando la diferencia entre la percepción del ritmo y la percepción de la expresividad temporal: “La identificación del ritmo intencional es un aspecto común a todos los oyentes, que tienen que encarar continuamente el potencialmente confuso fenómeno del rubato y los cambios graduales de velocidad. En contraste, la percepción precisa de las desviaciones desde la metricalidad es difícil y requiere de un entrenamiento muy específico. Es casi imposible para un ejecutante imitar a otro exactamente. Esto requiere que el escucha logre categorizar las duraciones de las notas que oye en crochetes, trémolos, etc., ... sin desear afirmar que la percepción categórica hace imposible discriminaciones temporales más refinadas. *Podemos* oír imprecisiones rítmicas y rubato con un entrenamiento apropiado, pero las diferencias de tiempo sutiles a menudo no son experimentadas como tal, sino como diferencias en la calidad (la “vida” u “oscilación”) de una interpretación” [Sloboda,1985].

“Por lo tanto la cuantización no desecha las desviaciones de una interpretación métrica estricta, sino que separa el tiempo en componentes estructurales y expresivos que puedan ser manipulados por procesos diferentes” [Desain,1992]. El método de inducción del pulso mediante las curvas de esperanzas, al tratar con expresividades temporales parece prometedor en función de la cuantización[26].

La percepción categórica es un fenómeno bien-formalizado en la investigación de la voz, pero no ocurre lo mismo en el dominio del ritmo donde se hace más difícil la demostración de su existencia. La línea general de los descubrimientos acerca de la percepción categórica del ritmo parece indicar que ésta es facilitada por el contexto lo que concuerda con la teoría de la esperanza expuesta[27].

26 La existencia de máximos y mínimos locales en la curva de esperanza sugiere algún tipo de segmentación del eje continuo de tiempo en regiones discretas asociadas a categorías rítmicas.

27 Las curvas de esperanza son más pronunciadas si se dispone de mayor contexto.

Capítulo 5 Conclusiones

La búsqueda de un método fiable y generalizable para la localización y seguimiento del pulso a partir del material acústico bruto es un tema abierto. Los métodos desarrollados en el trabajo podrían servir como punto de partida hacia la búsqueda de tal objetivo pero no son de ninguna manera conclusivos.

La segmentación propuesta, basada en los trabajos de Goto y Muraoka [Goto y Muraoka,1994], aunque parcialmente probada, debe ser sometida a una experimentación exhaustiva en presencia de diversos estímulos complejos, instrumentos, ruido, etc., y sintonizada para operar en ausencia de un entorno computacional tan sofisticado[1]. A pesar de los buenos resultados obtenidos en las simulaciones realizadas sería conveniente conectar este método de segmentación con otras áreas de la investigación acerca de la caracterización del material acústico[2]. Con un acercamiento en paralelo a las propiedades de la señal acústica es más efectiva la sintonización de los parámetros del algoritmo segmentador.

El modelo de inducción del pulso mediante las curvas de esperanza basada en los trabajos de Desain [Desain,1994], dada su descomponibilidad en componentes simples que modelan la percepción de los intervalos de tiempo, parece un buen candidato en la búsqueda de una base teórica común para muchas teorías incompatibles de percepción del ritmo y la memoria. Esta teoría enlaza elegantemente la esperanza proyectada hacia el futuro y el reforzamiento de eventos en el pasado por datos nuevos. Las predicciones siguiendo esta teoría son consistentes con algunas obtenidas en la percepción categórica del ritmo, pero requieren de una verificación empírica exhaustiva.

Aún queda por formalizar el empleo de las curvas de esperanza en los procesos cognoscitivos mencionados tales como la cuantización, inducción del

1 El método ha sido probado sólo en contextos monofónicos, ya sea para señales de espectro armónico e inarmónico.
2 Propiedades estadísticas, etc.

metro y del pulso, la ritmicidad y la similitud de ritmos. Está claro que una teoría completa de percepción del ritmo no puede ser basada únicamente en el tiempo y en la contribución de los eventos a la esperanza, sino que debe tener en cuenta otros parámetros musicales.

Para una simulación completa del sistema propuesto sería conveniente la realimentación de ambos métodos. La inducción del pulso propuesta al categorizar el ritmo podría condicionar las decisiones del segmentador. El segmentador está obligado a su vez a localizar eventos fiables para el adecuado funcionamiento de la inducción. La interacción entre ambos procesos potenciaría un sistema integral de localización e inducción fiable y robusto.

Para simular el funcionamiento del sistema se utilizó la herramienta MATLAB v4.0 for Microsoft Windows de Math Works, Inc., los Toolboxes Signal Processing suministrado por MATLAB® y Signal Processing & Communications suministrado por Dennis Brown y Monique P. Fargues de la Naval Postgraduate School, Monterey y un conjunto de programas específicos diseñados al efecto.

El material acústico empleado como dato de entrada al sistema fue generado sintéticamente con los programas Super Studio Session™ y SoundEdit™ para Apple Macintosh.

La continuación de esta investigación debe estar orientada a una comprobación empírica de los componentes del sistema y su completa integración en presencia de secuencias acústicas reales, a la sintonización modelo-entorno, la simulación detallada de los componentes teóricos perceptuales integrados en el modelo, la validación real del modelo, la incorporación de los nuevos avances teóricos en el dominio de la percepción del ritmo y la memoria y a su introducción en el dominio de la polifonía.

Apéndice A Ejemplos

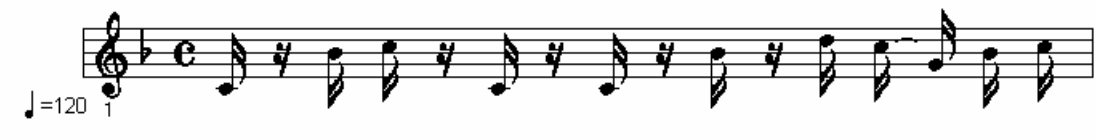
Ilustraciones de algunos de los ejemplos utilizados durante la simulación de los distintos componentes del sistema.

