Estimación de tempo en grabaciones de audio utilizando Octave

S.D. Bargas, M.T. Cassiet, J. Grinovero
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral

Resumen

En este documento se explican los procedimientos para la construcción de un algoritmo de estimación de tempo para grabaciones de audio en el programa Octave. Primero se explicarán conceptos musicales necesarios para comprender el tema y se presentarán los métodos ya existentes utilizados como base para la construcción del algoritmo. En segundo lugar se explicará en profundidad la implementación del algoritmo. Luego se mostrará un análisis de los resultados y se los comparará con otros métodos existentes. Por último, se cerrará con un comentario sobre las conclusiones obtenidas tras la finalización del trabajo.

1. Introducción

Las estaciones de trabajo de audio digital, los programas de mezcla musical y los programas de visualización musical utilizan el *tempo* de las canciones en la gran mayoría de sus funcionalidades. Poder obtenerlo de forma rápida y eficiente es una característica beneficiosa en la industria de software musical.

El *tempo* es la velocidad a la que se reproducen los patrones en una pieza musical, y depende de la cantidad de *beats* por unidad de tiempo (generalmente por minuto), así como también del *swing*. Un *beat* es un pulso que se repite de forma regular y constituye la base de un patrón musical. Entre *beats* se pueden encontrar pulsos menos significativos llamados *sub-beats*. El *swing* es una proporción de la separación entre dos *beats* y el *sub-beat* que se encuentra entre ellos.

El objetivo de este trabajo es el desarrollo de un sistema que permita identificar de manera automática el tempo de una canción a partir de una grabación de audio de la pieza musical.

Tomando lo desarrollado por Jean Laroche [1], para estimar el tempo de una canción el algoritmo debería utilizar principalmente características notorias dentro del contexto de la canción, como lo son el comienzo y el fin de notas, el cambio de una nota a otra o golpes de percusión. Esto se debe a que los *beats* suelen ocurrir en esos instantes. Una forma simple de realizarlo consiste en localizar variaciones rápidas de energía en la señal. Es preferible utilizar el dominio de la frecuencia de la señal en vez del dominio del tiempo ya que se evita que los sonidos agudos (de alta frecuencia) sean ocultados por fuertes sonidos graves (de baja frecuencia y alta amplitud). Esto se logra aplicando una función de compresión a la señal en el dominio del tiempo.

Según otro extracto de Laroche [2], una vez identificados los cambios de energía en función del tiempo, se procede a identificar los picos más significativos de esa señal obtenida. Se generan señales de probabilidad para diferentes posibles tempos donde se espera que se encuentren incrementos de energía y se la compara con la señal de flujo de energía para verificar la verosimilitud entre las dos señales. El tempo identificado de la canción es el tempo de la señal generada que consiga la mayor verosimilitud.

2. Métodos

2.1. Transformada de Fourier y cálculo del flujo de energía

Utilizando el método de Laroche [1], la señal se analiza utilizando ventanas sucesivas de 10 milisegundos, aplicándoles la transformada rápida de Fourier (FFT) a cada una de ellas y utilizando la magnitud de la transformada $X(f,t)$ para obtener la energía en ese fragmento de tiempo. A cada FFT se le aplica una simple función de compresión $C(x) = \sqrt{x}$ para evitar que se oculten sonidos como se explicó en el párrafo anterior.

Para identificar los cambios de energía más significativos $\hat{E}(i)$ entre ventanas consecutivas t_i de la magnitud de la FFT, se aplica una diferencia de primer orden entre los fragmentos tomando únicamente las frecuencias que se encuentran entre $f_{\text{mín}} = 100$ Hz y $f_{\text{máx}} = 10$ kHz.

$$\hat{E}(i) = \sum_{f=f_{\text{mín}}}^{f_{\text{máx}}} C(|X(f,t_i)|) - C(|X(f,t_{i-1})|) \tag{1}$$

A esta señal se le aplica una rectificación de media onda para obtener la señal positiva de flujo de energía $E(i)$:

$$E(i) = \begin{cases} \hat{E}(i) & \hat{E}(i) > 0 \\ 0 & \hat{E}(i) \leq 0 \end{cases} \tag{2}$$

2.2. Identificación de picos significativos

Utilizando el segundo extracto de Laroche [2], una vez identificados los cambios de energía en función del tiempo, se proceden a identificar los picos más significativos de esa señal obtenida.

En primer lugar, se encuentran todos los picos o máximos locales de la señal $E(i)$ utilizando la función de octave `[peaks, peak_locs] = findpeaks(E)` que retorna en `peaks` los valores de los máximos locales y en `peak_locs` sus posiciones.

