Sociedad de Ingeniería de Audio

Artículo de Congreso

Congreso Latinoamericano de la AES 2011
30 de Agosto a 1º de Septiembre de 2011
Montevideo, Uruguay

Análisis y Estimación de la Frecuencia Fundamental de la Melodía Principal en Piezas Musicales Reales

Haldo Spontón,[1] Gonzalo Gini[1] y Pablo Soubes[1]

[1] Facultad de Ingeniería, Instituto de Ingeniería Eléctrica, UdelaR
Montevideo, Uruguay

haldos@fing.edu.uy, elgongogini@hotmail.com, pabloss12@gmail.com

RESUMEN

La detección de melodía principal es uno de los problemas clásicos dentro del estudio de información musical. Con este objetivo, se dividió el problema en cuatro etapas que se detallan en este documento. En primer lugar, se realiza una estimación local de candidatos a frecuencia fundamental para cada fragmento o *frame* de audio. Luego, para estos candidatos calculados, se estudian y calculan algunas características que permitan aportar más información sobre los mismos. En una tercera etapa, se realiza el seguimiento temporal de posibles líneas melódicas, utilizando como entrada los resultados de las primeras dos etapas, y técnicas de filtrado adaptivo y programación dinámica. Por último, se utilizan técnicas básicas de síntesis sonora, de forma de mostrar los resultados obtenidos de una manera más amigable.
Se presentarán los porcentajes de acierto y error, comparando contra una base de datos prueba ó *Ground Truth*, tomada de *MIREX* (Music Information Retrieval eXchange).

INTRODUCCIÓN

La detección de melodía principal es uno de los problemas clásicos dentro del estudio de extracción de información musical [1, 2][1]. Consiste en obtener una representación simbólica de la melodía principal de una pieza musical de forma automática.

El interés por un sistema capaz de estimar la melodía principal de una pieza musical surge como una de las ideas básicas dentro del estudio musicológico, obteniendo una representación simbólica a partir de una señal de audio, manteniendo los aspectos perceptualmente significativos de dicha señal.

Objetivo

Como resultado se pretende obtener una estimación del valor frecuencial de la altura de la melodía principal para cada instante de tiempo de las señales de entrada. Además, se determina dónde la melodía principal

[1] MIR, Music Information Retrieval.

no está presente. Estos resultados se devuelven en forma de vector, guardando la información temporal correspondiente.

DESCRIPCIÓN DEL TRABAJO

Teniendo en cuenta los objetivos planteados, se dividió el problema en cuatro etapas:

- **Detección local de frecuencia fundamental:** Consiste en estimar para cada instante de tiempo los posibles valores que toma la frecuencia fundamental de la melodía principal. Esta estimación se realiza de manera local, ignorando lo que sucede en los instantes anteriores o posteriores.
- **Extracción y evaluación de características:** Una vez finalizada la estimación local, se obtienen candidatos a frecuencia fundamental para cada instante de tiempo. Se proponen diferentes características de los candidatos que permiten compararlos. Dichas características serán herramienta para poder vincular candidatos en diferentes instantes de tiempo.
- **Seguimiento temporal:** Se aplican técnicas de seguimiento temporal para encontrar tramos de melodía principal, utilizando el resultado de la estimación local de frecuencia fundamental y las características calculadas.
- **Aplicación (Síntesis Sonora):** Por último se crean archivos de audio sintetizado para mostrar de manera "amigable" al usuario los resultados de los diferentes algoritmos.

ESTIMACIÓN LOCAL DE CANDIDATOS A F0

En [3] se proponen estimadores conceptualmente simples de frecuencia fundamental $F0$ para señales de música polifónica, y se ofrece además una implementación computacionalmente eficiente para éstos. Dichos estimadores calculan la fuerza de un candidato a $F0$ como una suma ponderada de las amplitudes de sus armónicos parciales.

Se define *saliencia* o fuerza de un candidato a $F0$ en base a la siguiente ecuación:

$$s(\tau) = \sum_{m=1}^{M} g(\tau, m)|Y(f_{\tau,m})| \qquad (1)$$

donde $f_{\tau,m}$ es la frecuencia del m-ésimo parcial de un candidato con frecuencia fundamental f_τ.