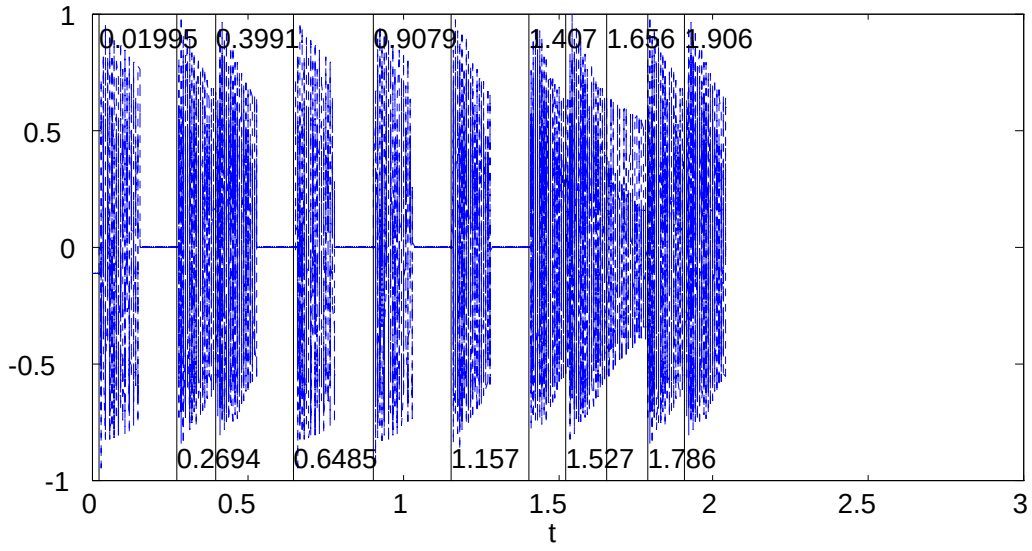
Para cada ejemplo se representa: pentagrama, resultado de la segmentación, pulso inducido. Para la segmentación en todos los casos se utilizó una ventana temporal de 30 mseg, una resolución frecuencial de 26.66 Hz y un desplazamiento entre ventanas de 10 mseg. En la inducción del pulso el intervalo de predicción en todos los casos fue de 1 seg.

Aunque el tempo indicado es ♩ = 120, en realidad para todos los experimentos ♩ = 60, ó 1 seg. Esto es debido al efecto de diezmado con factor 2 a que se sometieron las secuencias de prueba para poder simularlas en un entorno computacional aceptable. La frecuencia de muestreo empleado en la captación de estas señales fue de 11 025 Hz (reducida por el diezmado a 5 512.5 Hz). Para considerar los tiempos respecto a las señales originales es necesario multiplicarlos por el factor de diezmado 2.

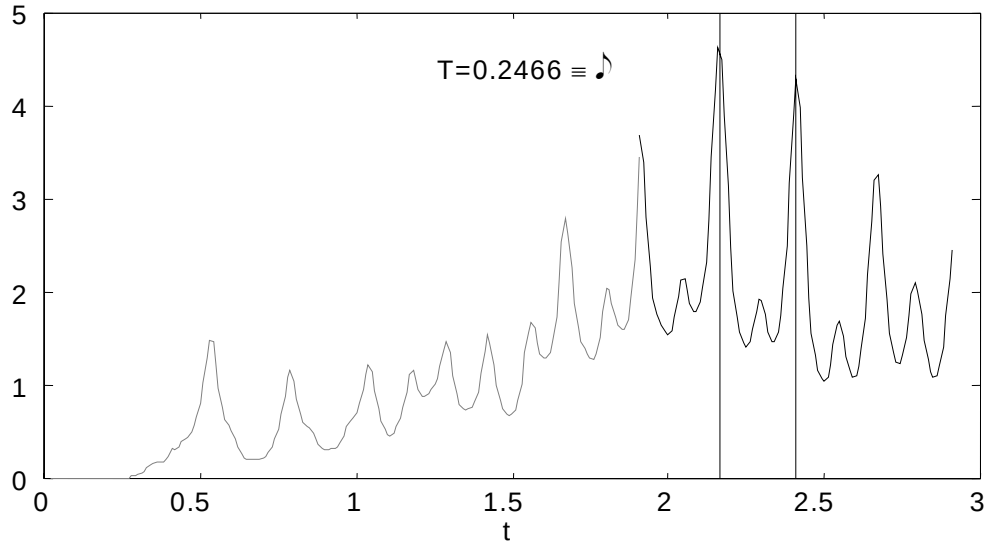
Los timbres empleados en la mayoría de los casos tienen ataques bruscos y relajación pronunciada y se corresponden a sonidos percutidos de espectro inarmónico (e.g., tumbas, güiros, timbales, etc.) y armónico (e.g., pianos, marimbas y bajos sintetizados). En algunos ejemplos se utilizaron combinaciones de más de un instrumento (e.g., en el Ejemplo 6 se utilizaron dos güiros diferentes, etc.).

Ejemplo 1.

