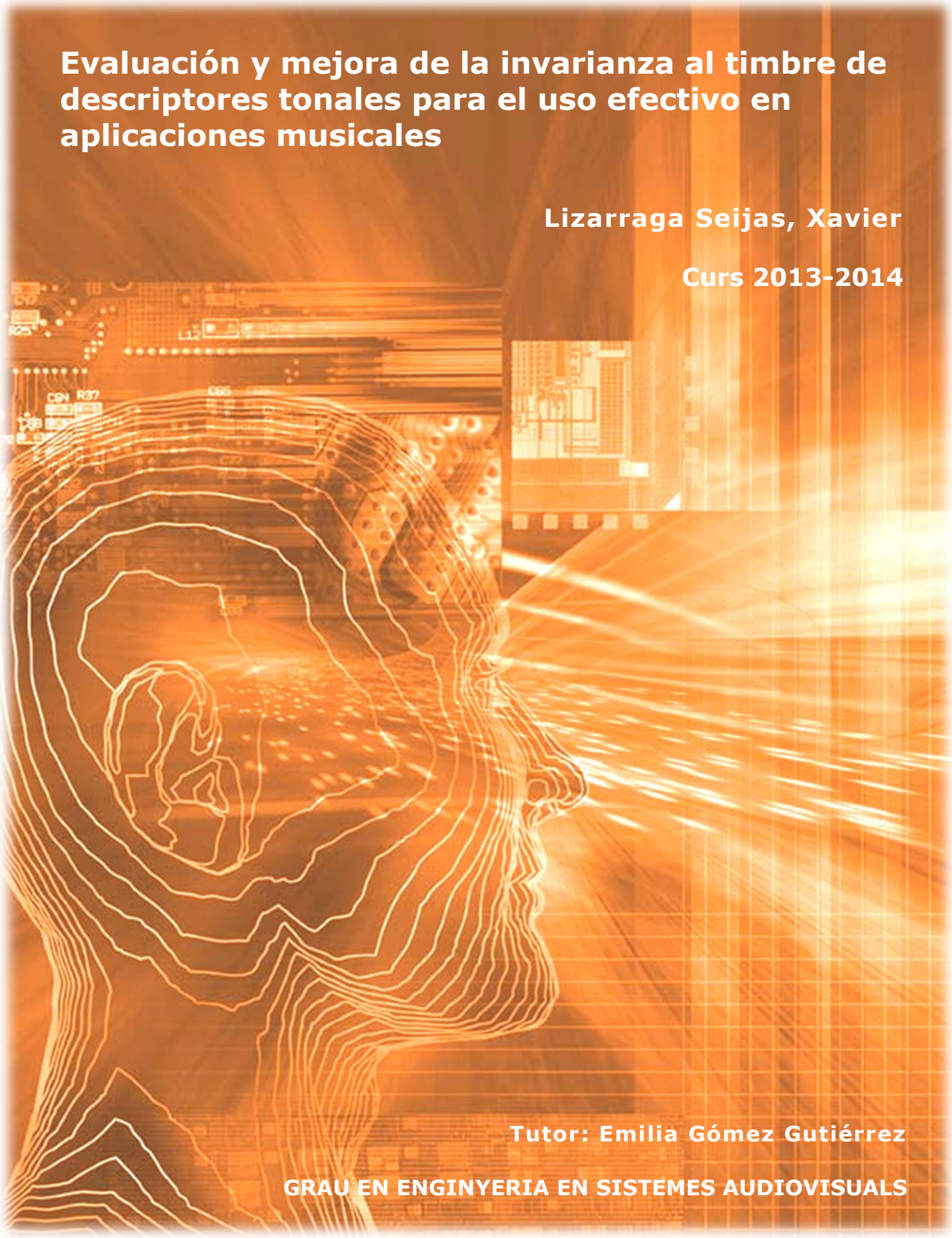
Evaluación y mejora de la invarianza al timbre de descriptores tonales para el uso efectivo en aplicaciones musicales

Lizarraga Seijas, Xavier

Curs 2013-2014



Tutor: Emilia Gómez Gutiérrez

GRAU EN ENGINYERIA EN SISTEMES AUDIOVISUALS

upf. Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Treball de Fi de Grau

Evaluación y mejora de la invarianza al timbre de descriptores tonales para el uso efectivo en aplicaciones musicales

Xavier Lizarraga Seijas

TRABAJO FINAL DE GRADO

GRADO DE INGENIERIA EN SISTEMAS AUDIOVISUALES

ESCUELA SUPERIOR POLITÉCNICA UPF

2013

DIRECTOR DEL TRABAJO

Emilia Gómez Gutiérrez

upf. Universitat Pompeu Fabra Barcelona
Escola Superior Politècnica

Para Montse,

Agradecimientos

Recuerdo que antes de saber que tema escogería para hacer el trabajo final de grado tenia claro que quería hacer algo sobre música, que implicara cálculo, procesado de señal, reconocimiento y adquisición de datos, visualización y programación en C++.

Después del trabajo realizado no tengo palabras suficientes para expresar lo agradecido que me siento a Emilia Gómez por ofrecerme la oportunidad de trabajar con ella y alcanzar mis expectativas. Por otro lado, me gustaría agradecerle sus enseñanzas, consejos y atención mostrada durante este trabajo y a lo largo de mi formación. Gracias por ser mi guía y corregirme.

Además me gustaría mencionar a Meinard Müller et al. por su colaboración y compromiso cediendo y permitiendo nos el uso de su colección de archivos de audio, sin ello hubiera sido difícil alcanzar los resultados conseguidos. A parte, me gustaría agradecerle la contribución de sus investigaciones y trabajos.

También recuerdo algunos momentos duros durante el desarrollo de las aplicaciones en los que no sabia que camino tomar o estaba atascado, es imposible olvidar la ayuda recibida de algunas personas en esos momentos. Por ello, estoy gratamente agradecido a Justin Salamon, por sus buenos consejos durante la implementación de los Vamp plugin y a Ricard Marxer por su ayuda y soporte durante el desarrollo de la estimación de tono con las frecuencias instantáneas, sin su colaboración no hubiera conseguido entender algunas ecuaciones y parámetros. Además me gustaría hacer mención y agradecer al Music Technology Group (MTG) y a sus miembros ofrecerme la posibilidad de trabajar en este proyecto. A parte me gustaría mencionar a varias personas que han contribuido en mi formación y que directa o indirectamente han contribuido a realizar este trabajo: Vicent Caselles, Waldo Nogueira, Sergi Jordà, Xavier Serra, Josep Blat, Xavier Binefa, Coloma Ballester, Marcelo Bertalmio, Daniel Arteaga, Jordi Bonada, Jordi Janer, Jordi Arqués, Enric Giné, Toni Mateos, Martin Haro, Giulio Cercale, Antoni Ivorra y Manel Jimenez.

También me gustaría mencionar a esas personas que me han acompañado a lo largo de esta larga empresa: José Emilio Lavilla, José L. Diez Antich, David Sánchez Blancas, Gerard Llorach, Patricia Vitoria y a muchos otros compañeros con los que he tenido el placer compartir buenos momentos. Por todos ellos.

En un sentido más personal, voy a estar eternamente agradecido con mi compañera, Montse Streit, por su enorme paciencia, consideración y cariño. Evidentemente, sin ella a mi lado no hubiera sido posible. Como no, mencionar y agradecer a mi querida familia: Merche, Eduardo, Mireia, Albert, Roser, Lluisa, Ramón y a tantos otros que siempre están en mis recuerdos. Por ultimo, darle las gracias a Jordi Funes por: su confianza y paciencia.

Resumen

La mayor parte de las aplicaciones musicales precisan de mecanismos inteligentes que nos permitan interactuar con el contenido musical de forma eficiente y ordenada. Una gran parte de los procedimientos automáticos utilizados en aplicaciones como la identificación o recomendación de música o el análisis de estructura o acordes se basan en la extracción automática de descriptores de croma, representativos del contenido tonal, mediante el análisis de señales musicales. Sin embargo, dichos métodos automáticos muestran ciertas deficiencias en presencia de ruido y cambios de timbre que acaban introduciendo errores en el sistema.

Principalmente, este trabajo trata sobre tonalidad y timbre. Concretamente, se centra en el descriptor de croma Harmonic Pitch Class Profile HPCP (Emilia Gómez, 2007, UPF). Nuestro objetivo principal es estimar su grado de variación al timbre y su efectividad respecto a otros métodos, además de evaluar varios procesos que proponemos para mejorar su rendimiento.

La estimación de la variación al timbre y la efectividad, se basa en el análisis de los cromagramas de 298 combinaciones de notas, interpretadas con diferentes instrumentos. Los métodos propuestos para mejorar el rendimiento son: la estimación de tono mediante las frecuencias instantáneas y la ecualización tímbrica basada en el filtrado cepstral. Además, presentamos diferentes experimentos con ambos métodos, en MATLAB y C++, que muestran la dependencia en los parámetros que controlan aspectos espectrales y su influencia en la extracción del croma.

Los resultados obtenidos de la evaluación del cromagrama nos permiten cuantificar y visualizar propiedades relacionadas con el timbre y la tonalidad. De esta forma conseguimos clasificar los métodos de extracción de cromagramas, según su efectividad. Los experimentos con los métodos de mejora que proponemos, concluyen en que la estimación del tono a partir de la frecuencia instantánea es un método eficaz, pero poco eficiente para el análisis de señales de audio polifónicas. Además muestra una gran dependencia de sus parámetros y un alto coste computacional. Sin embargo, los experimentos realizados sobre el filtrado de coeficientes cepstrum, muestran la posibilidad de modificar las modulaciones tímbricas, con un comportamiento similar al de una función de blanqueado espectral. Principalmente, se han analizado dos técnicas: la sustitución de coeficientes cepstrum por ceros (zeroing) y el filtrado de los coeficientes mas bajos mediante un filtro pasa altos.

Con los resultados obtenidos se han desarrollado dos plugin Vamp multi-plataforma: la tercera versión del HPCP y el espectrograma basado en la frecuencia instantánea, IF Spectrogram. Obviamente, estas dos aplicaciones finales quedarán a disposición de la comunidad para el análisis de señales de audio.

Abstract

Most music applications require intelligent mechanisms that allow us to interact efficiently and orderly with the musical content. A large part of the automatic procedures used in applications such as identification or music recommendation or analysis of structure or chords are based on the automatic extraction chroma features, representing tonal content, through the analysis of musical signals. However, these automated methods show some deficiencies in noise and timbre changes that introduce errors into the system.

Mainly, this work is about tonality and timbre. Specifically, it focuses on the chroma descriptor HPCP Harmonic Pitch Class Profile (Emilia Gómez, 2007, UPF). Our main objective is to estimate the degree of timbre invariance and its effectiveness over other methods, in addition to evaluating various processes we propose to improve its performance.

The appraisal of the timbre invariance and effectiveness, are based on the chroma features analysis of 298 combinations of notes, played with different instruments. The proposed methods for improving performance are the pitch estimation based on instantaneous frequencies and timbre equalization based on cepstral filtering. In addition, we present several experiments with both methods in MATLAB and C++, which show the dependence on the parameters that control spectral aspects and their influence on chroma extraction.

The results of the evaluation of croma approaches allow us to quantify and display properties related to the timbre and tonality. In this way we classify chroma extraction methods, depending on its effectiveness. Experiments with improved methods that we propose, conclude that the pitch estimation with the instantaneous frequency is effective, but not very efficient for the analysis of polyphonic audio signals. It also shows a considerable dependence of its parameters and a high computational cost. However, experiments with cepstrum filtering, show the possibility of modifying the timbre modulations, with a behavior similar to a spectral whitening. Mainly, two techniques have been analyzed with cepstrum: replacing coefficients by zeros and filtering using Gaussian function.