El valor τ representa el período para el cual se va a calcular la saliencia, medido en número de muestras. Por lo tanto, el período medido en segundos sería $\tau_{seg} = \frac{\tau}{f_s}$, donde f_s es la frecuencia de muestreo. Entonces, calcular la saliencia para un período en muestras τ equivale a calcular la saliencia para una frecuencia $f_\tau = \frac{1}{\tau_{seg}} = \frac{f_s}{\tau}$. Por último, la frecuencia del m-ésimo parcial de la frecuencia f_τ es $f_{\tau,m} = \frac{m f_s}{\tau}$.

La función $g(\tau, m)$ define el peso del m-ésimo parcial en la suma. $Y(f)$ es el espectro normalizado de la señal de audio que entra al algoritmo. Esa normalización es resultado de un proceso llamado "blanqueado espectral". Utilizando una gran cantidad de material de entrenamiento, se propone además una forma paramétrica para esta función $g(\tau, m)$.

Blanqueado Espectral

Con el objetivo de lograr que el sistema sea robusto frente a fuentes de sonido diferentes se trata de suprimir la información de "timbre" del sonido antes de estimar $F0$. Esto se logra a través del blanqueado espectral el cual estima la distribución en frecuencia de la energía de una señal y la aplana o empareja mediante filtrado inverso.

Se calcula la FFT $X(k)$ de la entrada $x[n]$ en un frame de tiempo enventanado (con ventana Hanning). Luego se simula un banco de 30 filtros pasa-banda donde cada sub-banda tiene una respuesta triangular en amplitud (figura 1).

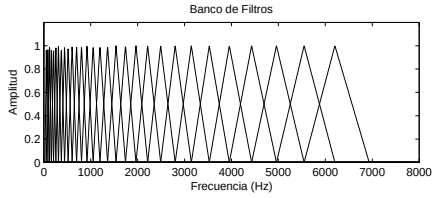

Figura 1: Banco de filtros simulado para el blanqueado espectral.

Las desviaciones estándar σ_b de cada sub-banda son calculadas aplicando el filtro $H_b(k)$ a la señal, donde K es el largo de la transformada de Fourier:

$$\sigma_b = \sqrt{\frac{1}{K}\sum_k H_b(k)|X(k)|^2} \qquad (2)$$

A partir de estos σ_b se calculan los coeficientes de compresión $\gamma_b = \sigma_b^{v-1}$ donde v es la cantidad de blanqueado espectral aplicado. Estos coeficientes γ_b, que representan una cierta inversa de la envolvente espectral estimada con los σ_b, son interpolados linealmente para obtener coeficientes de compresión $\gamma(k)$ para cada bin de frecuencia k. El espectro blanqueado se obtiene multiplicando el espectro de la señal de entrada por los coeficientes de compresión, es decir $Y(k) = \gamma(k)X(k)$. En la figura 2 se puede ver un espectro de ejemplo antes y después del proceso de blanqueado espectral.

Cálculo de Saliencia

Se utiliza un algoritmo de bipartición para estimar el período τ de mayor saliencia en cierto intervalo de interés. Este algoritmo no calcula la saliencia en todo el intervalo, sino que mediante bipartición estima

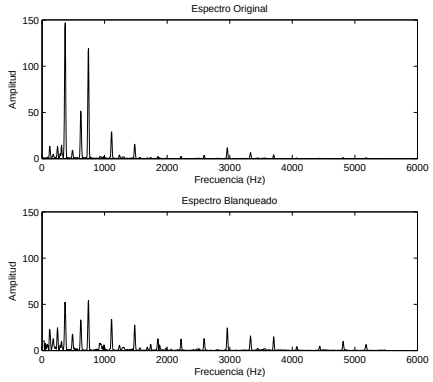

Figura 2: Ejemplo de espectro antes y después del proceso de blanqueado espectral.

en qué mitad del intervalo se encuentra el τ de mayor saliencia. Luego esta mitad se toma como un nuevo intervalo entero y se estima en cuál de sus mitades se encuentra el τ de mayor saliencia. Esto se repite hasta alcanzar la precisión deseada y con esto determinar el τ de mayor saliencia promediando los extremos del último intervalo. Por tanto se obtiene como salida el candidato τ con mayor saliencia $s(\tau)$.

En la figura 3 se puede apreciar un pseudo-código de este algoritmo, extraído de [3].

