Inferencia de Tactus con Fundamentos Estadísticos para Tap-dancing

Martin A. Miguel *

Laboratorio de Inteligencia Artificial Aplicada, FCEyN, UBA

Resumen El siguiente trabajo constituye un nuevo acercamiento al problema de la inferencia automática de *tactus* en un pasaje musical. El tactus se define como aquél pulso constante que la gente mantiene con su pie o mano al escuchar una canción. Este problema suele abordarse utilizando reglas musicales, como ser reglas que definen qué eventos están acentuados. En la solución aquí presentada dejamos estas reglas de lado y buscamos una definición funcional del tactus que nos ayude a lograr esta inferencia. De esta forma, el modelo desarrollado busca ser más simple y encarar el problema adicional de trabajar sobre música rítmicamente expresiva (en particular, *tap-dance*). El paper describe nuestro modelo cognitivo de tactus - denominado `Tactus Hypothesis Tracker` - y evaluamos su precisión en benchmarks presentados en trabajos anteriores en el tema. El modelo desarrollado presenta resultados similares en un benchmark que agrupa estilos musicales comúnmente usados en la bibliografía; y presenta mejoras en un benchmark de *tap-dance* desarrollado en este trabajo.
Keywords: Art and Music, Cognitive Modeling

1. Introducción

El placer de escuchar música es innegable y está presente en casi todas las culturas a lo largo del tiempo. Los orígenes y las formas de este placer todavía son un misterio. En algunos estilos musicales, como en la música percusiva africana, el interés se produce mediante pequeños corrimientos en la sincronización de los eventos musicales [9]. El *tactus*, ese pulso constante que un oyente lleva con su pie o con su mano, es clave para la comprensión de la estructura rítmica de un pasaje musical [3]. Esos corrimientos - detalles en la interpretación y juegos en la música - son elementos que contrastan con la estructura y es este contraste el que da lugar a la musicalidad.

La evidencia sobre la importancia del *tactus* se amontona. Un oyente promedio puede seguir este pulso casi sin pensarlo y distintos oyentes suelen acordar en cuál es pulso [4, 11]. Un mismo patrón rítmico es reconocido por dos personas como pasajes distintos si se les sugieren dos pulsos base distintos [1]. En la notación formal el tactus es la base sobre la que se escribe la música [7].

En la música tradicional occidental, la estructura métrica de una canción es únicamente un esqueleto sobre el cual se montan otros elementos que le darán expresividad a la música [9]. En aquellos estilos musicales donde el principal – sino el único – instrumento es percusivo, la expresividad debe obtenerse mediante otros recursos. Tal es

* m2.march@gmail.com

el caso del tap-dance, por ejemplo, donde las estructuras rítmicas son constantemente creadas y luego desafiadas para lograr pasajes musicales interesantes. La manipulación de las estructuras rítmicas observadas en el tap-dance es lo que denominamos *música rítmicamente expresiva*.

Queremos empezar a comprender cómo las estructuras rítmicas surgen en la mente del oyente y cómo se las hace interesantes al jugar con la expectativa del mismo. Para ello primero debemos inferir el pulso interno que una persona siente al escuchar este tipo de producciones musicales. En este trabajo desarrollamos el *Tactus Hypothesis Tracker*, un modelo de inferencia automática de *tactus*.

La principal motivación teórica del modelo desarrollado es lograr la inferencia intentando comprender el porqué del pulso inferido por el oyente. Para ello nos basamos en la propuesta realizada por Huron [4]. El el mismo, el autor relaciona los mecanismos de predicción y expectativa presentes en las personas con las distintas sensaciones evocadas en la música. Dada la importancia del tactus en la comprensión de los elementos rítmicos, podemos pensar al mismo como una herramienta de resumen y predicción de los eventos musicales. Nuestra definición funcional del tactus quedará expresada en un valor de confianza del mismo respecto de la canción.