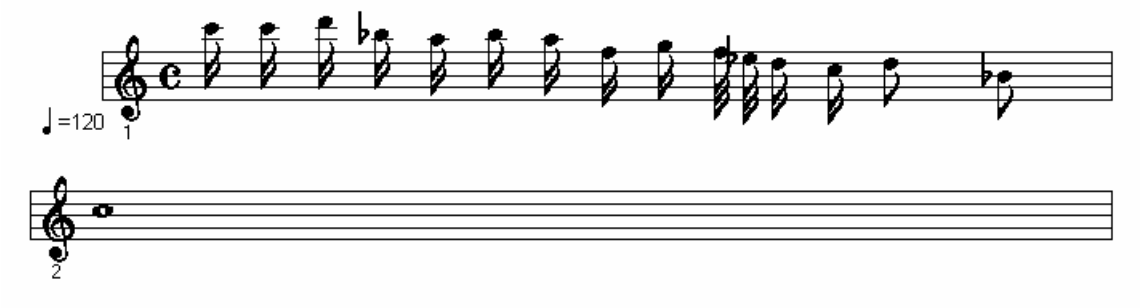


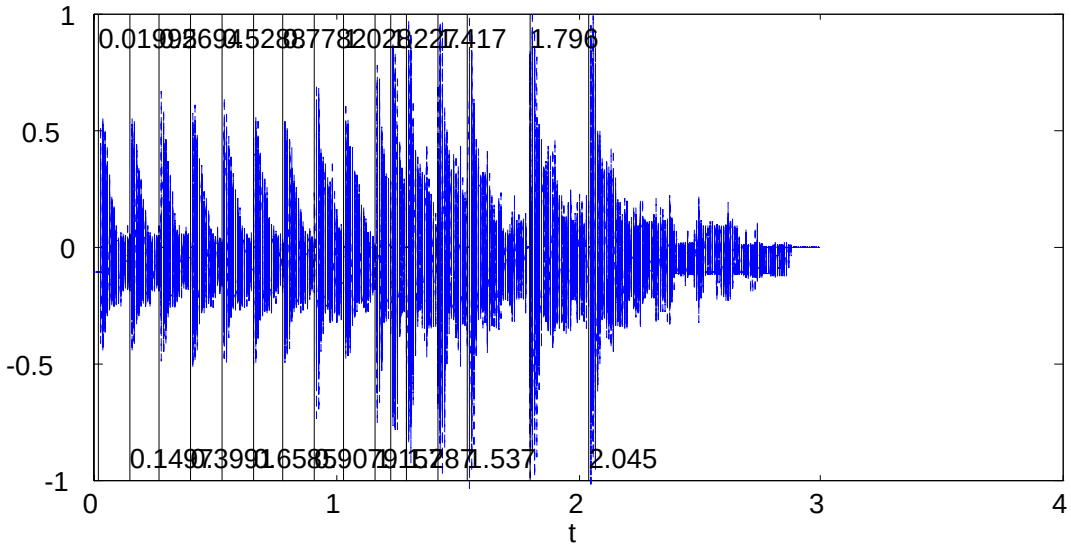


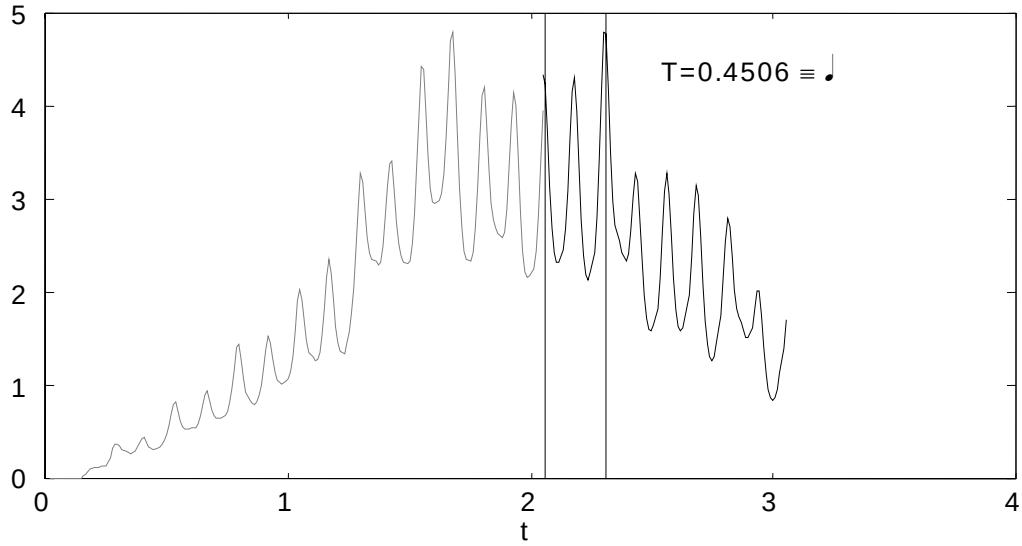
Véase el efecto de solapamiento que produce el ligado.



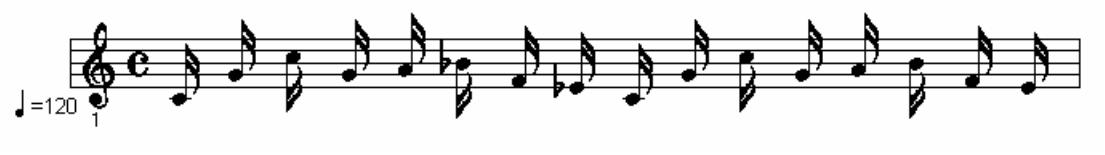
Ejemplo 2.