Luego, para extraer los picos más significativos se establece un umbral de tolerancia igual al 50 % del valor del pico máximo y se descartan todos los picos menores a ese umbral. Los picos que superan el umbral indican potenciales ubicaciones de beats en el fragmento de audio analizado.

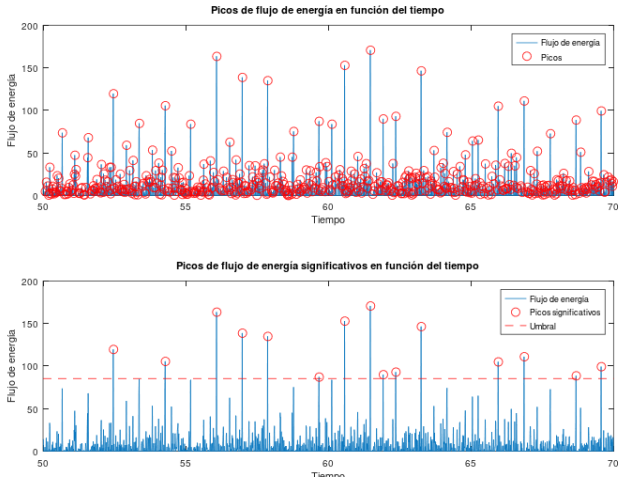


Figura 1: Identificación de picos significativos

2.3. Función de probabilidad

El *tempo*, el *swing* y la posición del primer *beat* definen un subconjunto de cuatro posiciones $\{b_0, b_1, b_2, b_3\}$ donde se estima una máxima probabilidad de encontrar picos significativos en el flujo de energía.

Este subconjunto se repite de forma periódica para todo el tamaño de la muestra formando el conjunto $b_i = \{b_0, b_1, b_2, b_3, \ldots, b_n\}$. Para cada una de las posiciones del conjunto se define una función de densidad de probabilidad gaussiana

$$p_t(t) = \sum_{i=0}^{n} p_i G((t \bmod T) - b_i) \tag{3}$$

donde t es el instante de tiempo que se evalúa, T es el tempo para el que se evalúa, G es la función gaussiana y p_i es un escalar que modifica la magnitud de la campana dependiendo del sub-beat.

Como la función de distribución de probabilidad es periódica con un periodo de medio beat (si hay swing) o un cuarto beat (si no hay swing), sólo se puede determinar la ubicación del beat hasta antes de ese medio-beat o cuarto-beat. Para eliminar esta ambigüedad, se agrega una simetría entre los cuatro sub-beats, haciendo que sea más probable que ocurra en el primer sub-beat, un poco menos probable en el tercer sub-beat y menos probable aún en el segundo y cuarto sub-beat. Los valores de p_i a utilizar son entonces $\{0{,}4, 0{,}15, 0{,}3, 0{,}15\}$.

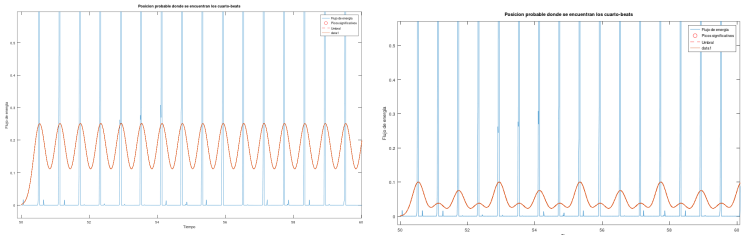


Figura 2: Beats, sub-beats y cuarto-beats. A la izquierda, igual probabilidad de aparicion en todos los sub-beats. A la derecha, mayor probabilidad de aparicion en los sub-beats 0 y 3

2.4. Máxima verosimilitud

Para calcular la medida en que la función de probabilidad generada p_t se asemeja al flujo de energía de la pista analizada se utiliza una función de verosimilitud L_t:

$$L_t(T, S, b_0) = \sum_{i=0}^{N-1} \log p_t(t_i) \tag{4}$$

donde T es el tempo, S es el swing y b_0 es la posición del primer beat.

En un escenario ideal, se calcularían todos los valores posibles de T, de S y de b_0 para la pista musical completa. Sin embargo, esto requeriría una inmensa cantidad de cálculos que, para realizarse en un período relativamente corto de tiempo, necesitarían un poder de cómputo incalculable. Es por este motivo que, a fines prácticos, se reducen y discretizan los rangos.