With the obtained results we have developed two multi-platform plugin Vamp: the third version of HPCP and spectrogram based on the instantaneous frequency, IF Spectrogram. Obviously, these two applications will be available to the community for the analysis of audio signals.

Indice

Resumen vii

Abstract ix

Listado de figuras xiv

Listado de tablas xvi

1. INTRODUCCIÓN Y MOTIVACIÓN 1

1.1 Música en formato digital y tecnologías de descripción . . . 1
1.2 Descripción de audio y niveles de descripción . . . 4
1.2.1 Descriptores de bajo nivel y su extracción . . . 5
1.2.2 Descriptores rítmicos y tonales . . . 7
1.3 Música y descriptores tonales . . . 8
1.3.1 Descriptores tonales invariantes al timbre . . . 11
1.4 Objetivos: evaluar y mejorar este tipo de descriptores . . . 13
1.5 Estructura del trabajo. . . . 14

2. DEFINICIONES, TÉCNICAS Y APLICACIONES 15

2.1 Introducción . . . 15
2.2 Definiciones . . . 15
2.2.1 Tono . . . 15
2.2.2 Sonido armónico y frecuencia fundamental . . . 16
2.2.3 Timbre . . . 17
2.2.4 Sonido polifónico . . . 18
2.2.5 Frecuencia de referencia . . . 18
2.2.6 Escala Mel y escala musical . . . 19
2.2.7 Escala temperada y escala cromática . . . 20
2.2.8 Análisis espectral . . . 20
2.2.8.1 Función ventana . . . 21
2.2.8.2 Rellenado de ceros . . . 22
2.2.8.3 DFT . . . 22
2.2.8.4 Coeficientes cepstrum . . . 23
2.3 Aplicaciones del cromagrama . . . 24
2.3.1 Análisis musical: estimación de acordes y tonalidad . . . 24
2.3.2 Similitud: detección de versiones o covers . . . 24
2.3.3 Clasificación automática. . . . 25
2.4 Algoritmos para la extracción de cromagrama: HPCP y CRP . . . 25

2.4.1 Harmonic Pitch Class Profile: HPCP . . . 25
2.4.1.1 Pre- procesado . . . 26
2.4.1.2 Análisis espectral . . . 26
2.4.1.3 Detección de picos espectrales: detección parabólica . . . 26
2.4.1.4 Cálculo del HPCP . . . 27
2.4.1.5 Normalización . . . 29
2.4.2 Chroma DCT Reduced Log Pitch: CRP . . . 30
2.4.2.1 Descomposición en bandas tonales . . . 30
2.4.2.2 Energía local y compresión logarítmica . . . 31
2.4.2.3 Cepstrum DCT . . . 31
2.4.2.4 Rellenado de ceros cepstral y inversa de la DCT . . . 31
2.4.2.5 Correspondencia tono a croma . . . 31
2.4.2.6 Normalización euclídea . . . 32
2.5 Objetivos y resultados esperados . . . 32
2.5.1 Evaluación de los descriptores tonales . . . 32
2.5.2 Mejoras y explotación en un contexto . . . 32
2.5.3 Implementación: de C++ a MATLAB . . . 33
2.5.4 Retos transversales . . . 34

3. METODOLOGIA **35**

3.1 Comparativa de aproximaciones para el cálculo de croma . . . 35
3.1.1 Colección de sonidos . . . 35
3.1.2 Medidas de evaluación . . . 37
3.1.2.1 Grado de varianza al timbre . . . 37
3.1.2.2 Grado de eficiencia . . . 38
3.2 Método propuestos . . . 39
3.2.1 Filtrado de coeficientes cepstrum . . . 40
3.2.2 Estimación del tono mediante la frecuencia instantánea . . . 41
3.3 Mejora de invarianza al timbre . . . 43

4. RESULTADOS **45**

4.1 Resultados de la evaluación del HPCP y el CRP . . . 45
4.1.1 Estimación del grado de varianza al timbre . . . 45
4.1.2 Estimación del grado de eficiencia . . . 45
4.2 Resultados de los métodos propuestos . . . 52
4.2.1 Experimentos en MATLAB . . . 52
4.2.1.1 Experimentos con el filtrado de coeficientes cepstrum . . . 52

4.2.1.2 Experimentos para el cálculo de las frecuencias instantáneas . . . 54
4.2.2 Implementación de Vamp plugins en C++. 56
4.2.2.1 HPCP 3.0 56
4.2.2.2 Espectrograma IF 59
4.3 Evaluación de la nueva versión del HPCP 63
4.4 Usuarios de prueba 66

5. CONCLUSIONES **69**

Bibliografía **71**

ANEXO **75**

I. Experimentos con el filtrado de coeficientes cepstrum 75

II. Experimentos para el cálculo de las frecuencias instantáneas 78

III. Archivos digitales 87

Listado de Figuras

Figura 1. Evolución de las ventas de música digital vs formato físico en EEUU 1
Figura 2. Ejemplo del sistema de indexación automática de Google 2
Figura 3. El sistema de indexación automática de Freesound.org 3
Figura 4. Sistema de identificación de contenido musical 3
Figura 5. Proceso de digitalización de una señal sinusoidal 5
Figura 6. Esquema de la extracción de descriptores de bajo nivel 5
Figura 7. Descriptor de bajo nivel: energía o RMS 6
Figura 8. Representación grafica de un *onset* 7
Figura 9. Cromagrama del acorde C 8
Figura 10. Relación frecuencia fundamental y tono musical de la octava del C_4 central 9
Figura 11. Relaciones cromáticas 10
Figura 12. La escala cromática 10
Figura 13. Descriptor tonal HPCP y sus diferencias tímbricas 11
Figura 14. El espectro de la frecuencia de dos instrumentos 12
Figura 15. Señal sinusoidal de 1 KHz en el dominio temporal y frecuencial 15
Figura 16. Señal de audio polifónica en el dominio temporal y frecuencial 16
Figura 17. Magnitud del espectro de un sonido armónico 17
Figura 18. Muestra la magnitud del espectro con diferentes instrumentos 18
Figura 19. Relación entre tono musical y frecuencia fundamental 18
Figura 20. Relación entre frecuencias y escala Mel 19
Figura 21. Diagrama del análisis espectral 20
Figura 22. Un fragmento de una señal de audio en el calculo de la STFT 21
Figura 23. Función ventana Blackman 21
Figura 24. Cálculo de la FFT 22
Figura 25. Diagrama del cálculo de coeficientes cepstrum 23
Figura 26. Cepstrum vs magnitud del espectro 23
Figura 27. Cromagrama de un acorde de C mayor 24
Figura 28. Esquema del proceso de extracción del HPCP 25
Figura 29. Detección de picos espectrales sobre la magnitud del espectro 27
Figura 30. Función de pesos del HPCP 28
Figura 31. Esquema de la extracción del CRP 30
Figura 32. Magnitud en decibelios del banco de filtros del CRP 30
Figura 33. La energía local del CRP 36
Figura 34. Esquema de la evaluación de los descriptores tonales 37
Figura 35. El perfil acorde del acorde C sin normalizar 38
Figura 36. Perfil tonal del acorde C mayor normalizado 39
Figura 37. Función gaussiana definida por el parámetro alfa 41

Figura 38. Frecuencia vs. la frecuencia instantánea de una señal de voz 42
Figura 39. Resultados de la evaluación del HPCP por instrumentos 46
Figura 40. Resultados del HPCP con un acorde Am de un sonido MIDI de saxo 46
Figura 41. Resultados de la evaluación del CRP por instrumentos 47
Figura 42. Diferencias entre el espectro original y procesado cepstrum 47
Figura 43. Resultados de la evaluación por acorde del HPCP y CRP(I) 48
Figura 44. Resultados de la evaluación por clase acorde del HPCP y CRP (II) 48
Figura 45. Gráfico de cajas de la correlación de las clases acorde del HPCP 49
Figura 46. Gráfico de cajas de la correlación de las clases acorde del CRP 50
Figura 47. Ampliación de la vista la correlación de las clases acorde (duetos) 51
Figura 48. Resultados del CRP para la clase acorde 88 con diferentes instancias 52
Figura 49. Diferencias: espectro original y procesado cepstrum de un sonido polifónico 53
Figura 50. Rellenado de ceros cepstral como blanqueador espectral 53
Figura 51. Esquema del relleno de ceros de los coeficientes cepstrum más bajos 54
Figura 52. Esquema para la el cálculo de la frecuencia instantánea 55
Figura 53. Comparación del espectro original y el espectro procesado 57
Figura 54. Configuración del HPCP 3.0 en Sonic Visualizer 58
Figura 55. Resultados del croma de la primera melodía del Bolero de Ravel (I) 58
Figura 56. Resultados del croma de la primera melodía del Bolero de Ravel (II) 59
Figura 57. Resultados del croma de la primera melodía del Bolero de Ravel (III) 59
Figura 58. Configuración del Espectrograma IF en Sonic Visualizer 61
Figura 59. Resultados Espectrograma IF (I) 61
Figura 60. Resultados Espectrograma IF (II) 61
Figura 61. Resultados Espectrograma IF (II) 62
Figura 62. Resultados Espectrograma IF (IV) 62
Figura 63. Espectrograma IF de un sonido polifónico con tamaño de la FFT de 1024 62
Figura 64. Espectrograma IF de un sonido polifónico con tamaño de la FFT de 4096 63
Figura 65. Varianza al timbre para el cepstral liftering 63
Figura 66. Grado de eficiencia para el cepstral liftering (correlación) 64
Figura 67. Grado de eficiencia para el cepstral liftering (distancia coseno) 64
Figura 68. Varianza al timbre para el cepstral filtering 65
Figura 69. Grado de eficiencia para el cepstral filtering (correlación) 65
Figura 70. Grado de eficiencia para el cepstral filtering (distancia coseno) 66

Listado de tablas

Tabla 1. Funciones MATLAB implementadas para la evaluación de descriptos tonales 38
Tabla 2. Resultados de la estimación del grado de varianza al timbre 45
Tabla 3 Estimación del grado de eficiencia a partir del coeficiente de correlación 45
Tabla 4. Estimación del grado de eficiencia a partir de la distancia del coseno 46
Tabla 5. Clasificación cualitativa de las distancia en semitonos 50
Tabla 6. Resultados de la estimación del grado de varianza al timbre HPCP 3.0 63

1. Introducción y motivación

Este trabajo se centra en la extracción de características del contenido de audio, basados en descriptores de bajo nivel, como el cromagrama para contenido musical. En el informe tratamos de evaluar su eficiencia y su invarianza al timbre. Este documento propone una herramienta de evaluación de la identificación automática de contenido musical. Además proponemos algunos métodos para mejorar la extracción de características de bajo nivel basadas en el cromagrama que mejoren su invarianza al timbre. Estas mejoras pueden beneficiar a aplicaciones dedicadas a la descripción o identificación de archivos musicales. En los experimentos realizados mostramos su influencia en la extracción de estos descriptores tonales. También, mostramos los resultados obtenidos en una aplicación final de estos métodos, implementado en un Vamp Plugin (C++), y evaluaremos su aportación en la cadena de procesado del descriptor tonal HPCP [1]. En este capítulo trataremos de introducir el contexto en el que se realiza este trabajo, además de explicar la motivación que nos ha llevado a realizarlo. El capitulo concluye con los objetivos establecidos, junto con un resumen del trabajo realizado.