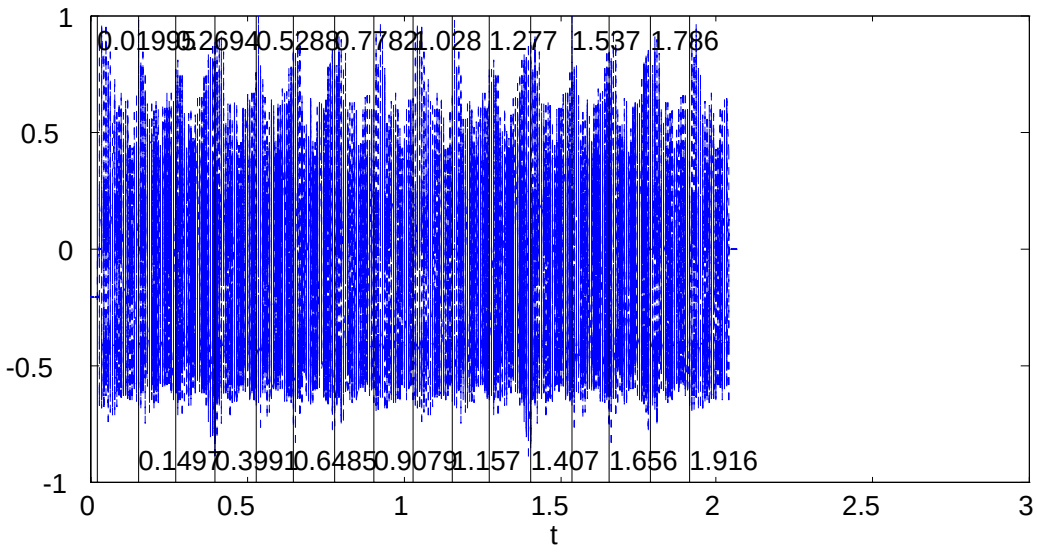

♩=120
1
2

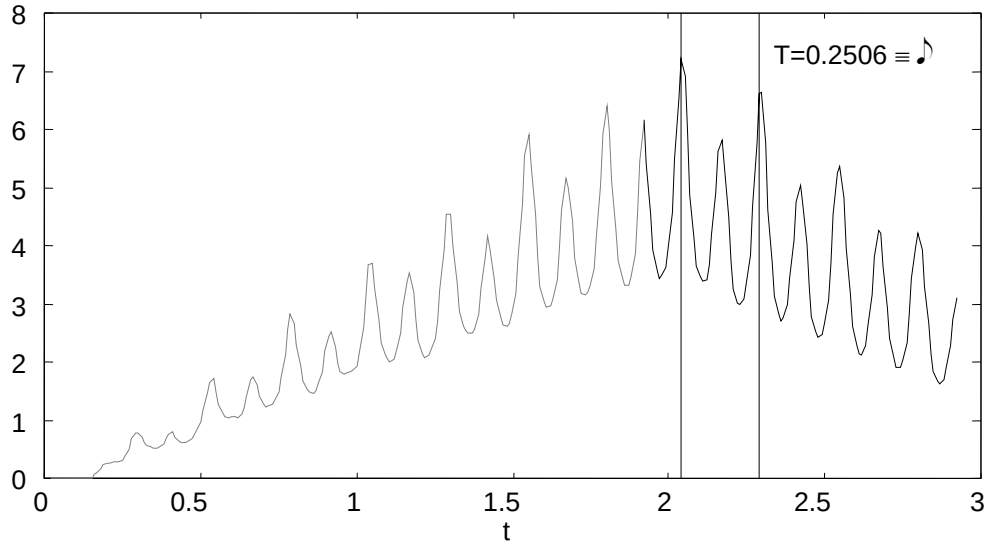

1
0.5
0
-0.5
-1
0
1
2
3
4
t
0.019
1.796
0.149
1.537
2.045


5
4
3
2
1
0
0
1
2
3
4
t
T=0.4506 ≡ ♩

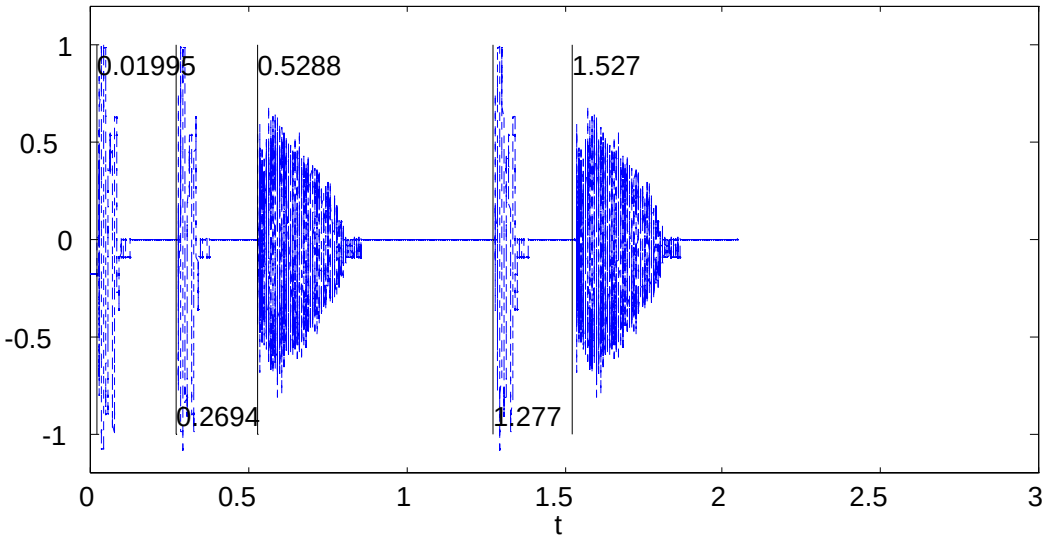
Ejemplo 3.