Si bien la determinación del tempo depende de un parámetro objetivo, existe una componente subjetiva atada al contexto. Al comparar un tempo $T_1 = n$ BPM con otro tempo $T_2 = 2n$ BPM (con $n \in \mathbb{R}$), los beats de T_1 coincidirán con la mitad de los beats de T_2. Es por este motivo que en canciones con tempo T_1, se podría tomar como válido su doble T_2. Por lo tanto, se puede limitar el rango de T a $[t_{\text{inicio}}, 2t_{\text{fin}}]$. Como la mayoría de las canciones tienen un tempo cercano a 100 BPM, se utiliza el intervalo $[70, 140]$. Aquellas canciones con $T < 70$ serán identificadas con $2T$ y aquellas canciones con $T > 140$ serán identificadas con $\frac{1}{2}T$. Además, se discretiza el rango utilizando sólo valores enteros.

Como el mínimo tempo identificable es 70 y la gran mayoría de las canciones tiene 4 beats por compás, cada compás dura como máximo aproximadamente 3,4 segundos. Por lo tanto, en vez

de tomar la pista musical completa, se puede tomar un fragmento de 4 segundos para asegurarse de incluir un compás. Cuanto mayor sea el tamaño del fragmento utilizado, más fiable será el algoritmo.

Para cada tempo a calcular, la ubicación del b_0 puede encontrarse entre $t = 0$ y $t = P$ donde $P = \frac{1}{T}$ es el periodo del beat. Se divide este intervalo en 32 valores discretos, que serán los valores de b_0 a calcular.

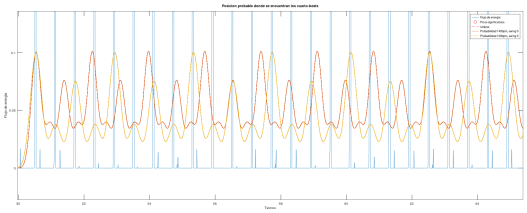


Figura 3: Audio de un metrónomo de 100BPM: en rojo, 140BPM. En amarillo, 100BPM

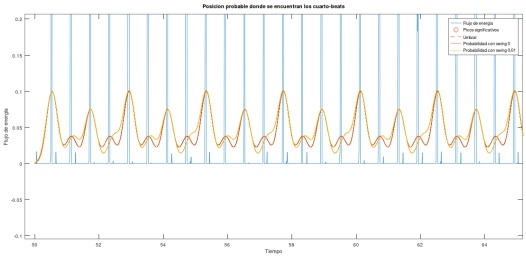


Figura 4: Audio de un metrónomo de 100BPM: en rojo, $S = 0\,\%$. En amarillo, $S = 1\,\%$

2.5. Optimización

A pesar de haber acotado y discretizado los rangos para reducir la cantidad de cálculos a realizar, el algoritmo siguió tomando largos periodos de tiempo en ejecutarse. Por este motivo se intentaron implementar optimizaciones como:

- A partir de las ubicaciones de los picos significativos, calcular los tamaños promedios de los intervalos entre dos picos significativos, para luego convertir esa medida en BPM y realizar la búsqueda de máxima verosimilitud en un rango de tempos menor, establecido como $T_{\text{aprox}} \pm 20$ BPM.
- Realizar una aproximación inicial con $T = \{70, 80, 90, 100, 110, 120, 130, 140\}$ que retorne un tempo aproximado T_0 y realizar nuevamente el cálculo para $T = [T_0 - 9, T_0 + 9]$.

Sin embargo, estas aproximaciones introdujeron errores que afectaban a la efectividad del algoritmo, por lo que fueron descartadas.