1.1 Música en formato digital: técnicas de descripción y indexación automática

Durante los últimos años, el desarrollo de la tecnología ha estado presente en todos los ámbitos. El sector audiovisual no ha sido un excepción, concretamente en el caso de la música, la digitalización ha supuesto un gran cambio en todos los sentidos. La digitalización de la música junto la proliferación de reproductores portátiles y la expansión de Internet, permite nuevas posibilidades de creación, venta y promoción. Hoy en día la venta de música digital online representa una buena parte de las ventas de música anuales. Portales dedicados exclusivamente a la venta de música ponen a nuestra disposición un total de 13 millones de canciones [2]. Además, la venta de música a la carta a precios accesibles ha generado una nueva atracción, aunque hoy por hoy, el fenómeno *streaming* se esta imponiendo y se sitúa como el más escogido por los usuarios. Los estudios presentados por la plataforma '*Global Enterteinment*' [3], muestran que la tendencia se mantendrá durante los próximos años y la venta de música disminuirá aunque la venta de música digital continuará creciendo.

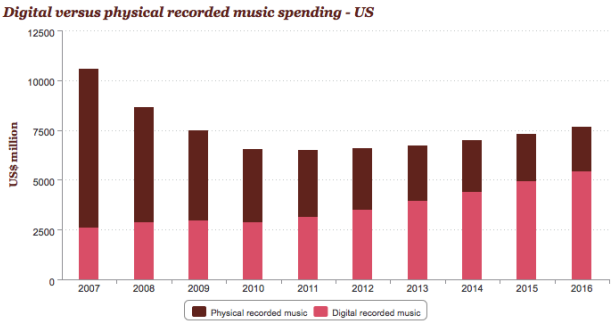


Figura 1. Evolución de las ventas de música digital vs formato físico en EEUU[1].

[1] http://www.pwc.com/gx/en/global-entertainment-media-outlook/segment-insights/music.jhtml

Además hemos de considerar la proliferación de repositorios y colecciones de sonidos, gratuitas o de pago, que ofrecen sus contenidos en diferentes formatos de audio[2]. Todas estas aplicaciones precisan de mecanismos inteligentes, que nos permitan interactuar con el contenido musical de forma eficiente y ordenada, basándose en sistemas de reconocimiento automático. Gracias a Internet, un gran numero de colecciones y aplicaciones musicales se encuentran disponibles desde cualquier parte del mundo. La oferta de estas aplicaciones se ha multiplicado y la tendencia indica que continuara creciendo en los próximos años. Ante este escenario, cada día toman más importancia los sistema de reconocimiento capaces de clasificar e indexar colecciones de música digital automáticamente y que respondan con cierta eficiencia las peticiones de los usuarios. El crecimiento de las colecciones de música y la accesibilidad de la que disponemos mediante tabletas electrónicas, teléfonos inteligentes, ordenadores, televisión digital,… obliga al sector a continuar mejorando sus sistemas de indexación y a interesarse por la investigación y desarrollo de nuevos mecanismos inteligentes que mejoren la recuperación y identificación del contenido musical.

La indexación automática se define como "la selección de un conjunto de términos que representan un documento mediante un programa informático[3]" y es la herramienta más importante de un motor de búsqueda.

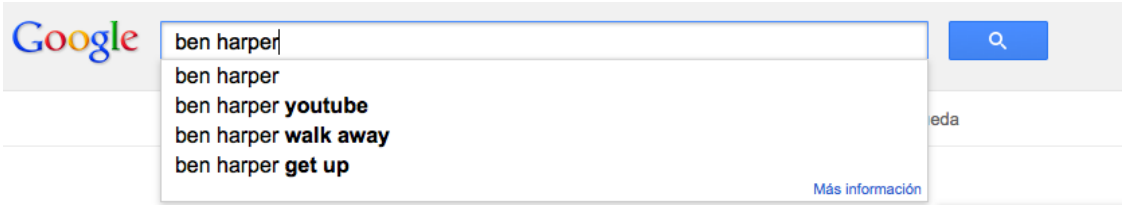


Figura 2. Ejemplo del sistema de indexación automática de Google[4]. El sistema de recomendación responde con un listado de conceptos relevantes y relacionados con el artista.

Una parte de las aplicaciones que precisan de indexación automática de archivos de audio, utilizan sistemas basados en la información que introduce el usuario mediante metadatos, anotaciones textuales y etiquetas[5] que describen el estilo o genero musical al que pertenece el contenido. Además, permiten incluir información relacionada con el ritmo, la escala musical y cualidades o emociones que podemos relacionar con el sonido, como se explica en [4]. En un principio, puede resultar un sistema atractivo para el usuario pero se trata de un sistema ineficiente para la identificación, ya que el etiquetado ofrece un grado de variación demasiado elevado entre los diferentes usuarios. Sin embargo, ofrece buenos resultados en sistemas de recomendación y buscadores web.

Generalmente, en la clasificación de colecciones de audio, además de los metadatos, se emplean estrategias basadas en el análisis y el procesamiento de la señal para extraer características o atributos capaces de describir el contenido por si mismos.

[2] www.freesound.org
[3] http://es.wikipedia.org/wiki/Indizaci%C3%B3n_autom%C3%A1tica
[4] https://www.google.com/
[5] Conocido como tag.

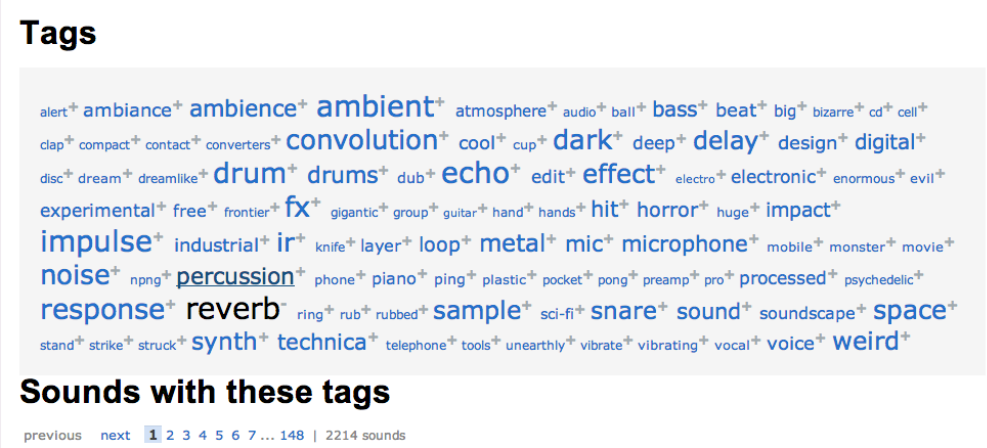


Figura 3. El sistema de indexación automática de Freesound.org. Incorpora un sistema de votaciones para mejorar su eficacia. Las etiquetas más relevantes se representan mediante la expresividad de la fuente tipográfica.

El secreto para conseguir una buena identificación de contenido musical es encontrar una huella sonora única [5], que permita distinguir una instancia entre las demás, cuando existe una petición en el sistema. La eficacia de un descriptor depende de que sus resultados sean robustos ante la presencia de ruido o posibles distorsiones que puedan introducirse en el sistema

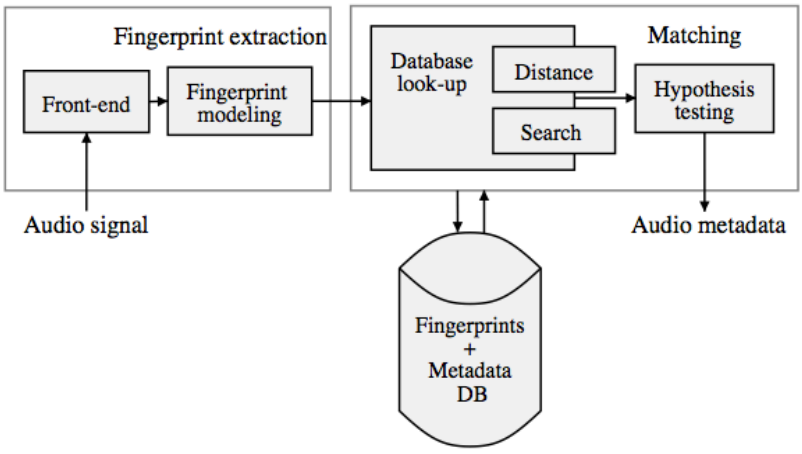


Figura 4. Estructura básica de un sistema de identificación de contenido musical.

Una gran parte de los procedimientos automáticos utilizados en el reconocimiento de música, se basan en el análisis tonal y en la extracción instantánea de una secuencia de características del material sonoro, como por ejemplo el sistema implementado en la aplicación Shazam [6]. Sin embargo, estos métodos muestran ciertas deficiencias en presencia de ruido y cambios de timbre que introducen errores en el sistema.

1.2 Descriptores de audio y niveles de descripción

La necesidad de disponer de técnicas o estrategias que permitan a los sistemas clasificar y identificar archivos de audio automáticamente, ha impulsado a una parte de la comunidad científica a desarrollar nuevos descriptores y continuar investigando en este ámbito.

Los descriptores de audio se clasifican en tres niveles de abstracción:

- Descriptores de bajo nivel
- Descriptores de medio nivel
- Descriptores de alto nivel

Los descriptores de bajo nivel se centran en aspectos de la señal y su análisis. Acostumbran a calcularse para cada fragmento de audio, con lo que aportan una secuencia, y aportan información instantánea de la señal. Mediante el análisis de estos descriptores podemos extraer información musical. Para los usuarios este tipo de descriptores aportan muy poca información, pero por otra banda son fáciles de implementar en un sistema informático.

Los descriptores de nivel medio se centran en el objeto o el documento en si, generalizando el contenido y etiquetándolo por genero musical o tonalidad. Este tipo de descriptores aportan al usuario información semántica del contenido sonoro. Para etiquetar el contenido necesitan aplicar un modelo de aprendizaje (HMM, KNN, GMM, SVM,…) que hay que entrenar previamente con un conjunto de datos [7].

Los descriptores de alto nivel se centran en el usuario final y tratan de describir su carácter o comportamiento mediante un modelo de aprendizaje. Se basan en descriptores de medio o bajo nivel extraídos del contenido musical relacionado con el usuario. A partir de ellos, se clasifica al usuario con uno de los modelos que ofrecen el sistema, como se sugiere en el documento [8]. Aunque se basan en descriptores de bajo nivel, no son útiles para etiquetar colecciones musicales, ya que describen su contenido de modo general.

Como ya hemos comentado, las huellas sonoras son los descriptores mas indicados para la identificación de contenido musical. Éstas se basan en la extracción de un conjunto de características del material sonoro o, lo que es lo mismo, en descriptores de bajo nivel. La extracción y modelado de las características musicales requieren de la unión de diferentes disciplinas: procesado del señal, inteligencia artificial, reconocimiento de la información y ciencia cognitiva.

1.2.1 Descriptores de bajo nivel y su extracción

La digitalización de una señal musical nos permite aplicar procesos para analizar su contenido de diferentes maneras. A partir de una señal de audio digital podemos extraer muchas características de bajo nivel. El procesado del señal nos provee de varias técnicas que pueden ser útiles para la extracción de estas características.

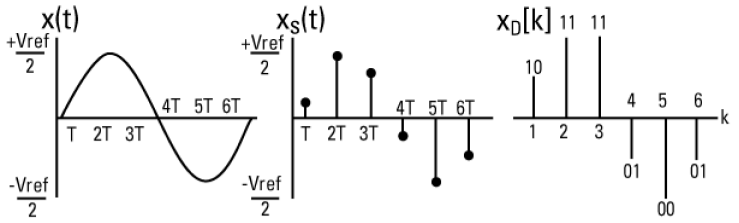


Figura 5. Este esquema representa el proceso de digitalización de una señal sinusoidal.

Podemos extraer características de un archivo de audio en función del tiempo o de la frecuencia. En el dominio temporal podemos asignar una característica por muestra, por fragmentos, por objeto. La extracción por muestra no es relevante, en cambio la extracción por fragmento, permite obtener un valor representativo para una parte de la señal o un valor global que representa el conjunto de sus muestras. De esta forma podemos asociar características a un fragmento del contenido musical.