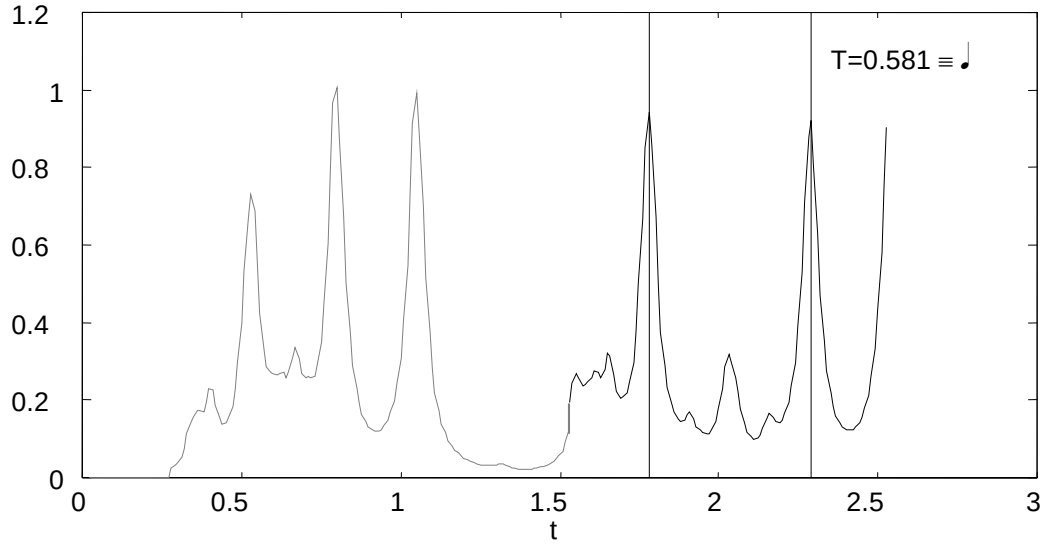

♩=120


0.01995
0.2694
0.5288
0.7782
1.028
1.277
1.537
1.786
0.1497
0.3991
0.6485
0.9079
1.157
1.407
1.656
1.916
t

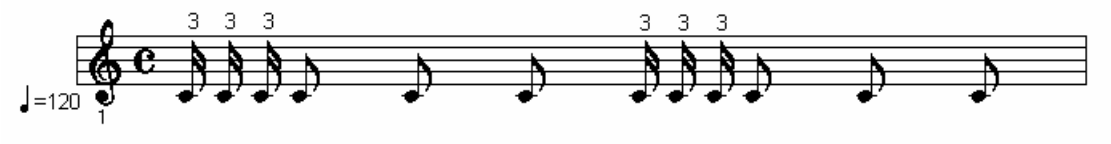

T=0.2506 ≡
t

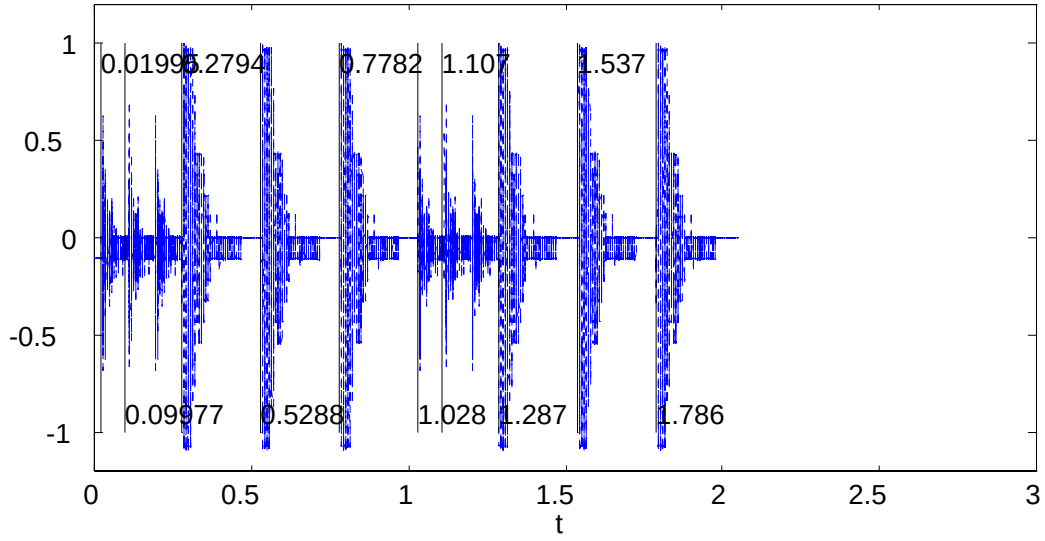
Ejemplo 4.

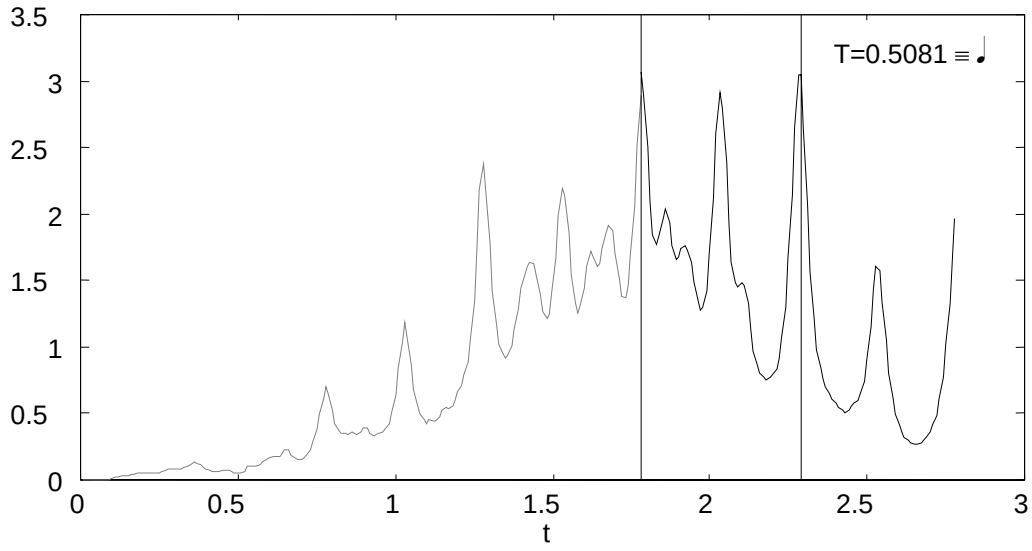




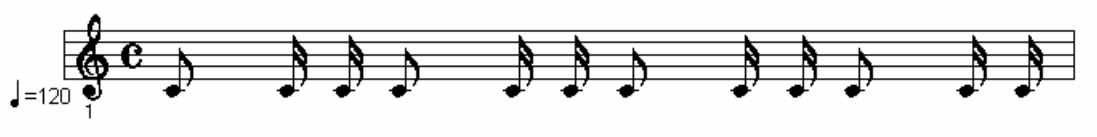
Ejemplo 5.