Algorithm 1: Fast search of the maximum of $\hat{s}(\tau)$

$Q \leftarrow 1$; $\tau_{\text{low}}(1) \leftarrow \tau_{\min}$; $\tau_{\text{up}}(1) \leftarrow \tau_{\max}$; $q_{\text{best}} \leftarrow 1$;
while $\tau_{\text{up}}(q_{\text{best}}) - \tau_{\text{low}}(q_{\text{best}}) > \tau_{\text{prec}}$ **do**
  # Split the best block and compute new limits
  $Q \leftarrow Q + 1$
  $\tau_{\text{low}}(Q) \leftarrow (\tau_{\text{low}}(q_{\text{best}}) + \tau_{\text{up}}(q_{\text{best}}))/2$
  $\tau_{\text{up}}(Q) \leftarrow \tau_{\text{up}}(q_{\text{best}})$
  $\tau_{\text{up}}(q_{\text{best}}) \leftarrow \tau_{\text{low}}(Q)$
  # Compute new saliences for the two block-halves
  for $q \in \{q_{\text{best}}, Q\}$ **do**
    Calculate $s_{\max}(q)$ using Equations (3)-(4)
    with $g(\tau, m) = \frac{f_s/\tau_{\text{low}}(q)+\alpha}{m f_s/\tau_{\text{up}}(q)+\beta}$
      $\tau = (\tau_{\text{low}}(q) + \tau_{\text{up}}(q))/2$
      $\Delta\tau = \tau_{\text{up}}(q) - \tau_{\text{low}}(q)$
  end
  # Search again the best block
  $q_{\text{best}} \leftarrow \arg\max_{q \in [1,Q]} s_{\max}(q)$
end
Return $\hat{\tau} = (\tau_{\text{low}}(q_{\text{best}}) + \tau_{\text{up}}(q_{\text{best}}))/2$
  $\hat{s}(\hat{\tau}) = s_{\max}(q_{\text{best}})$

Figura 3: Pseudo-código del Algoritmo de Cálculo de Saliencia.

Se implementó además una estimación iterativa y cancelación donde cada candidato detectado es eliminado de la mezcla y $s(\tau)$ es actualizado correspondientemente antes de estimar el siguiente $F0$.

EXTRACCIÓN DE CARACTERÍSTICAS

Se utiliza como característica principal una variante de la saliencia, eliminando la dependencia inversa con el número de parcial m. Se vio que este cambio hace que disminuyan sensiblemente algunos errores de detección como errores de octavas. En resumen, la saliencia como característica (y no como función de costo a maximizar) se calcula como:

$$s(f) = \sum_{m=1}^{M} \frac{f+\alpha}{f+\beta} |Y(mf)| \qquad (3)$$

donde se sustituyó $g(f, m)$ por $\frac{f+\alpha}{f+\beta}$.

La saliencia es una función que depende del tiempo y de la frecuencia en que se esté calculando. Esto hace que su resolución dependa de la resolución de la representación en tiempo-frecuencia del audio disponible, la cual está fijada por el largo de la Transformada de Fourier utilizada para calcular el espectrograma de cada pieza musical. Para obtener mayor resolución, se propone la "saliencia interpolada". Ésta se calcula usando como entrada un espectrograma interpolado en frecuencia (se aumenta la cantidad de bins del espectrograma mediante interpolación lineal). En la figura 4 se ve un ejemplo de mapa de saliencia utilizando un espectrograma interpolado.

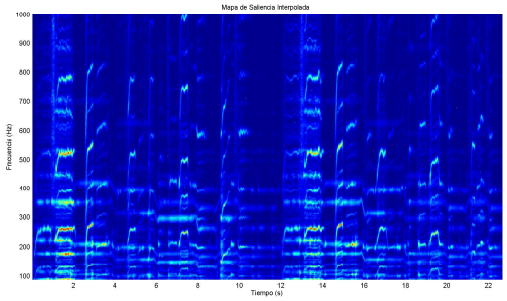

Figura 4: Mapa de saliencia interpolada del archivo $pop1.wav$.

SEGUIMIENTO TEMPORAL (TRACKING)

Lo que se busca entonces con el seguimiento temporal es aprovechar esa estrecha relación entre la frecuencia fundamental de la melodía principal en un frame y la de los siguientes o anteriores, así como la continuidad temporal de la saliencia. Para esto se estudiaron dos técnicas: Filtro de Kalman y Programación Dinámica.

Filtro de Kalman

Se diseñó un filtro de orden 5, utilizando el Ground Truth para entrenar los coeficientes del mismo. Una vez completa la estimación local de $F0$ y la extracción de características, se busca un punto de arranque para el tracking en base a buscar un candidato que tenga mayor saliencia como característica. A partir de este punto se aplica Kalman hacia atrás y hacia adelante, buscando en cada paso candidatos en un entorno de la predicción. Se van uniendo así diferentes candidatos para formar tramos de melodía. El algoritmo consta además

con condiciones de parada de forma de intentar detectar principios y finales de líneas melódicas. Luego se busca un nuevo punto de arranque, y se continúa el proceso hasta que se cumpla alguna de las condiciones de parada globales.