En el dominio de la frecuencia, la unidad de análisis más pequeña es lo que se conoce como un fragmento (frame). Sobre estos fragmentos se aplican técnicas de procesado como la transformada de Fourier (FFT) o un banco de filtros que permiten obtener características del contenido en cada banda frecuencial. Además, el dominio de la frecuencia relaciona las tres dimensiones: tiempo, tono y magnitud. En realidad, la relación matemática que asocia los tres ejes es una variante de la FFT, conocida con la transformada de Fourier (STFT) en tiempos cortos, que acostumbra a utilizarse en audio para analizar la evolución de cambios locales en frecuencia y fase.

En el proyecto CUIDADO [9] se describen varios descriptores que se extraen siguiendo las asunciones anteriores y resume en el siguiente esquema:

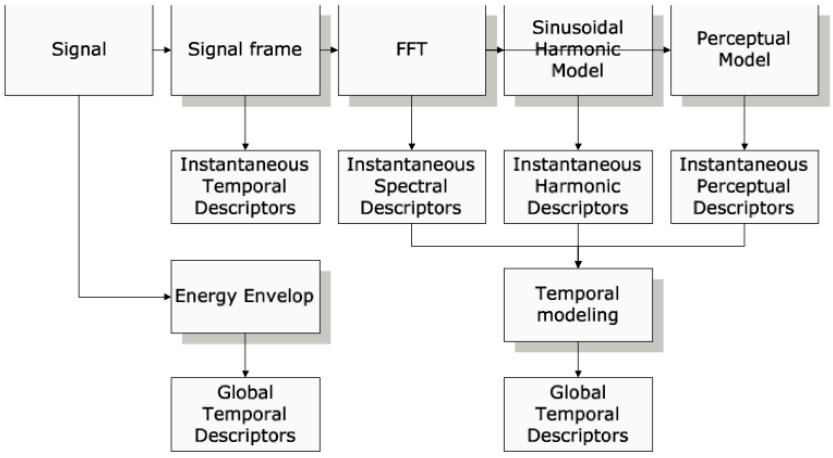


Figura 6. Esquema de la extracción de descriptores de bajo nivel (Peeters, 2004).

Siguiendo este esquema clasificamos los descriptores de bajo nivel como:

- Características de la forma de onda en el dominio temporal: son características que se extraen de la forma de onda o de su envolvente. Representen valores globales que describen o que hacen referencia a todo un archivo. Entre ellos destaca: Attack-Time Temporal o la energía del valor de pico.

- Características temporales: normalmente se asigna una característica a cada fragmento. Destacan: la auto-correlación, la covarianza, el ratio de cruces por cero, log-attack time y el centroide temporal.

  El centroide temporal es uno de los descriptores más destacados de las características temporales. Indica el centro de gravedad de un fragmento de audio en el dominio temporal.

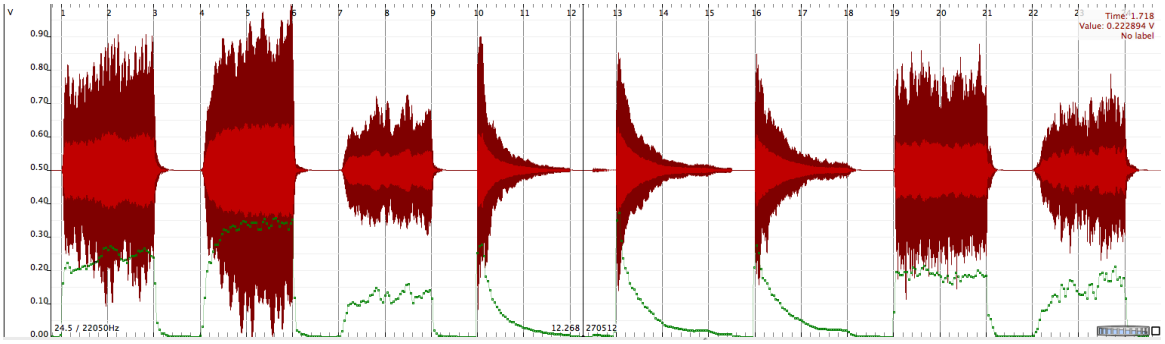


Figura 7. Descriptor de bajo nivel: energía o RMS.

La mayor gran parte de los descriptores mencionados en este trabajo están implementados en la aplicación Sonic Visualizer[6], registrada bajo licencia Open Source por el departamento de audio de la Universidad Queen Mary de Londres[7]. Esta aplicación esta pensada especialmente para el análisis de audio y la visualización de descriptores. Permite diferentes modos de visualización pero generalmente el valor del descriptor o de la característica se asocia al eje de ordenadas y el tiempo o fragmento al eje de abscisas. Además permite cargar librerías dinámicas externas conocidas como Vamp plugins y implementadas por terceros.

- Características energéticas: se centran en la energía de cada fragmento de audio. Algunos ejemplos son: energía RMS por fragmento, valor de pico, factor cresta, etc.

- Características de la forma de la envolvente espectral: son características instantáneas que se calculan para cada fragmento a partir del espectro obtenido en la STFT. Algunos ejemplos son: centroide espectral, coeficiente cepstrum en escala mel, etc.

[6] http://www.sonicvisualiser.org/
[7] http://c4dm.eecs.qmul.ac.uk/

- Características armónicas: se calculan para u modelado de la señal a partir del modelo sinusoidal o armónico (Serra, X. 1990). Algunos ejemplos son: ratio armónico/ruido, desviación armónica, etc.

Estas características aplican el modelo sinusoidal para extraer los picos del espectro de cada fragmento y buscar relaciones establecer medidas entre las frecuencia correspondientes a los picos detectados. La detección de pico se basa en la detección parabólica de picos desarrollada por Serra, X [10]. Pueden extraer características musicales si aplicamos ciertos procesos que relacionen la frecuencia con el tono musical, en los capítulos siguiente hablaremos sobre ello.

- Características perceptuales: utilizan un modelo perceptual que simula la percepción humana. Algunos ejemplos son: Sharpness, Spread, Spectral Flatness y factor cresta.

1.2.2 Descriptores rítmicos y tonales

Mediante los descriptores de bajo nivel podemos relacionar audio con diversas cualidades musicales como el ritmo y la tonalidad, analizando e introduciendo mecanismos inteligentes que permitan al sistema presentar estas características con cierto orden y elegancia. En la música existen cualidades esenciales como el tono y el ritmo y podemos clasificar los descriptores para música siguiendo esta distinción.

Los descriptores rítmicos se centran en la evolución de la energía a lo largo del tiempo y en la detección de transitorios que se asocian a intervalos de tiempo o a las frecuencias en que se detectan los picos de energía más destacados. El punto exacto donde comienza un transitorio se conoce como *onset*. Hay versiones tanto en el dominio temporal como en el de la frecuencia. De esta forma, podemos conocer aspectos rítmicos como el tempo o el compás de una composición musical.

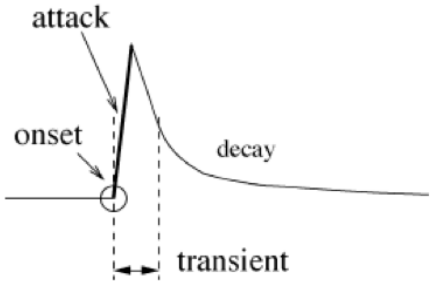


Figura 8. Representación grafica de un *onset*.

Las técnicas que se basan en el contenido en frecuencia permiten conocer eventos temporales, como los picos del espectro detectados en un fragmento de audio. No obstante, aportan información de las frecuencias o tonos que predominan durante ese instante. Lo que es muy relevante para el análisis musical, pues permite relacionar frecuencia y tiempo.

Los descriptores tonales hacen referencia a las notas, más en concreto al pitch (altura) representada por la frecuencia fundamental, su magnitud acústica relacionada. Acostumbran a describir el contenido tonal de una pieza o fragmento musical. Entre los descriptores de tonalidad, los más representativos son la extracción de melodía y el cromagrama. Los descriptores de melodía extraen la melodía principal de una pieza o fragmento musical, centrándose en la voz o en instrumentos predominantes como el bajo eléctrico.

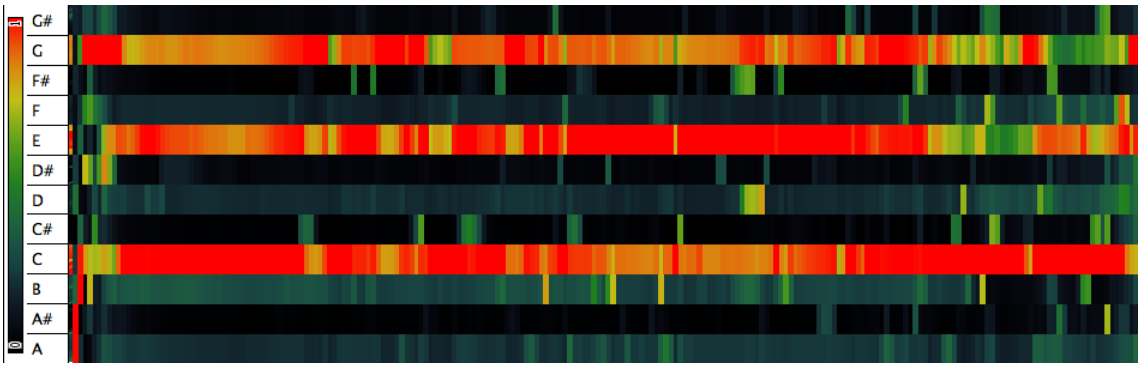


Figura 9. Cromagrama del acorde C.

El cromagrama es un descriptor de tonalidad que define la estructura tonal de un fragmento musical y define la intensidad relativa de los grados de una octava. Permite reducir la dimensionalidad de la escala musical proyectando la intensidad de cada una de las notas detectadas en una sola octava. El eje de ordenadas indica la tonalidad, que puede expresarse en semitonos o céntimos semitono[8], y el eje de abscisas indica su intervalo de tiempo. Ofrece una visión global del contenido musical y permite extraer información rítmica, tonal y de estructura. Además, permite identificar estrofas, puentes o partes solistas. Se basa en el análisis de la frecuencia para cada fragmento, extrayendo una secuencia de características relacionadas con el tono. De esta forma, ofrece una visión compacta y comprimida del contenido, muy valida para un sistema de indexación de música automática.

1.3 Música y descriptores tonales

El principal objetivo del análisis musical con descriptores es mostrar relaciones significativas entre diferentes extractos musicales de un conjunto de datos que acaben representando piezas musicales por si mismas. Estos descriptores se utilizan en tareas de sincronización, análisis, estructura musical, emparejamiento de música con un perfil similar y identificación de versiones o covers, donde se dan diferencias en el timbre y la instrumentación.

Como acabamos de ver, la información o características musicales que presenta la señal de audio están relacionadas con descriptores de bajo nivel, lo parte importante es interpretarlos con cierto sentido musical o saber situarlos en un contexto musical que relaciona los datos obtenidos, que no deja de ser números, con cualidades musicales.

[8] 12 o más intervalos iguales.

Los aspectos más cruciales para definir la música son el ritmo, la melodía y la armonía. El ritmo se suele relacionar a aspectos como el tempo, en pulsaciones por minuto, o el compás. La melodía y la armonía se asocian directamente a la frecuencia o visto desde un punto de vista musical, con la altura o tono. El tono se asocia a la altura por que muchas veces se habla de subir o bajar el tono y es una forma de relacionarlo a una idea pedagógica. La melodía describe la secuencia de tonos interpretada por un instrumento y la armonía estudia la progresión de diversas notas superpuestas o de acordes. Analiza los intervalos entre tonos o acordes que forman la estructura de una composición y los modos que intervienen. La única posibilidad de extraer las características que describan estos aspectos es mediante el análisis en frecuencia que permite asociar estos aspectos musicales un intervalo de tiempo y de esta forma definir la tonalidad o escala musical de los fragmentos analizados.