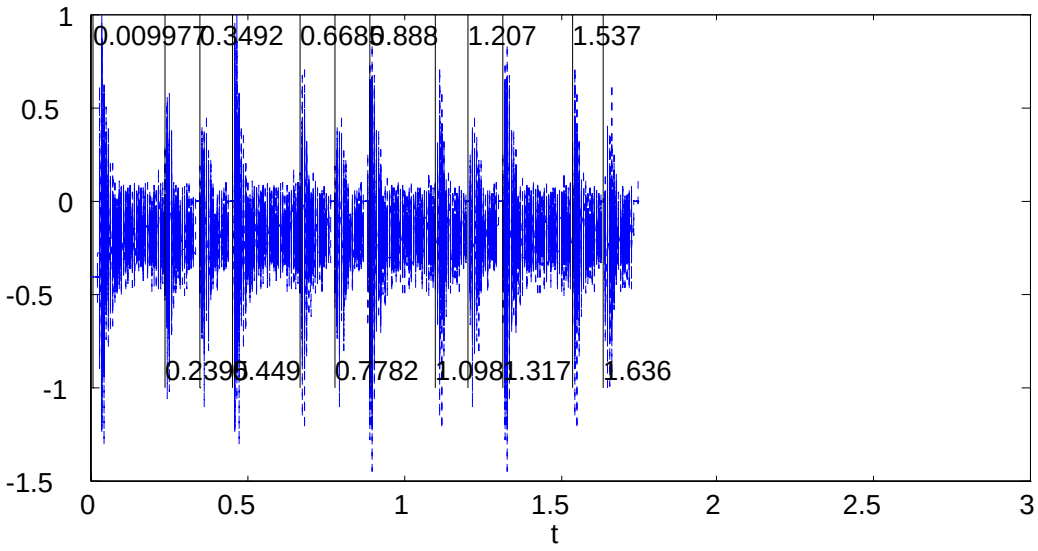

3 3 3
3 3 3
=120
1

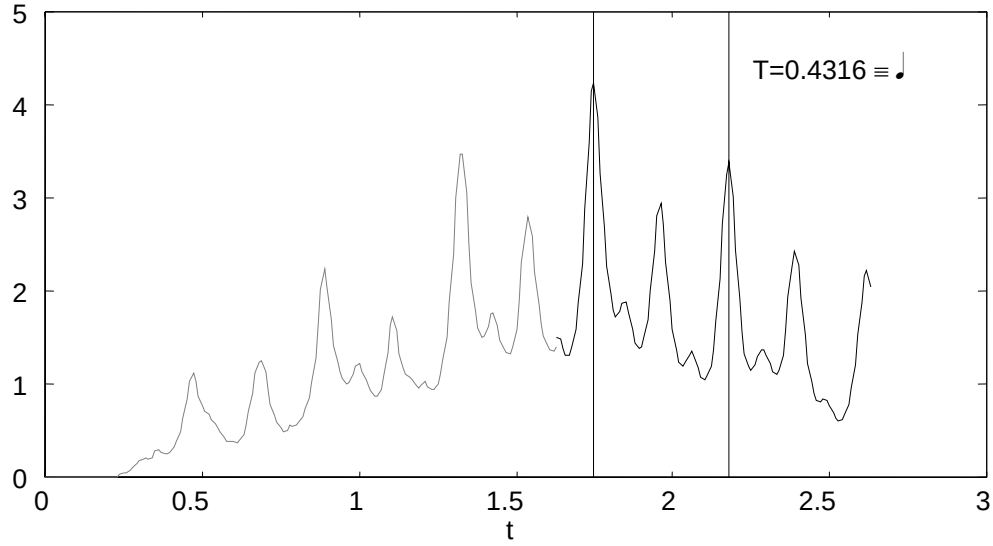

1
0.5
0
-0.5
-1
0.01995
0.2794
0.7782
1.107
1.537
0.09977
0.5288
1.028
1.287
1.786
0
0.5
1
1.5
2
2.5
3
t


3.5
3
2.5
2
1.5
1
0.5
0
T=0.5081 ≡
0
0.5
1
1.5
2
2.5
3
t

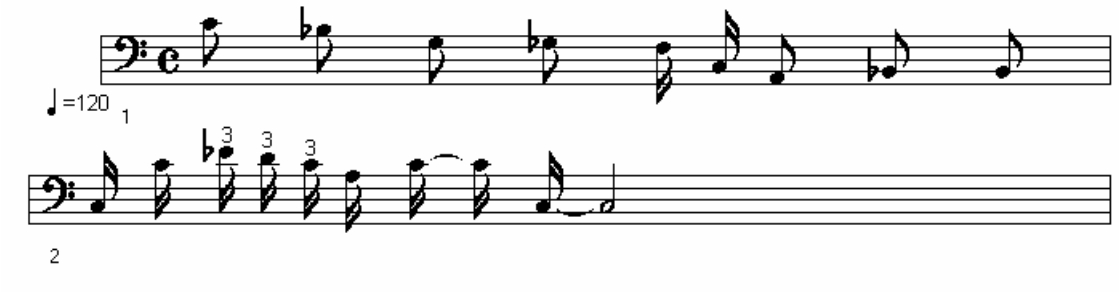
Ejemplo 6.