Al visualizar el *tactus* como una herramienta de predicción de los eventos musicales en el tiempo buscamos liberamos de los preconceptos musicales utilizados por otros autores. A partir de estos conceptos más simples esperamos llegar a un modelo que se adapte mejor a música no tradicional. Por otra parte, la evaluación aquí realizada pretende dar a conocer la dependencia que tienen estos preconceptos musicales en el proceso de inferencia.

Otra decisión aplicada en este trabajo es el modelado del tactus como un reloj preciso - una serie de pulsos isócronos. Esta decisión es importante para poder detectar las pequeñas fluctuaciones del momento en que ocurren los eventos respecto del pulso perfecto. En el futuro, detectar las fluctuaciones será importante para modelar otras características de expresividad. Argumentamos la decisión considerando que los músicos aprenden y practican utilizando un metrónomo como acompañamiento, el cual es un reloj preciso.

Finalmente es necesario aclarar que dado que el sistema busca dar información sobre el proceso cognitivo mientras se escucha una canción, el mismo provee información de forma continua sobre su proceso de inferencia. A partir de esta información buscamos en un trabajo futuro detectar cambios discretos en la velocidad del tactus. Cambios continuos - realentizaciones y aceleraciones suaves - no se encuentran en el alcance del trabajo actual.

1.1. Trabajo previo

El enfoque normalmente utilizado en la temática busca inferir los distintos niveles de la estructura métrica. Cada nivel de la estructura métrica se define como una serie de momentos en el tiempo, donde cada nivel superior es un subconjunto de los elementos del nivel inmediato inferior. Los momentos de un nivel que aparecen en el nivel superior suelen estar a igual distancia entre sí. Los distintos niveles pueden luego asociarse a distintas partes de la estructura métrica: tactus, compás e hipermetro. Este tipo de análisis rítmico suele ser el paso siguiente a la inferencia de tactus. Ejemplos de sistemas que

realizan este tipo de análisis son los desarrollados por [1], [9], [8] y [2]. Temperley [11] presentó el sistema denominado `Melisma`, que además del métrico realiza análisis de clave, armonía y contrapunto.

Sistemas como [8] y [11] definen la posición de cada pulso del tactus localmente, normalmente expandiendo una hipótesis ya existente sobre la estructura de la canción. La hipótesis, compuesta por momentos ya seleccionados, se expande agregando un nuevo evento más adelante en el tiempo. Estos sistemas tienden a ubicar los pulsos del tactus directamente sobre los eventos, de ser posible y coherente. Como resultado, suelen tener mucha variabilidad en el intervalo entre eventos (inter-beat-interval). La complejidad de estos modelos y la dificultad para luego observar imperfecciones de la interpretación musical nos direccionó en nuestra decisión de utilizar un reloj preciso.

La mayoría de los sistemas citados son sistemas basados en reglas: funcionan a partir de un conjunto de reglas con las que toman decisiones respecto de la generación, preservación y evolución de hipótesis. Muchas de ellas son o se asemejan a las enunciadas en *Generative Theory of Tonal Music* [6]. Estas reglas fueron pensadas para la música tonal, que es sistema de organización de la música principalmente utilizado en la cultura occidental. La motivación estadística de nuestro trabajo nos permite un modelo más simple (menos reglas) y más genérico (menor influencia de un estilo en la decisiones tomadas).

De los trabajos mencionados anteriormente, todos salvo [2] y [9] trabajan con representaciones simbólicas de la música. Estas representaciones informan la secuencia de eventos musicales: qué nota se produjo, cuando y cuanto duró. Un formato comúnmente utilizado es el *midi*. Goto y Schloss, en cambio, realizan análisis de señales sobre el audio de las grabaciones. En nuestro caso trabajamos con grabaciones `midi` que fueron adaptadas a nuestro modelo - poseen una única voz y nos desinteresamos de la información tonal.

La mayoría de los trabajos relacionados aquí mencionados fueron puestos a prueba con ejemplos musicales occidentales. Rosenthal y Temperley trabajan con música para piano. Goto utiliza canciones populares con métrica en 4/4. Schloss se enfoca en audios de percusión en batería. Povel y Essens crearon patrones rítmicos específicos para su experimento. En este trabajo estaremos evaluando nuestro modelo tanto sobre ejemplos de música occidental como sobre pasajes de tap.