En la figura 5 se ve el resultado de la estimación de frecuencia fundamental de la melodía principal luego de aplicarse el filtro de Kalman discreto diseñado. Se ve como se reducen notablemente los errores que se mencionan anteriormente, lográndose un resultado que varía suavemente, muy similar al Ground Truth.

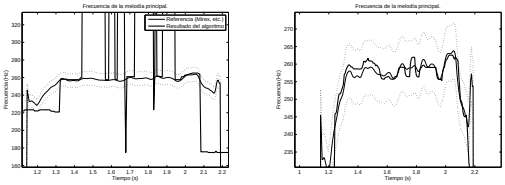

Figura 5: Análisis de desempeño del filtro de Kalman, archivo $pop1.wav$.

Programación Dinámica

La programación dinámica [4] reune un conjunto de técnicas que permiten determinar de manera eficiente las decisiones que optimizan el comportamiento de un sistema que evoluciona a lo largo de una serie de estados. En nuestro problema parece intuitivo reconocer a los diferentes instantes de tiempo como etapas y a los candidatos a $F0$ en cada instante junto con su saliencia asociada como los estados.

El algoritmo se compone de los siguientes módulos:

- Determinación de puntos seguros.
- Elaboración de candidatos.
- Seguimiento del mejor camino.

En síntesis dicho algoritmo debe seguir el mejor camino entre puntos seguros de la melodía a partir de la evaluación de ciertas características de los candidatos obtenidos anteriormente.

En la elección de la función de costo a utilizar, se busca que ésta:

1. Penalice saltos en frecuencia para frames consecutivos.
2. Mantenga continuidad en la saliencia.
3. Seleccione caminos de saliencia alta.

En la figura se aprecia un ejemplo del desempeño de PD para el seguimiento de melodía principal.

SÍNTESIS SONORA

La síntesis sonora se realiza de una manera simple, cambiando la fase de una función sinusoidal de manera de que la derivada de dicha fase sea igual a la frecuencia

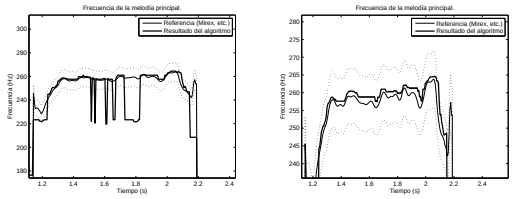

Figura 6: Análisis de desempeño de PD, archivo $pop1.wav$.

instantánea estimada en el sistema. Más formalmente, se crea una señal de audio $y(t)$ tal que:

$$y(t) = \alpha cos(\Phi(t)) \tag{4}$$

donde

$$\frac{d\Phi(t)}{dt} = 2\pi f(t) \tag{5}$$

siendo $f(t)$ la frecuencia instantánea estimada en el sistema y α un factor multiplicativo menor que 1 para evitar la saturación de la señal de audio sintetizada.

RESULTADOS

Se muestran los porcentajes de acierto (banda de $3\,\%$) de la estimación local de frecuencia fundamental de la melodía principal, utilizando la base de datos de prueba de MiReX:

	Primer candidato	Algún candidato
STFT	60,91 %	78,82 %
CQT	55,97 %	76,23 %

Cualitativamente, tanto el Filtro de Kalman como PD, ayudan a mejorar notoriamente el resultado primario de la estimación local de $F0$. Se logran resultados muy positivos en varias piezas musicales populares. Se pueden apreciar que los resultados son buenos utilizando el algoritmo de síntesis sonora.

REFERENCIAS

[1] Emilia Gómez, Anssi Klapuri, and Benoît Meudic, "Melody description and extraction in the context of music content processing," *Journal of New Music Research*, vol. 32, 2003.

[2] Haldo Spontón, Pablo Soubes, and Gonzalo Gini, "Melodía: Análisis y estimación de la melodía principal en piezas musicales reales," *Proyecto de Fin de Carrera, IIE, Facultad de Ingeniería.*, 2004.

[3] Anssi Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *in ISMIR*, 2006, pp. 216–221.

[4] Martin L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley-Interscience, April 1994.