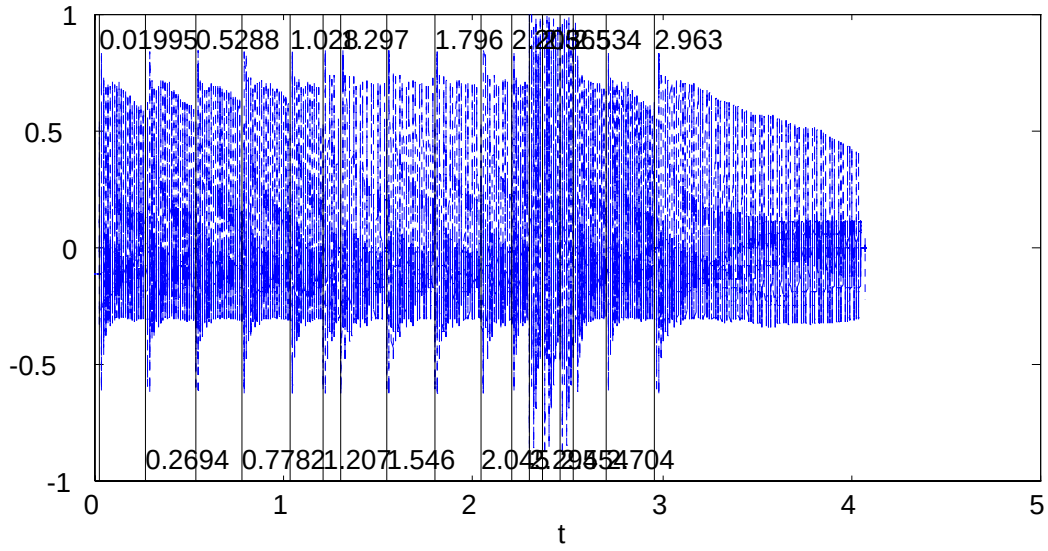

♩=120
1

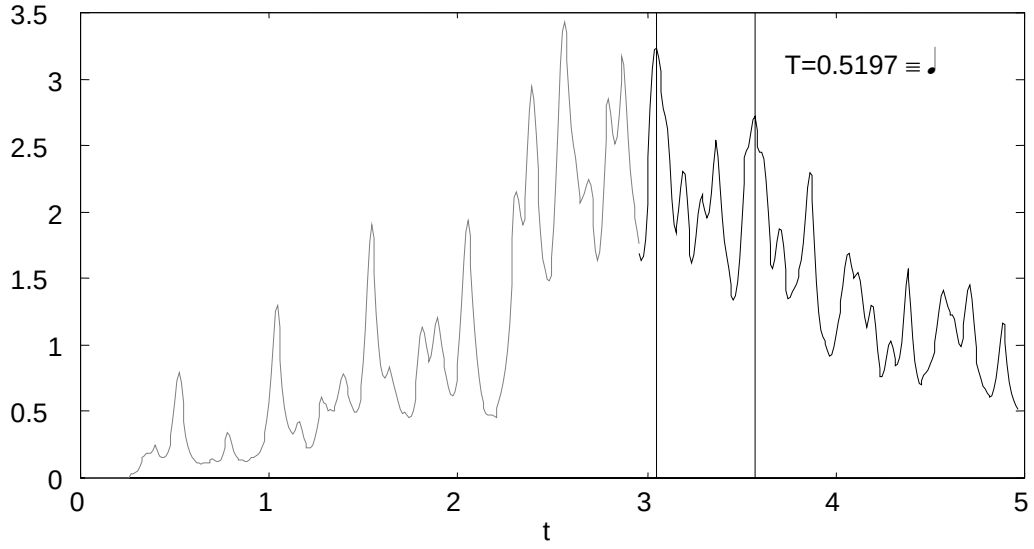

0.009977
0.3492
0.6685
0.888
1.207
1.537
0.2396
0.449
0.7782
1.098
1.317
1.636
1
0.5
0
-0.5
-1
-1.5
0
0.5
1
1.5
2
2.5
3
t


T=0.4316 ≡ ♩
5
4
3
2
1
0
0
0.5
1
1.5
2
2.5
3
t

Ejemplo 7.


♩=120
1
3
3
3
2


1
0.5
0
-0.5
-1
0
1
2
3
4
5
t
0.01995
0.5288
1.028
1.297
1.796
2.963
0.2694
0.7782
1.207
1.546
2.045
2.704


3.5
3
2.5
2
1.5
1
0.5
0
0
1
2
3
4
5
t
T=0.5197 ≡ ♩

Referencias

La bibliografía contiene todos los libros y artículos referenciados en el texto. Muchas de la referencias han sido tomadas indirectamente a través de alguno de los autores citados.

Allen, J.B., & Rabiner, L.R., "A Unified Approach to Short-Time Fourier Analysis and Synthesis," *Proc. IEEE,* Vol. 65, 1558-1564, Nov. 1977.

Clarke, E., Levels of structure in the organization of musical time. *Contemporary Music Review,* 5 (1), 1-30.

Cooper, G., & Meyer, L., *The Rhythmic Structure of Music*. Chicago: University of Chicago Press, 1960.

Crochiere, R.E., & Rabiner, L.R., *Multirate Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

Desain, P., & Honing, H., The quantization od musical time: a connectionist approach. *Computer Music Journal,* 1989, 13 (3), 56-66.

Desain, P., A connectionist and a traditional AI quantizer, symbolic versus sub-symbolic models of rhythm perception. In I. Cross (De.), Proceedings of the 1990 Music and the Cognitive Sciences Conference, *Contemporary Music Review.* London: Harwood Press, 1990.

Desain, P., & Honing, H., The quantization of musical time: a connectionist approach. *Computer Music Journal*, 1989, 13 (3), 56-66. [Fue reimpreso en P.M. Tood & D. G. Loy (Eds.), *Music and connectionism. Cambrige*, MA: MIT Press, 1991]

Desain, P., A (de)composable theory of rhythm perception. *Music Perception,* 1992, 9 (4), 439-454.