2. *Tactus Hypothesis Tracker*

Desarrollamos el `Tactus Hypotheses Tracker`, un modelo que infiere y evalúa hipótesis de tactus usando solo información rítmica del pasaje musical.[1] Nuestro principal interés es la música de *tap*, por lo que nuestro modelo solo considera un instrumento y un evento del mismo a la vez. De esta forma podemos definir la forma en que nuestro modelo ve la música como una lista ordenada (m_i) de milisegundos - la ubicación de los eventos musicales. En esta versión del modelo no estamos trabajando con cambios continuos en el pulso (realentizaciones y aceleraciones). Con esto en

[1] Una versión funcional del modelo puede encontrarse en `https://github.com/m2march/tht`

cuenta podemos modelar una hipótesis de tactus como un valor en milisegundos ρ - la fase o ubicación del pulso - y un valor en milisegundos δ - el intervalo entre pulsos.

El sistema recibe como entrada un archivo *midi* y expone como resultado el seguimiento de cada hipótesis de tactus sobre el tiempo. El seguimiento incluye los cambios en los valores de la hipótesis (ρ, δ) y las actualizaciones de su valor de confianza. Midis polifónicos fueron adaptados a nuestro modelo.

2.1. Generación de hipótesis

Si consideramos (m_i) la ubicación ideal de los eventos musicales que el músico posee en su mente, llamaremos (r_i) a la ubicación de los mismos cuando estos se interpretan en la realidad. De esta forma tenemos $r_i = m_i + e_i$ con e_i la diferencia entre el ideal y la realidad producida por el interprete. Esta diferencia surge por el error natural en cualquier ejecución motriz en conjunto con irregularidades realizadas adrede como parte de la expresividad musical de la ejecución.

Considerando que inferimos el tactus para poder resumir la canción, el mismo deberá coincidir con al menos dos eventos musicales. Es así que podemos decir que las hipótesis a considerar serán aquellas $h = (\rho, \delta)$ donde $\rho = r_k$ y $\delta = r_j - r_k$ para algún $k < j$.

En una reproducción sin errores ($e_i = 0$), alguna de las hipótesis h definidas representaría el tactus correcto. Siendo que este no es el caso en una reproducción real, deberemos corregir nuestras hipótesis iniciales de forma que se ajusten lo mejor posible al pasaje. Esta corrección la denominaremos Δh. Para continuar con nuestro análisis debemos primero definir la proyección de una hipótesis sobre una reproducción:

$$p((\rho, \delta), (r_i)) = (p_k^{\rho,\delta}) \text{ con } p_k^{\rho,\delta} = \rho + k \times \delta$$

Los valores de k son tales que $min(p_k^{\rho,\delta}) \geq min((r_i)) - \frac{\delta}{2}$ y $max(p_k^{\rho,\delta}) \leq max((r_i)) + \frac{\delta}{2}$. Esto es, representan un recorte de la proyección infinita de la hipótesis de tactus de forma que no se extienda demasiado por sobre la canción. A partir de ahora dejamos de lado el superíndice $^{\rho,\delta}$.

Las correcciones de hipótesis Δh se calculan como una regresión lineal de la línea constante 0 sobre la *confianza del error de predicción*. La misma se define como:

$$pe_k = m \times (r_{p_k} - p_k) \times d^{\frac{|p_k - r_{p_k}|}{\delta}}$$

con r_{p_k} el evento musical más cercano a p_k. El parámetro multiplicativo m define cuanto afecta el error a la corrección y el parámetro de decaimiento d define que tan rápido un error se considera falta de coincidencia entre el evento y la predicción p_k. Esto es importante ya que existen momentos en la música donde el pulso está presente pero no hay ningún evento musical. Estos son los silencios de la música. Tales situaciones no son errores de predicción, por lo que no deben afectar la corrección. Los parámetros se establecieron mediante prueba y error.