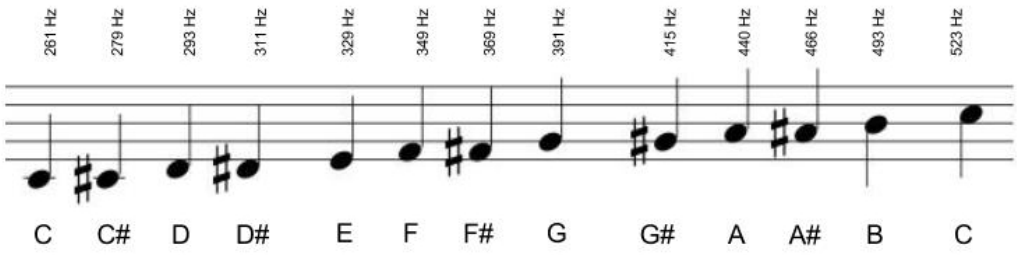


Figura 10. Relación frecuencia fundamental y tono musical de la octava del C_4 central.

Un sonido armónico presenta una tonalidad o tono musical, y es evidente, que existe una clara relación entre tono, frecuencia y nota musical. La escala temperada, referenciada en un LA_4 (440Hz), es la escala musical utilizada en la música moderna occidental. Cuando nos referimos a las notas, un intervalo se define como la distancia entre dos notas. La distancia más pequeña entre dos notas que podemos reproducir con un instrumento musical corresponde un semitono (se descompone en 100 céntimos). Por otro lado, un conjunto de doce notas forman una octava y el espectro está formado por 10 octavas. Por lo tanto, en la música occidental disponemos de 120 notas que justamente coincide con el numero de notas MIDI[9]. En realidad, hay muchas definiciones musicales que podemos relacionar con los descriptores de tonalidad como es el ritmo melódico, el análisis cualitativo de intervalos: segunda menor, cuarta justa, séptima menor, etc.

Podemos extraer características musicales si aplicamos estas conversiones o procesos que permiten relacionar la frecuencia con el tono musical. El ejemplo más representativo y en el que se centra este trabajo, es el cromagrama, que muestra relaciones entre la frecuencia y el tono musical. El cromagrama es un descriptor de tonalidad que responde a la mayor parte de las definiciones musicales relacionadas con el tono. Cuando hablamos de croma, en un contexto musical, estamos haciendo referencia a las notas musicales, y en un contexto más artístico hace referencia al color. Croma es una palabra que proviene del griego y que significa color[10].

[9] http://en.wikipedia.org/wiki/MIDI
[10] http://www.angelfire.com/de/nestsite/modbiogreek.html

La escala cromática o diatónica fue introducida por Pitágoras en la Antigua Grecia. Se fundamenta en la razón 3/2 y en el circulo de quinta justa. Según algunos estudios [11] muestran que esta escala se basa en la división del intervalo de una cuerda manteniendo la proporción aurea.

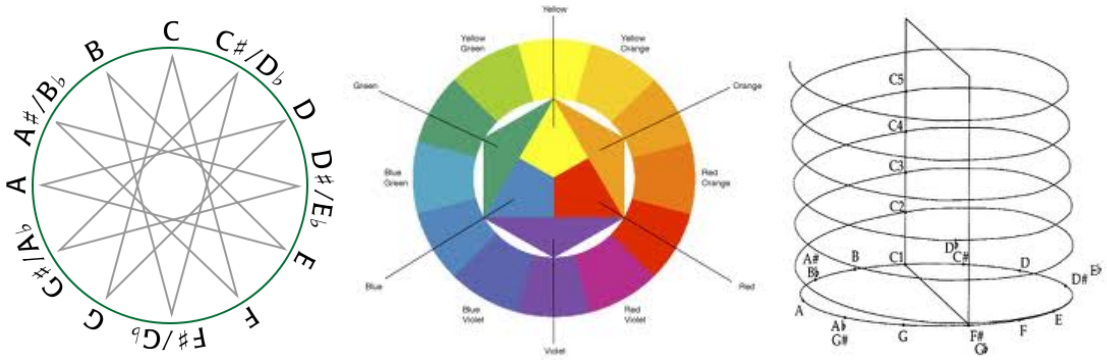


Figura 11. Relación entre el dodecaedro cromático, círculo de los 12 tonos del color[11] y una hélice que representa la repetición de las notas cromáticas.

Por esta razón, la escala cromática[12] se conoce como la que contiene todas las notas de la escala diatónica temperada. Muchas veces, se representa colocando las notas sobre un círculo o sobre una hélice, ya que las notas forman una secuencia que se va repitiendo cuando sube o baja la octava (figura 11). Es importante tener en cuenta que la relación entre las notas es lineal, sin embargo en frecuencia no es lineal. El intervalo de frecuencia entre dos octavas, se ajusta al doble en frecuencia, con lo que se relacionan con una potencia de dos.

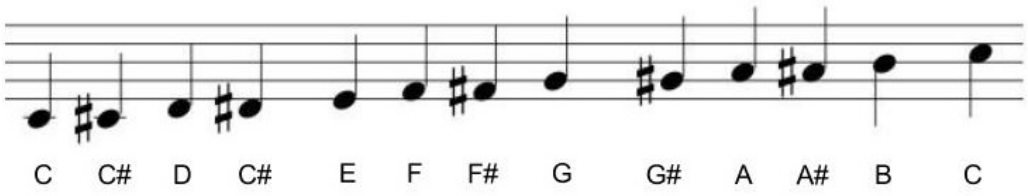


Figura 12. La escala cromática. Se compone de totas las notas de la escala temperada. C, D, E, F, G, A, B, C representan las notas musicales en el sistema internacional.

En música, se habla de cromatismos para referirnos al intervalo que hay entre dos notas adyacentes, es decir un semitono. Entre E a F hay un semitono, igual que de B a C, en el resto hay un tono de separación entre las notas sin alteración (#).

Al fin y al cabo, el cromagrama nos presenta el contenido musical referenciado en la escala diatónica para todas las notas que intervienen en la composición. En la figura 13 se muestra el cromagrama resultante de un archivo de audio que contiene diversas instancias del acorde de C mayor, con diferente instrumentación, que nos permitirá introducir el siguiente apartado.

[11] http://aphototeacher.com/2008/09/13/the-art-of-color/
[12] http://en.wikipedia.org/wiki/Chromatic_scale

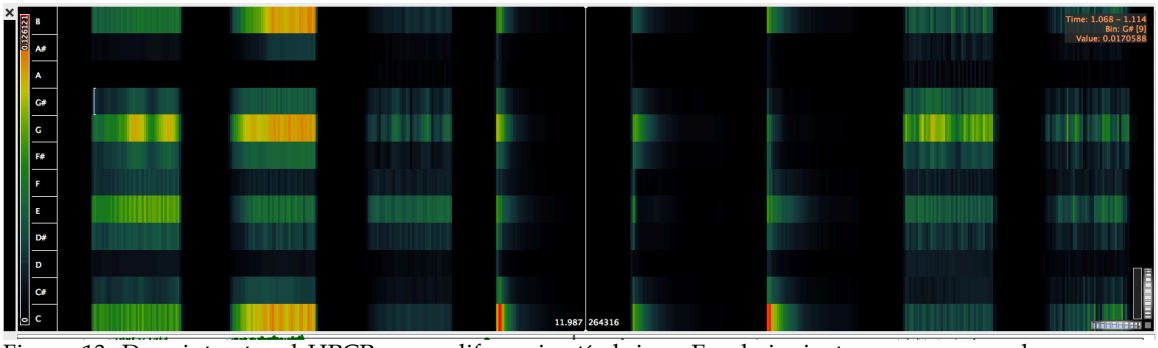


Figura 13. Descriptor tonal HPCP y sus diferencias tímbricas. En el siguiente cromagrama observamos que existe variación entre los resultados para tratarse del mismo acorde con diferentes instrumentos.

El cromagrama permite representar de forma compacta todo el contenido tonal sobre una única octava para definir con claridad el perfil de pitch o el acorde que contiene un fragmento. Cabe recordar que una octava musical se compone de 12 notas y que el espectro y las notas MIDI se componen de 10 octavas.

El cromagrama presenta el contenido tonal de una canción o fragmento de una señal de música polifónica de forma compacta, teniendo en cuenta todos los instrumentos que intervienen para informar en que escala musical está interpretada la pieza. Evidentemente, los descriptores tonales han de presentar cierta robusteza ante las distorsiones que puedan introducir, la presencia de ruido y el cambio de instrumentación o timbre [12]. Hoy por hoy, podemos encontrar diferentes implementaciones de extracción de cromagrama que se centran en el análisis del contenido tonal.

1.3.1 Descriptores tonales invariantes al timbre

Para conseguir su objetivo los descriptores tonales deben cumplir ciertos requisitos, como presentar resultados robustos a cambios de dinámica, ruido y cambios de instrumentación. En general, las diferentes versiones de cromagramas son invariantes al ruido y a los cambios de dinámica pero delante de cambios de instrumentación acostumbran a mostrar cierta deficiencia y sus resultados muestran como la intensidad relativa de los armónicos contenidos en cada nota musical influyen en el perfil de pitch estimado, como se ve en la figura 13.

El timbre es la cualidad del sonido que nos permite diferenciar un instrumento de otro. El timbre se define por la relación entre frecuencias (armónicos) que caracterizan el sonido (color) de un instrumento. El color es una manera más artística o musical de referirse al timbre.

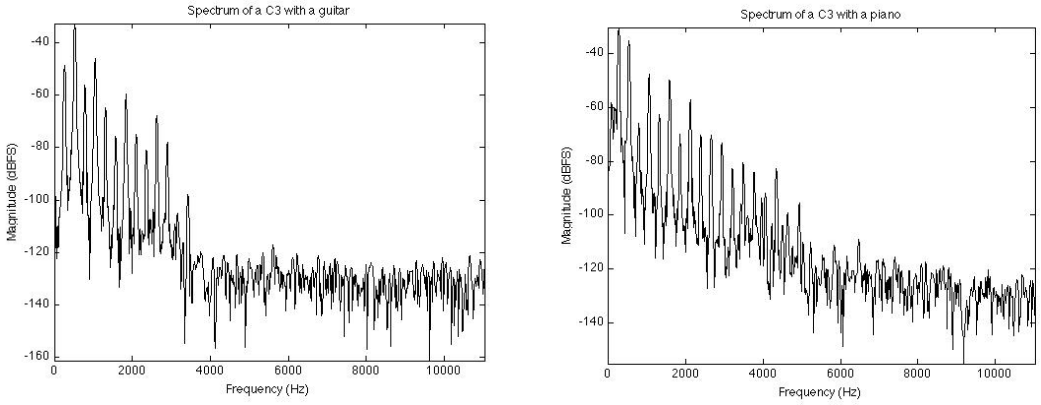


Figura 14. El espectro de la frecuencia de dos instrumentos de la nota C3. En en ambos casos, mantienen cierta relación entre armónicos que participan en la composición de cada uno de los sonidos. Sin embargo, presentan diferencias en la magnitud de cada uno de los armónicos. Observamos que la serie armónica que contiene el espectro de un sonido de piano contienen más armónicos superiores que el sonido de una guitarra clásica. El timbre es la cualidad perceptual del sonido que diferencia un instrumento de otro.