Desain, P., & Honing, H., Advanced issues in beat induction modeling: syncopation, tempo and timing. *ICMC Proceedings,* 1994, 92-94.

Drake, C., Botte, M.C., & Gérard, C.A., A perceptual distortion in simple musical rhythms. *Proceedings of the International Society for Psychophysics Fifth Annual Meeting*, Cassis, France, 1989.

Fraisse, P., Rhythm and tempo. In D. Deutsch (Ed.), *The psychology of music,* Orlando, FL: Academic Press, 1982.

Gabrielsson, A., Performance of rhythmic patterns, *Scandanavian Journal of Psychology*. 1974, 15, 63-72.

Gabrielsson, A., & Bengtsson, Y., & Garielsson, B., Performance of muscial rhythm in 3/4 and 6/8 meter. *Scandanavian Journal of Psychology,* 1983, 24, 193-213.

Gordon, J., *Perception of Attack Trascients in Musical Tones*, Ph.D. Dissertation, Department of Music, Stanford University, 1984.

Goto, M., & Muraoka, Y., A beat tracking system for acoustic signals of music, School of Science and Engineering, Waseda University, 1994, 365-372.

Goto, M., & Muraoka, Y., & Rosenthal, D., Rhythm tracking using multiple hypotheses, *ICMC Proceedings,*1994, 85-87.

Gray, J., *An Exploration of Musical Timbre*, Ph.D. Dissertation, Stanford University, también Department of Music Report No. Stan-M-2 (1975).

Henderson, M.T., Rhythm organization in artistic piano performance. In Carl E. Seashore, Ed., *University of Iowa Studies in the Psychology of Music*, Vol. IV. Iowa City: University of Iowa Press, 1936.

Hirsh, Ira J., Auditory perception of temporal order. *Journal of the Acoustical Society of America,* 1959, 31 (6), 759-767.

Jones, M. R., Only time can tell: On the topology of mental space and time. *Critical Inquiry,* 1981, 7, 557-576.

Jones, M.R. & Boltz, M., Dynamic attending and responses to time. *Psychologycal Review,* 1989, 96 (3), 459-491.

Lee, C.S., The rhythmic interpretation of simple musical sequences: towards a perceptual model. In R. West, P. Howell, & Y. Cross (Eds.), *Musical Structure and Cognition*, 1985, 53-69, London: Academic Press.

Longuet-Higgins, H.C., The perception of melodies. *Nature,* 1976, 263, 646-653.

Longuet-Higgins, H.C., & Lee, C.S., Perception of musical rhythms. *Perception*, 1982, 11, 115-128.

Lunney, H.W.M., Time as heard in speech and music. *Nature*, 1974, 249:592.

Moorer, J.A., *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*, Ph.D. Dissertation, Stanford University, 1975.

Nawab, S.H., & Quatieri, T.F., "Short-Time Fourier Transform," in *Advanced Topics in Signal Processing*, J.S. Lim and A. V. Oppenheim, Eds., Prentice-Hall, Englewood Cliffs, NJ, 1988.

Oppenheim, A., & Schafer, R., Discrete-Time Signal Processing, Prentice-Hall, Englewood Cliffs, NJ, 1989.

Palmer, C., & Krumhansl, C.L., Mental representations for musical meter. *Journal of Experiment Psychology: Human Perception and Performance*, 1990, 16 (4), 728-741.

Papoulis, A., *Probability, Random Variables, and Stochastic Processes*. 3ra Ed., McGraw-Hill Book Company, New York, 1991.

Patterson, J., & Green, D. M., Discrimination of transient signals having identical energy spectra. *Journal of the Acoustical Society of America,* 1970, 48 (4), 894-905.

Piszczalski, M., & Galler, B., "Predicting Musical Pitch from Component Ratios," *Journal of the Acoustical Society of America*, 1979, 66 (3), 710-720.

Povel, D. J., Time, rhythms and tension: In search of the determinants of rhythmicity. Internal Report 84FU11, University of Nijmegen. 1984.

Povel, D., & Essens, P., Perception of temporal patterns. *Music Perception,* 1985, 2, 411-440.

Rabiner, L.R., & Schafer, R. W., *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, 1978.

Riemann, H., *Musikalische Dynamik und Agogik,* Hamburg, 1884.

Rioul, O., & Vetterli, M., Wavelets and signal processing. *IEEE SP Magazine,* 1991, October, 14-38.

Rosenthal, D., Emulation of human rhythm perception. *Computer Music Journal,* 1992, 16 (1), 64-76.

Schloss, W. A., On the automatic transcription of percussive music - from acoustic signal to high-level analysis, 1985, *Ph.D. Thesis,* CCRMA, Stanford University.

Seashore, C. E., Ed. *University of Iowa Studies in the Psychology of Music*, Vol I. *The Vibrato*, Vol III. Psychology of the Vibrato in Voice and Instrument, Vol IV. *Objective Analysis of Musical Performance*, Iowa City: University of Iowa Press, 1932, 1936.

Seeger, C., “An Instantaneous Music Notator”, *Journal of the International Folk Music Council* III: 103-106, 1951.

Shaffer, L.H., Performances of Chopin, Bach, Bartok: Studies in Motor Programming. *Cognitive Psychology*, 1981, 13, 326-376.

Sloboda, J.A., *The musical mind: The cognitive psychology of music*. Oxford: Oxford University Press, 1985.

Stautner, J.P., *The Auditory Transform*, MS thesis, MIT, 1982.

Sternberg, S., Knoll, R.L., & Zukofsky, P., Timing by skilled musicians. In D. Deutsch (De.), *The psychology of music,* Orlando, FL: Academic Press, 1982.

Vorberg, D.J., & Hambuch, R., On the temporal control of rhythmic performance, In J. Requin, ed. *Attention and Performance*, VII, 1987,

Yeston, M., *The Stratification of Musical Rhythm*, New Haven: Yale University Press, 1976.