Dentro de universo de hipótesis consideradas, existirán muchas que serán equivalentes, ya que en el caso ideal los dos eventos base utilizados se encuentran en la proyección de otra hipótesis. Se definió un índice de similitud entre hipótesis que tiene en

cuenta cómo el error afecta la proyección de las hipótesis. Mediante este índice, hipótesis suficientemente parecidas se unen en una sola descartando la más anterior de las dos.

Otra consideración realizada en la generación de hipótesis es limitar el valor de δ a estar entre 187ms y 1500ms (320 y 40 bpm, respectivamente). Limitar el intervalo entre pulsos a no ser demasiado largo o corto es común en la bibliografía. Se ha demostrado que un intervalo entre 600ms y 750ms es en el que las personas tienen mejor capacidad de mantener un pulso constante [4]. A partir del corpus presente en el toolkit `music21`[2] observamos que los percentiles 9 % y 91 % en la distribución de intervalos del pulso para la corchea es de 60 y 200 bpm respectivamente. Nuestros valores son conservadores respecto de esta distribución. El corpus contiene 1700 partituras.

2.2. Confianza de una hipótesis de tactus

Cuando escuchamos música buscamos comprender, resumir y predecir los eventos que sucederán. Siendo un ritmo una serie de eventos en el tiempo, el tactus - un pulso continuo - es una forma muy concisa de representar los eventos. Un tactus que logre un buen resumen de la música será aquél que prediga tantos eventos musicales como sea posible y evite predecir eventos que no sucedieron.

Queremos saber para cada valor de la predicción p_k si coincide con algún evento musical. En este trabajo decidimos no utilizar un criterio fuerte para esta decisión. En cambio, decidimos calcular un valor de confianza para la coincidencia. Definimos la confianza de coincidencia de un evento de la predicción p_k como:

$$conf(p_k, r_{p_k}) = 0{,}01^{\frac{|p_k - r_{p_k}|}{\delta}}$$

En particular nos interesa la confianza al evento más cercano en la reproducción r_{p_k}. La confianza vale 1 cuando $p_k = r_{p_k}$ y decae exponencialmente hacia 0 con la distancia entre los valores.

Para capturar el concepto de resumen realizado por la hipótesis de tactus queremos saber cuantos eventos de la canción pueden ser explicados por la hipótesis y que porcentaje de las predicciones concuerdan con la canción. Definimos la confianza de una hipótesis h de la siguiente manera:

$$conf(h, (r_i)) = \sum_k \frac{\text{concordancias de la hipótesis}}{\text{predicciones de la hipótesis}} \times \frac{\text{concordancias de la hipótesis}}{\text{eventos musicales}}$$

Con las siguientes definiciones formales:

$$\begin{aligned} \text{concordancias de la hipótesis} &= \sum_k conf(p_k, r_{p_k}) \\ \text{predicciones de la hipótesis} &= |(p_j)| \\ \text{eventos musicales} &= |(r_i)| \end{aligned}$$

[2] `http://web.mit.edu/music21/`

2.3. Evaluación continua

Debido a que el sistema está pensado para asemejarse a la percepción humana, el análisis se realiza de izquierda a derecha sobre el pasaje musical, evolucionando continuamente. Asumiendo que la canción se escuchó hasta el milisegundo d, las únicas hipótesis consideradas son aquellas donde $\rho + \delta < d$; sin contar aquellas descartadas por asimilarse mucho con alguna otra ya generada. Con cada nuevo evento musical r_k descubierto, nuevas hipótesis son generadas, todas las hipótesis son corregidas con el cálculo de Δh y se vuelven a buscar hipótesis similares para descartar. Además, la confianza de todas las hipótesis se recalcula. Esta evolución se registra ya que conforma parte de la devolución del sistema. Cuando se calcula la confianza de la hipótesis se realiza solo sobre la proyección limitada al último $r_k < d$ observado.