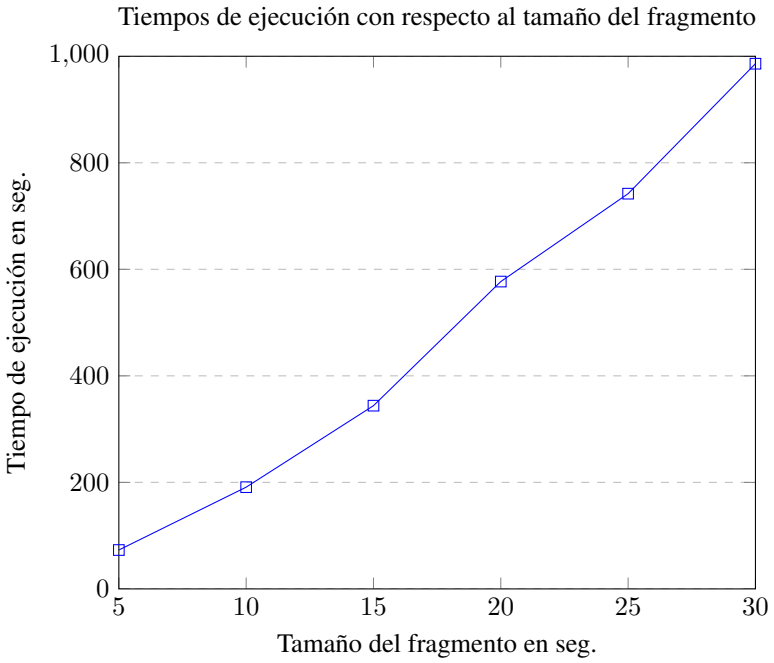
3. Resultados

La medida de efectividad del algoritmo a utilizar será el porcentaje de canciones para las que se haya calculado el tempo de forma correcta. Se considera correcta una estimación con una tolerancia a un error relativo del 5 %. Además, se tomará como correcta una un tempo estimado equivalente a la mitad o al doble del tempo real.

Se utiliza la recopilación de datos de algoritmos de detección de beats realizada por Alonso, M. [3] para comparar con el algoritmo implementado BCG (Bargas-Cassiet-Grinovero):

Método	Acierto
BCG (5 seg)	37,8 %
BCG (10 seg)	31,1 %
BCG (15 seg)	33,3 %
BCG (20 seg)	34,4 %
BCG (25 seg)	37,8 %
BCG (30 seg)	40,0 %
Paulus	56,3 %
Scheirer	67,4 %
SP	63,2 %
AC	73,6 %
SP-SEF	84,0 %
AC-SEF	89,7 %

Como se puede observar, el algoritmo implementado no iguala ni supera a los otros métodos presentados. Incrementar el tamaño del fragmento de la canción a analizar no mostró una mejora significativa, y tomó una cantidad cada vez más grande de tiempo en finalizar.



Para el desglose por género musical, la tabla de comparación es la siguiente:

Género	BCG	PLS	SCR	SP	AC	SP-SEF	AC-SEF
Clásica	40,0	46,0	46,2	48,2	70,8	71,5	82,4
Jazz	50,0	57,0	70,9	62,0	69,8	78,4	86,0
Latina	18,2	70,3	81,1	62,1	70,3	91,8	94,5
Pop	20,0	57,5	70,0	75,0	85,7	92,5	92,5
Rock	55,6	40,9	84,1	61,3	84,4	81,8	88,6
Reggae	40,0	76,7	86,7	86,6	76,9	96,6	100
Soul	50,0	50,0	87,5	70,8	76,7	100	100
Rap	30,0	75,0	85,0	75,0	56,5	100	100
Techno	40,0	69,6	56,3	65,2	95,0	95,6	100

4. Conclusiones

Si bien no se logró igualar o mejorar la certeza de los métodos implementados por otros investigadores, este trabajo contribuyó a una comprensión más profunda del proceso de estimación de tempo en grabaciones de audio y la puesta en marcha de un algoritmo en Octave. Además pudimos aplicar los conocimientos adquiridos en el cursado de la materia Procesamiento Digital de Señales, para poder interpretar las señales de audio en el software de calculo Octave, establecer ventanas de procesamiento, tanto como en el dominio del tiempo como en el dominio de las frecuencias mediante el cálculo de la Transformada Discreta de Fourier y realizar un filtrado de los valores de la señal

Consideramos que no se logró una buena certeza en cuanto al reconocimiento de tempo debido al umbral determinado al momento de definir los picos significativos, el 50 % del valor máximo descartaba muchos picos, lo cual proporcionaba un buen tiempo de ejecución pero un bajo nivel de certeza. También puede deberse a la mala selección del fragmento de audio. Es posible que en ese intervalo de tiempo, el tempo varíe entre varios valores, haya pausas largas en la señal, o la canción no tenga eventos significativos en cada beat, como puede suceder con un solo de violín.

Referencias

[1] J. Laroche. Efficient tempo and beat tracking in audio recordings. In *J. Audio Eng. Soc*, 2003. vol. 51, no 4, p. 227.

[2] J. Laroche. Estimating tempo, swing and beat locations in audio recordings. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001. Prosody and Emotion 5.

[3] M. Alonso. Tempo and beat estimation of musical signals. In *Int. Soc. for Music Information Retrieval Conf*, 2004. Prosody and Emotion 5.