Hoy en día es muy frecuente encontrar versiones o covers de las composiciones musicales más representativas de cada estilo musical, donde en muchos de los casos la instrumentación utilizada cambia. Por ejemplo, en la música clásica acostumbra a interpretarse una composición con diferente orquestación, en el rock, se versiona mucho '*Smoke on the Water*' de Deep Purple, bandas de funk versionan '*Sex Machine*' de James Brown, bandas de reggae versionan '*Jammin'* de Bob Marley. Aunque el contenido melódico y armónico sigue siendo el mismo (mismas notas y acordes), el cambio en la instrumentación afecta al timbre de la composición original y lo altera. Además, entre todo el material sonoro encontramos diferentes codificaciones y grabaciones que según su calidad modifican el contenido en frecuencia. Esto muchas veces afecta al contenido alterando el timbre, lo que provoca errores en la detección y identificación de grabaciones de música mediante el cromagrama. Además, la mayor parte del contenido musical durante el proceso de grabación, mezcla y masterización , se ve sometido a procesos como la ecualización, compresión, efectos de reverberación y retardos que varían la respuesta en frecuencia de los instrumentos y por lo tanto su timbre original. Asimismo, en los grandes estudios y la mayoría de los discos de cierto cache, realizan estos procesos con dispositivos analógicos que no siempre mantienen una respuesta lineal en todo el espectro de frecuencias. De esta forma, introducen modificaciones en la frecuencia y distorsión armónica en la señal musical. Según qué marca y modelo de dispositivos se utilice en cada uno de los procesos el timbre se verá alterado en cierto modo.

Teniendo en cuenta todas estas consideraciones, es evidente que precisamos de un enfoque de extracción de cromagrama inalterable ante los procesos y situaciones que acabamos de comentar. Si el resultado de un cromagrama muestra diferencias para la misma nota con un piano y con una guitarra clásica, sus resultados para una grabación de música, como huella sonora, de los procesos aplicados en el proceso de producción, también dependerá del formato del archivo y sobre todo de la instrumentación. En este contexto, conseguir un grado alto de invarianza al timbre en una extracción de cromagrama se considera todo un éxito para la industria, puesto que mejoraría la identificación de covers o la descripción del perfil de pitch.

En resumen, el cromagrama puede ser una herramienta muy potente para la similitud de música Occidental armónica, si mejoramos su robusteza a los cambios de timbre. Por esta razón, se investigan nuevos procedimientos que enfaticen esta capacidad y que consigan extraer un identificador único, robusto a los cambios de instrumentación y los demás requisitos que se le exigen a un descriptor tonal [12]. Encontrar una solución a esta cuestión puede resolver gran parte del paradigma de la clasificación y la identificación automática de grabaciones de música, puesto que la mayor parte de los descriptores tonales muestran ciertas deficiencias en cuanto a la independencia al timbre y el instrumento.

1.4 Objetivos: evaluar y mejorar los descriptores tonales

El objetivo principal de este trabajo es cuantificar el grado de invarianza al timbre y el grado de eficiencia de la estimación de acordes de un descriptor tonal. Con el grado de eficiencia nos referimos a el potencial que presentan al extraer el contenido tonal. Para todo ello, creamos un método que mediante la comparación de los cromagramas con la detección ideal o perfil teórico que nos permite comparar los resultados entre las diferentes versiones de cromagrama.

Para la evaluación utilizaremos los cromagramas más representativos y relevantes, Harmonic Pitch Class Profile HPCP [13] y Chroma DCT Reduced Log Pitch CRP [14]. Además, analizamos diferentes técnicas como: la estimación del tono mediante la frecuencia instantánea y el filtrado de coeficientes cepstrum para corregir la envolvente espectral, que se pueden adaptar a la implementación en C++ del HPCP[13]. El objetivo final es implementar estos métodos en un Vamp plugin, que es una pequeña aplicación externa que podemos desarrollamos en C++ y que nos permite visualizar los resultados con Sonic Visualizer cargándolo como una librería dinámica. La detección de picos en frecuencia mediante la frecuencia instantánea lo implementamos como plugin independiente que puede utilizarse para el análisis en frecuencia. En cambio el filtrado de coeficientes cepstrum lo introducimos en la cadena de procesado del HPCP como función de blanqueado espectral y creamos una nueva versión del HPCP.

Para conseguirlo, hace falta desarrollar una herramienta en MATLAB capaz de evaluar la eficiencia y la varianza al timbre de los descriptores tonales. Una vez hecho esto, extraemos conclusiones de los resultados de la evaluación de descriptores. El análisis de las nuevas técnicas requiere su implementación en MATLAB, para poder analizarlas con detalle y rapidez. Finalmente, añadimos estas técnicas al HPCP o desarrollarlas en C++ como una librería dinámica, lo que requiere introducirnos en un entorno de desarrollo y familiarizarnos con la Vamp Plugin SDK[14]. Esta parte de desarrollo une diferentes disciplinas, puesto que hay que trabajar con comandos en la consola, el procesado de señal, la programación en C++ y la evaluación del resultado final.

[13] http://mtg.upf.edu/technologies/hpcp
[14] http://vamp-plugins.org/code-doc/main.html

1.5 Estructura del trabajo

El trabajo se compone de 6 capítulos, en el primero de ellos, como habéis podido comprobar, se centra en introducir de la mejor forma posible el problema que estudiamos, está dividido en cuatro partes que, con una pequeña introducción que resume la idea general de lo que se quiere hacer o de lo que se espera de este trabajo. En el siguiente apartado, se trata de situar la problemática de la clasificación de información dentro de un contexto musical y se contrasta con referencias actuales para poder atraer la atención del lector. En el siguiente capitulo introducimos conceptos más técnicos pero indispensables para llegar a una entera comprensión. Explicaremos conceptos como el tono musical, la frecuencia de referencia, la melodía, la armonía, el timbre, la serie armónica, función ventana, STFT, etc.

En el tercer capitulo se explica con detalle la evaluación de los descriptores tonales (HPCP y CRP) además de mostrar y comentar los resultados. También se presentan los métodos de mejora que planteamos para el HPCP y algunos de los experimentos realizados. En el cuarto capitulo mostramos los resultados de la implementación de las mejoras propuestas. Además se muestran los resultados de la evaluación del grado de eficiencia y el grado de invarianza al timbre de la nueva versión del HPCP. Finalmente en el quinto capitulo presentamos la conclusión final de este trabajo.

2. Definiciones, técnicas y aplicaciones

2.1 Introducción

En este capitulo presentamos el contexto en el que se encuentra este trabajo y revisamos la literatura relacionada. También, explicamos los conceptos básicos y relevantes que ayudan a comprender el propósito de este trabajo como: es estudiar la invarianza al timbre en el análisis de la tonalidad de la literatura consultada o como cuáles son los descriptores tonales más sobresalientes en este sentido y que se emplean para la descripción y identificación de grabaciones de música.

2.2 Definiciones

2.2.1 Tono

El tono de un sonido es la propiedad de percepción que nuestro sistema auditivo asocia a la frecuencia. El sonido más simple que podemos generar es una señal sinusoide, que se modela con la siguiente ecuación:

$$x(t) = A \cdot sin(2 \cdot \pi \cdot f \cdot t + \phi) \tag{2.1}$$

donde A corresponde a la amplitud, f a la frecuencia, t al instante de tiempo en segundos y ϕ a la fase instantánea definida entre 0 y 2π.

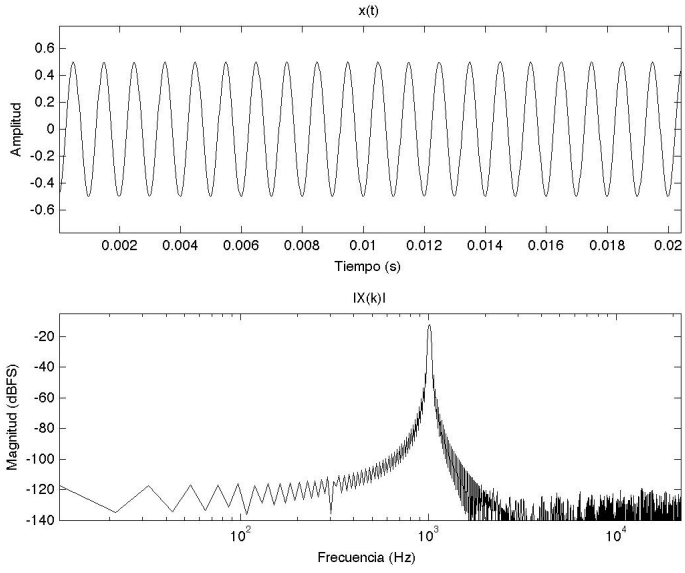


Figura 15. En la parte superior mostramos una señal sinusoidal de 1 KHz en el dominio temporal. En la parte inferior, la magnitud del espectro en el domino frecuencial.

Normalmente, en una situación real, el sonido acostumbra a ser complejo y podemos descomponerlo en N señales sinusoidales:

$$s(t) = \sum_{n=1}^{N} A(n) \cdot sin(2 \cdot \pi \cdot f(n) \cdot t + \phi(n)) \qquad (2.2)$$

donde A corresponde a la amplitud, t el instante de tiempo en segundos y ø a la fase instantánea de cada una de las sinusoides que lo componen.

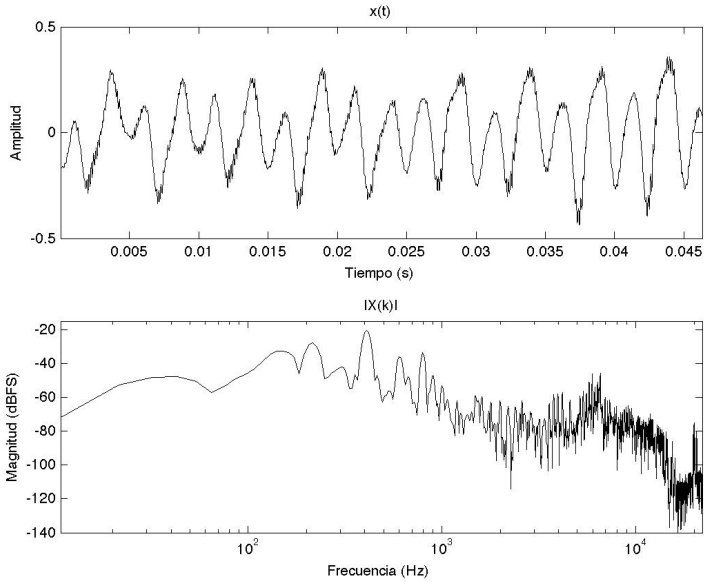


Figura 16. En la parte superior, se muestra un pequeño fragmento de una señal de audio compleja. Se distingue el periodo de la señal y su riqueza armónica. En la parte inferior, vemos que su contenido en frecuencia se extiende en todo el espectro.

2.2.2 Sonido armónico y la frecuencia fundamental

Cuando se trata de un sonido simple, como una señal sinusoidal, asociamos el tono directamente a su frecuencia, en el caso de sonidos complejos, que presentan cierta tonalidad, normalmente solemos asociar el tono a su frecuencia fundamental (f_o). Un sonido complejo con tonalidad, se conoce como un sonido armónico cuando presenta cierta relación entre sus frecuencias y todos ellos acostumbran a ser múltiplos de la frecuencia fundamental. Las frecuencias que no son múltiplos enteros de f_o se les denomina parciales o inarmónicos. El sonido de la nota de un instrumento se compone de su f_o y una serie armónica que le dan al instrumento un timbre en particular, definido por la amplitud de cada uno de los armónicos. Esto no quiere decir que todos los instrumentos no contengan parciales (inarmonía y disonancia).

$$s(t) = \sum_{n=0}^{N} A(n) \cdot \sin(2 \cdot \pi \cdot n \cdot f_o \cdot t + \phi(n)) \qquad (2.3)$$

A corresponde a la amplitud, f a la frecuencia, t al instante de tiempo en segundo y ø a la fase instantánea. En un sonido armónico, podemos definir la f_o como el mínimo común divisor del conjunto de frecuencias que corresponden a los picos espectrales.

$$f_o = mcm(f(n)) \tag{2.4}$$

Aplicando la detección de picos sobre el espectro en frecuencia podemos calcular el mínimo común divisor de las frecuencias relacionadas a los picos. Podemos encontramos algoritmos dedicados a esta tarea en el dominio temporal (auto-correlación o YIN) o en el dominio frecuencial (detección parabólica o las frecuencias instantáneas). En la naturaleza encontramos sonidos inarmónicos donde la relación entre el contenido en frecuencia no responde a una serie de frecuencias múltiples y que no se puede modelar con la ecuación anterior.