3. Resultados

En esta sección presentamos la evaluación de la precisión de inferencia de nuestro sistema. Deseamos evaluar la capacidad de obtener el valor de δ correcto para el tactus de la canción. Estaremos comparando contra el sistema `Melisma` [11]. Calcularemos el valor inferido δ_i de dos formas distintas para cada sistema. Dado el valor del intervalo esperado δ_c, consideraremos que el sistema infirió correctamente el tactus si $|\delta_i * k - \delta_c| < \epsilon_c$ para algún entero k. ϵ_c se estableció en $1{,}5$.

En su trabajo, Temperley [11] utiliza los corpus KP y KP-Perf para evaluar su sistema de inferencia métrica. Tanto su sistema de inferencia como los corpus mencionados se encuentran a disposición online. [3] Para suscribirnos a la propuesta de Temperley, utilizamos sus corpus para nuestra propia evaluación. En este trabajo generamos además un nuevo corpus, denominado *tap*, que también se encuentra disponible online para su libre uso [4] .

El corpus KP es un conjunto de extractos del libro de ejercicios de Kostka y Payne [5]. El corpus KP-Perf es una selección de estos extractos que son solo piano, ejecutados por una pianista experimentada [11]. Nuestro corpus, *tap*, se conforma de transcripciones de patrones de tap, algunos de ellos producidos dos veces a distintas velocidades.

tht	weighted_tht	melisma_ma	melisma	
0.89	0.87	0.83	0.85	KP
0.39	0.44	0.61	0.61	KP Perf
0.50	0.40	0.20	0.10	tap

Cuadro 1. Precisión en la detección de tiempo por parte de los sistemas de inferencia.

En la tabla 1 mostramos los resultados de la evaluación de ambos sistemas. Todas las canciones en los corpus tienen un tempo teórico constante. Esto quiere decir que, según la notación, la velocidad del tactus no cambia. En el corpus KP los midis son

[3] `http://www.link.cs.cmu.edu/melisma/`

[4] `https://github.com/m2march/tht`

una ejecución precisa de la notación ($e_i = 0$). Los corpus *tap* y KP-Perf contienen ejecuciones reales, que contienen correcciones expresivas en el tiempo de los eventos así como error humano. No poseen cambios en su velocidad más allá de estas pequeñas perturbaciones.

Los resultados aquí presentados son el porcentaje de extractos para los cuales δ_c fue correctamente inferido según el criterio explicado previamente. En el corpus KP, δ_c fue obtenido de la información midi. Para los corpus KP-Perf y *tap*, se los obtuvo manualmente.

Para cada sistema evaluado, se utilizaron dos formas distintas de calcular δ_i. El sistema `THT` da como resultado la mejor hipótesis en cada momento de la canción. En el campo *tht* de la tabla calculamos δ_i como el δ más común entre todas las hipótesis ganadoras momento a momento. *weighted_tht* es el δ más común pesado positivamente acorde a que tan temprano aparece como hipótesis ganadora en la canción. Esto se realizó para representar el *principio de consistencia* [10].

El sistema `Melisma` tiene por resultado la lista de momentos en los que los pulsos de cada nivel métrico suceden. En la columna *melisma* calculamos δ_i como el promedio de los intervalos entre pulsos para el nivel métrico de tactus de la salida. En la columna *melisma_ma* presentamos una media movil de los mismos datos. Este cálculo se hizo para intentar compensar por las pequeñas diferencias entre los intervalos entre pulsos (ver Trabajo Relacionado).

En la tabla de resultados observamos resultados equivalentes para el corpus KP, un poco de peor calidad por parte de nuestro sistema en KP-Perf, y algo mejor en el corpus de *tap*.

4. Discusión

El actual trabajo desarrolla el modelo de inferencia de tactus en pasajes musicales llamado `Tactus Hypothesis Tracker`. Este modelo considera el tactus de una canción como una forma de resumir la misma, de forma de evitar preconceptos musicales utilizados en sistemas equivalentes. El sistema `THT` está orientado a trabajar con música rítmicamente expresiva, en particular *tap dance*.

Vemos que ambos sistemas tienen alta precisión en el corpus KP. Esto es esperado ya que no existe error en las reproducciones, lo que representa la principal dificultad en esta tarea. Vale aclarar que en este trabajo no estamos buscando inferir exactamente el tactus inferido por un oyente sino que nos conformamos con un múltiplo del mismo. Esta búsqueda se relaciona con preguntas de psicología cognitiva que definen la preferencia a centrarse en un pulso en lugar de en su subdivisión. Abordar estas preguntas es remanente para trabajo futuro.

La alta precisión del sistema `THT` en el corpus KP indica que el criterio de confianza definido es una buena medida de aptitud del tactus inferido. La leve ventaja de nuestro sistema sobre `Melisma` se debe a nuestro modelado del tactus como un reloj preciso, lo que elimina el ruido de las fluctuaciones del intervalo entre pulsos. Además, en todos los casos donde el δ_i no es estrictamente un múltiplo de δ_c, observamos que la relación entre ambos era aproximadamente $\frac{3}{2}$. Esto es equivalente a tener el nivel del tactus de

δ_c como los pulsos principales de un compás 3/4 por encima del pulso descripto por δ_i. Esta situación podría también considerarse correcta.

En el corpus KP-Perf nuestro sistema no obtuvo resultados tan buenos. Los valores esperados de δ se definieron a mano escuchando los pasajes musicales. Tales definiciones resultaron no ser tan fáciles de realizar, lo que indica que definir la correctitud de una inferencia tampoco es tan directo. Considerando que `Melisma` funciona dando preferencia a los eventos musicales en sí para elegir el momento de los pulsos del tactus, puede que esto le haya dado más solidez al sistema al enfrentar la expresividad que aparece en KP-Pref. En `THT`, la información para adaptarse a estos cambios se deriva de los valores de confianza calculados al avanzar sobre el pasaje. En el futuro se pretende hace mejor uso de esta información mediante una mejor definición del *principio de consistencia*.

Finalmente, el sistema `THT` muestra mejores resultados que `Melisma` en el corpus *tap*. Los pasajes de este corpus no suscriben a las convenciones rítmicas de la música clásica y desafían la estructura rítmica de distinta forma. Además, al ser música percusiva, no hay información relevante de altura o duración de las notas (que `Melisma` sí utiliza).

En resumen, podemos concluir que podemos inferir el intervalo del pulso constante de un patrón musical pensando al tactus como un agente de resumen del mismo. Además, esto puede realizarse sin información de altura o reglas para determinar eventos musicales acentuados. Esta concepción estadística del tactus funciona tanto para música occidental tradicional como para música percusiva no tradicional, como ser el *tap dance*.

Referencias

[1] Essens, P.J., Povel, D.J.: Metrical and nonmetrical representations of temporal patterns. Perception & Psychophysics 37(1), 1–7 (1985)

[2] Goto, M.: An audio-based real-time beat tracking system for music with or without drum-sounds. Journal of New Music Research 30(2), 159–171 (2001)

[3] Honing, H.: Without it no music: beat induction as a fundamental musical trait. Annals of the New York Academy of Sciences 1252(1), 85–91 (2012)

[4] Huron, D.B.: Sweet anticipation: Music and the psychology of expectation. MIT press (2006)

[5] Kostka, S., Payne, D., Schindler, A.: Workbook for Tonal harmony, with an introduction to twentieth-century music. McGraw-Hill (1995), `https://books.google.com.ar/books?id=7L43AQAAIAAJ`

[6] Lerdahl, F., Jackendoff, R.: A generative theory of tonal music. MIT press (1985)

[7] Martineau, J.: The Elements of Music: Melody, Rhythm, and Harmony. Wooden Books, Walker (2008), `https://books.google.com.ar/books?id=fyKdLgAACAAJ`

[8] Rosenthal, D.F.: Machine rhythm–computer emulation of human rhythm perception. Ph.D. thesis, Massachusetts Institute of Technology (1992)

[9] Schloss, W.A.: On the automatic transcription of percussive music: from acoustic signal to high-level analysis. No. 27, Stanford University (1985)

[10] Steedman, M.J.: The perception of musical rhythm and metre. Perception 6(5), 555–570 (1977)

[11] Temperley, D.: The cognition of basic musical structures. MIT press (2001)