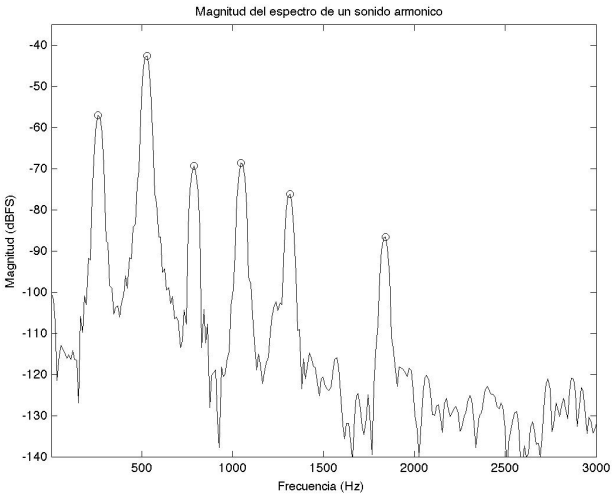


Figura 17. Magnitud del espectro de un sonido armónico. Concretamente de la nota C3 con una guitarra acústica.

En la figura se observa el espectro de la nota C3 de una guitarra. Vemos que cada uno de los picos espectrales son múltiplos de la frecuencia 265Hz, es decir múltiplos de la frecuencia fundamental.

2.2.3 Timbre

El timbre es la cualidad del sonido que nos permite diferenciar entre diferentes instrumentos. El timbre de un instrumento depende de su contenido espectral, concretamente del numero de armónicos que lo componen y de su magnitud. Podemos asociar una serie armónica al timbre de un instrumento.

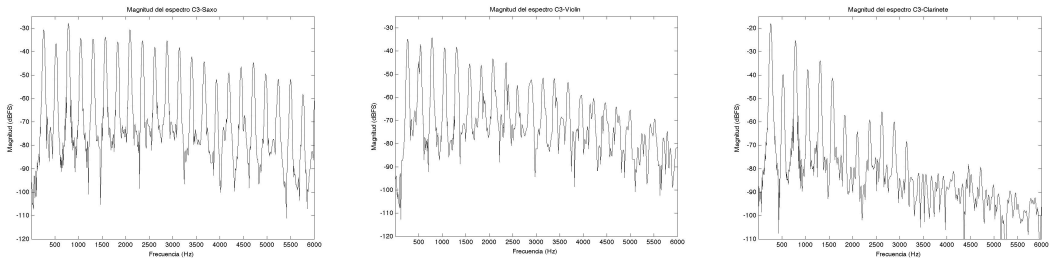


Figura 18. Muestra la magnitud del espectro de la nota C3 con diferentes instrumentos: saxofón, violín y clarinete. Se observa que cada instrumento tiene diferente timbre y se asocia diferentes series harmónicas.

El sonido de saxofón presenta una mayor contribución de los armónicos superiores que se puede asociar a su estridencia (mas de 22 armónicos). Mientras que un sonido más meloso, como el de un clarinete, conserva la mayor parte de su energía en los armónicos inferiores (11 armónicos). El caso del violín vemos que las amplitudes de la serie armónica asociada a su timbre es muy distinta.

2.2.4 Sonido polifónico

Un sonido polifónico se define como aquel que contiene más de una nota o más de un tono, sean fruto de uno o varios instrumentos. Podemos hablar de instrumentos polifónicos, que son aquellos que permiten reproducir varias notas en el mismo instante (guitarra, piano, arpa, acordeón, etc.). En el caso de los instrumentos musicales electrónicos esta propiedad se define como voces de polifonía. Durante este trabajo, por lo general, cuando nos referimos a un sonido polifónico haremos referencia a grabaciones musicales.

2.2.5 Frecuencia de referencia

En lenguaje musical cuando nos referimos al tono de un instrumento monofónico estamos haciendo referencia a la nota. El estándar de referencia para afinar un instrumento en la escala musical temperada es el LA_4 (440Hz). Sirve de referencia para la afinación de instrumentos musicales. En 1936, se estableció como estándar internacional, anteriormente se utilizaban otras frecuencias como 432Hz[15]. Aun así, según el diseño y el material empleado en la construcción de los instrumentos musicales, podemos encontrar casos que presentan desviaciones de la frecuencia de referencia en algunos tonos, como ocurre en el piano.

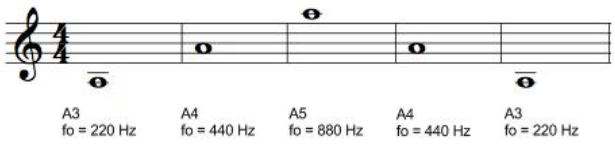


Figura 19. Relación entre tono musical y frecuencia fundamental

[15] Ciertos sectores defienden el 432Hz como estándar musical internacional por su relación con la resonancia de la Tierra y las ondas alfa (8Hz). http://www.schillerinstitute.org/music/rev_verdituning.html

En un contexto musical el tono de un sonido se asocia a su nota. El intervalo que esta compuesto por 12 notas se define como una octava. Subir una octava corresponde a doblar la frecuencia y bajar una octava corresponde a dividir la entre dos. De esta forma, deducimos que el tono musical y la frecuencia se relacionan con una potencia de dos. Un semitono es igual a la doceava parte de una octava y su proporción geométrica corresponde a la siguiente expresión:

$$semitono = \sqrt[12]{2} = 1.05946 \tag{2.5}$$

Como se puede observar la división de una octava en 12 notas no acaba de ser perfecta, puesto que el intervalo entre semitonos no es entero. Pero es la única manera de asegura que todas las notas tengan la misma desafinación respecto a la frecuencia de referencia.

2.2.6 Escala Mel y escala musical

El oído humano percibe el sonido de forma no lineal, tanto amplitud como tono. En cuanto al tono se refiere, 5 de las 10 octavas posibles, se concentran entre 30 y 1000Hz. Las 5 octavas restantes se extienden en un rango de 1000 a 20000 Hz. Por ese motivo, en algunas aplicaciones se pasa la frecuencia a escala Mel o a la escala temperada.

En los primeros 1000Hz, se concentran cinco octavas, donde la correspondencias entre frecuencias y la escala Mel responde a una función prácticamente lineal. En las octavas superiores están mucho mas separadas y la correspondencia entre frecuencias y escala Mel es logarítmica.

Figura 20. Relación entre frecuencias y escala Mel[16]

$$m = 2595 \cdot \log_{10}\left(1+\frac{f}{700}\right) \tag{2.6}$$

La parte del contenido espectral que define el timbre se concentra en las primeras octavas, donde la relación entre armónicos y sus contribuciones es más notable. Normalmente los armónicos superiores al octavo armónico apenas tienen energía y no contribuyen de la misma forma en las octavas superiores.

Para mapear de frecuencia a tono musical, nos encontramos con una relación en base 2, dado que la relación entre octavas es del doble. El intervalo entre un pico espectral (frecuencia) y la frecuencia de referencia, medida en semitonos β, se define de la siguiente manera:

[16] http://en.wikipedia.org/wiki/File:Mel-Hz_plot.svg

$$\beta = 12 \cdot log_2(\frac{f}{440}) \quad (2.7)$$

β es negativo para frecuencias inferiores a la frecuencia de referencia y viceversa. Otra relación importante a tener en cuenta y que permite mapear de frecuencia a nota MIDI, es la que mostramos a continuación:

$$p = 69 + 12 \cdot log_2(\frac{f_i}{440}) \quad (2.8)$$

donde p es el número de nota MIDI en un rango de 0 a 127.

2.2.7 Escala temperada y escala cromática

La escala temperada es la escala musical utilizada en la actualidad. Esta escala musical divide una octava en 12 semitonos iguales. Como hemos mencionado antes, la división de una octava en 12 notas no coincide con un semitono exacto, en realidad es un poco más de un semitono (1.059). De esta forma, aseguramos que todas las notas tengan la misma desafinación o temperamento. Una propiedad importante de escala es que permite transportar las composiciones de tono sin cambiar los intervalos musicales.

La escala cromática es la escala musical que contiene las 12 notas de la escala temperada. De la misma forma que la escala temperada, es la escala musical que mantiene sus notas equidistantes.

2.2.8 Análisis Espectral

En el dominio temporal se representa el sonido en dos dimensiones mediante lo que se denomina forma de onda. El eje de abscisas se asocia al fragmento de tiempo y el eje de coordenadas a la energía en cada instante *x(n)*. Una de las técnicas que nos permite analizar el contenido musical de una grabación es el análisis espectral. La transformada de Fourier permite transmitir muestras del dominio temporal al de la frecuencia mediante un cambio de base de *N* dimensiones, donde *N* es el numero de frecuencias.

Antes de aplicar la FFT sobre el fragmento de audio se acostumbran a realizar ciertos procesos. Para realizar el análisis espectral de una señal de audio las etapas principales en el procesado de la señal son: enventanado, relleno de ceros y el cálculo de la transformada discreta de Fourier.

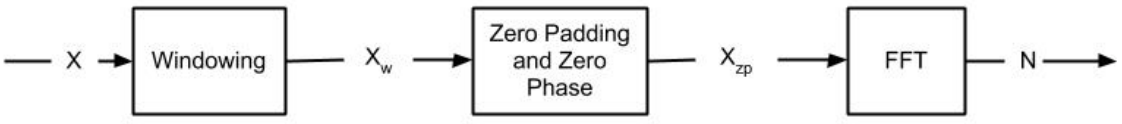


Figura 21. Esquema del análisis espectral

2.2.8.1 Enventanado

Es el primer paso a realizar en un análisis espectral. Se centra en multiplicar cada uno de los fragmentos de audio por una función ventana *w(n)*. Cada función de ventana muestra diferente comportamiento en el dominio de la frecuencia: mayor contribución de los lóbulos secundarios según el ancho de banda del lóbulo principal. Las funciones más utilizadas en el análisis espectral son Hann, Hamming y Blackman. El proceso es equivalente a la siguiente ecuación.

$$x_w(n) = x(n + l \cdot N_{hop}) \cdot w(n) \tag{2.9}$$

donde l corresponde al número de fragmento, N_{hop} es el tamaño de la transformada de Fourier y $x(n + l \cdot N_{hop})$ es un fragmento de audio indeterminado.

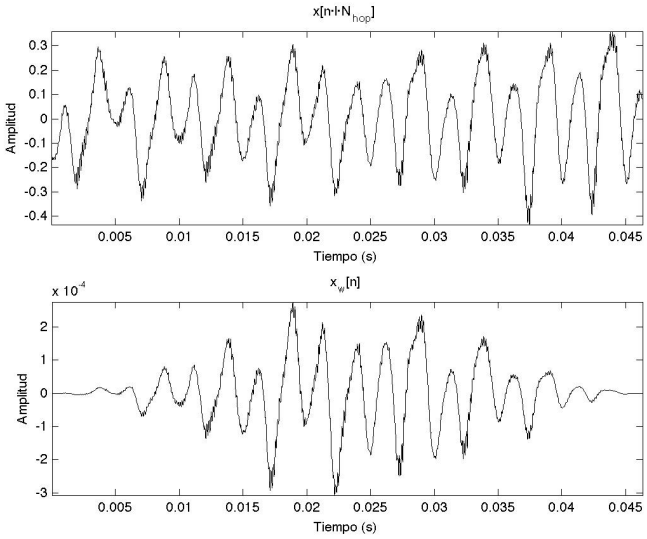


Figura 22. Un fragmento de una señal de audio en el cálculo de la STFT.

La resolución temporal del proceso es proporcional al tamaño de la ventana.

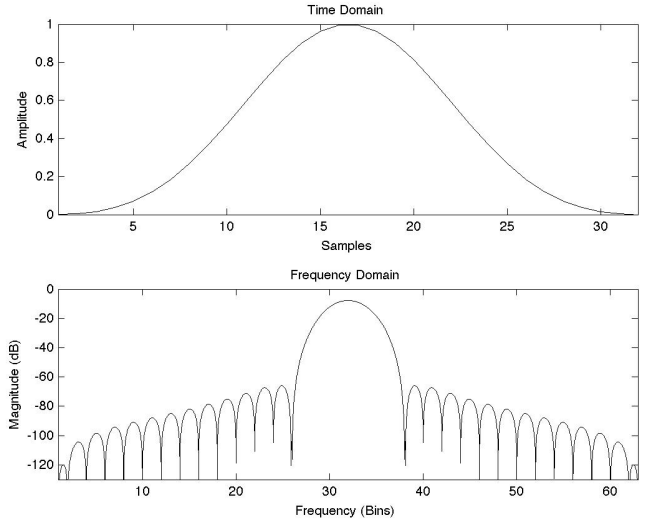


Figura 23. Función ventana Blackman. Arriba dominio temporal, abajo frecuencial.

2.2.8.2 Relleno de ceros

Este proceso se centra en rellenar el fragmento de audio con ceros hasta alcanzar cierto tamaño. Este proceso mejora el análisis espectral, ofreciendo un espectro de con mayor resolución en frecuencia, puesto que aumentar el tamaño del fragmento con ceros a una interpolación parabólica en el espectro. Seguidamente se aplica la fase cero.

$$x_{zp}(n) = \begin{cases} 0 & si \quad n = -\frac{N_{FFT}}{2} \ldots -\frac{N_{frame}}{2} - 1 \\ x_{wc}(n) & si \quad n = -\frac{N_{FFT}}{2} \ldots \frac{N_{frame}}{2} - 1 \\ 0 & si \quad n = \frac{N_{FFT}}{2} \ldots \frac{N_{frame}}{2} - 1 \end{cases} \tag{2.10}$$

2.2.8.3 DFT

Seguidamente, se calcula la transformada de Fourier donde obtenemos el espectro complejo *X(k)* de un fragmento de audio como se explica en el trabajo de McCleallan, Schaffer y Yoder [30].

$$X(k) = \sum_{n=-\frac{N_{FFT}}{2}}^{\frac{N_{FFT}}{2}-1} x_{zp}(n) \cdot e^{-j2\pi nk/N_{FFT}} \tag{2.11}$$

donde $k = 0, 1 \ldots N_{frame-1}$. El espectro contiene una parte real y otra imaginaria. Calculamos la magnitud y la fase del espectro de la siguiente manera:

$$|X(k)| = \sqrt{X_r(k)^2 + X_i(k)^2} \tag{2.12}$$

$$\phi(k) = arctan \frac{X_i(k)}{X_r(k)} \tag{2.13}$$

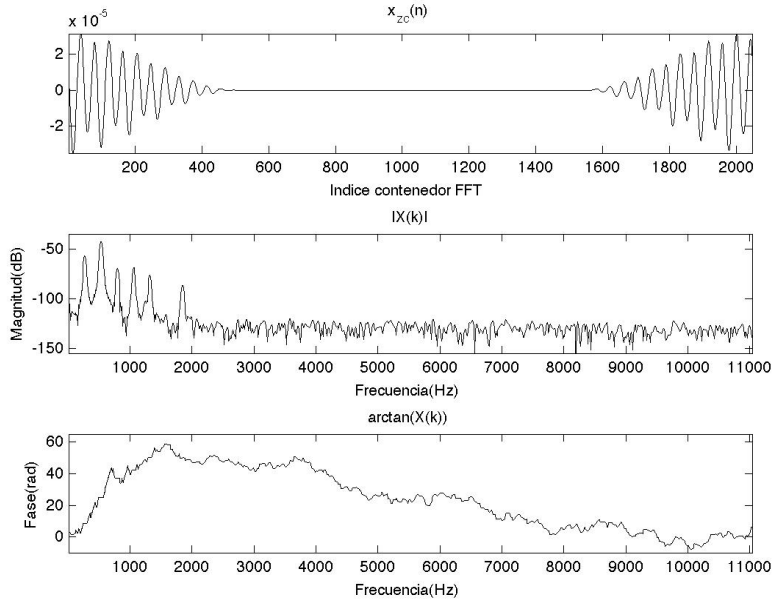


Figura 24. Calculo de la FFT. Arriba, el fragmento enventanado después del aplicar fase cero y relleno de ceros. En el centro, la magnitud del espectro y abajo la fase.

2.2.9 Cepstrum

Esta técnica permite analizar y procesar el contenido espectral. Se calcula mediante la FFT del logaritmo de la magnitud del espectro. Este proceso se hereda del reconocimiento de habla [15].

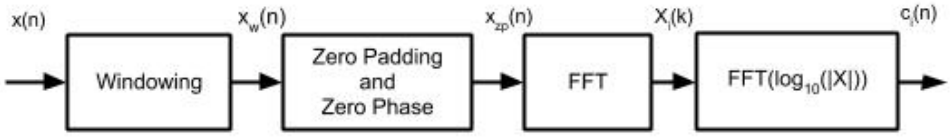


Figura 25. Diagrama del cálculo de los coeficientes cepstrum

Cada coeficiente cepstrum se asocia a la frecuencia mediante f_s/N, donde f_s es la frecuencia de muestro y N es el número de coeficiente. Normalmente el pico mas destacado que aparece en los coeficientes cepstrum corresponde a la frecuencia fundamental (figura 26). Se asocia el coeficiente cepstrum a la frecuencia con la siguiente expresión:

$$f = \frac{f_s}{N} \tag{2.14}$$

donde N, es el índice del coeficiente cepstrum y f_s, la frecuencia de muestreo. En la figura 21, se muestran los coeficientes cepstrum de un sonido de clarinete de la nota C_4 (261.62Hz) con frecuencia de muestreo 20050Hz. Podemos apreciar que el primer pico corresponde al bin 85, lo que se equivale a la frecuencia 20050/85 = 259.4118Hz.

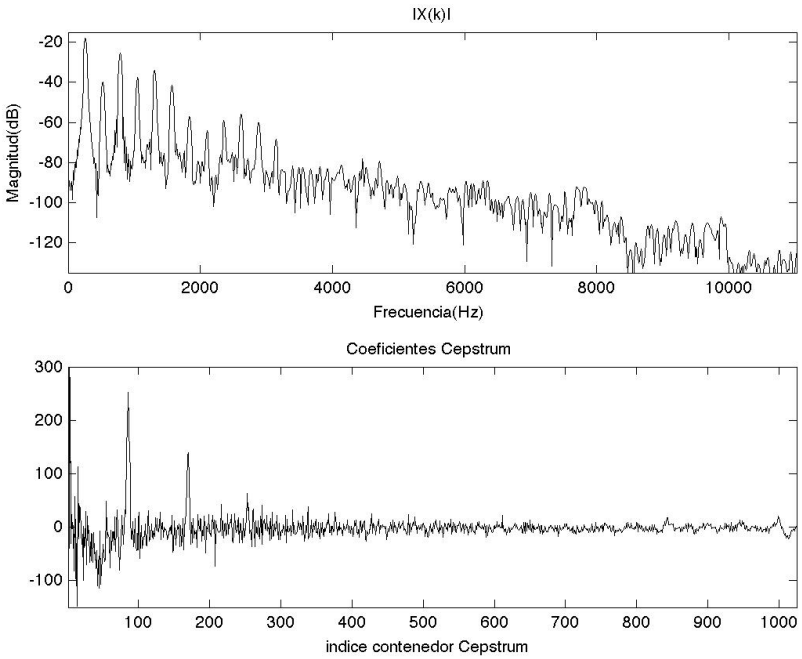


Figura 26. Coeficientes cepstrum vs. magnitud del espectro. Arriba, la magnitud del espectro y abajo, sus coeficientes cepstrum.

2.3 Aplicaciones del cromagrama

Como ya hemos mencionado en el capitulo anterior el cromagrama se utiliza en diversas aplicaciones musicales. El principal objetivo del análisis musical con cromagrama es mostrar relaciones significativas entre diferentes extractos musicales de un conjunto de datos que acaben representando piezas musicales por si mismas. Estos descriptores se utilizan en tareas de análisis musical, emparejamiento y similitud de música con un perfil similar y identificación de versiones o *covers*, clasificación por genero o recomendación.

2.3.1 Análisis musical: estimación de acordes, tonalidad, etc.

Mediante el cromagrama y mecanismos inteligentes se consigue estimar los acordes, la tonalidad de la pieza o su estructura musical. Basándose en métodos de clasificación y reconocimiento de datos como el modelo oculto de Markov (HMM) o el modelado de la mezcla de gaussianas (GMM) conseguimos aproximar la tonalidad o la estructura estimada a los valores teóricos o reales[16] .

La estimación de acordes se puede conseguir de forma muy sencilla. Haciendo uso de una base de datos, que contenga el perfil de cada uno de los acordes, y de medidas como la distancia euclideana o del coseno, se puede encontrar el perfil acorde de mayor similitud, que corresponde al perfil que presente la mínima distancia respecto al cromagrama. Para encontrar la tónica se hace un desplazamiento circular en todas las posibles posiciones utilizando medidas como la distancia o métodos mas sofisticados (HMM o GMM).

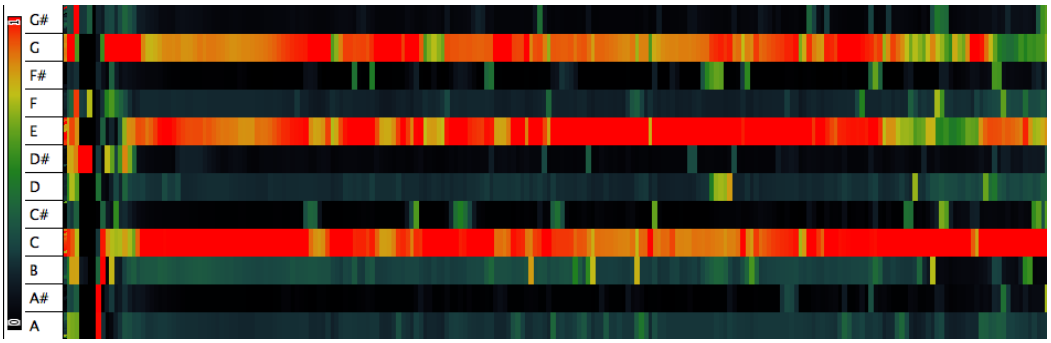


Figura 27. Cromagrama de un acorde de C mayor.

2.3.2 Similitud: detección de versiones o covers

Los sistemas de similitud se basan en el análisis y la comparación de diferentes grabaciones para encontrar relaciones que puedan asociarlas [17]. El cromagrama permite acceder al contenido tonal de cada fragmento de tiempo y esto lo convierte en una potente herramienta de detección de versiones, puesto que permite comparar su armonía y estructura.

2.3.3 Clasificación por género, por compositor, por estado de ánimo o recomendación musical.

La clasificación de grabaciones musicales según distintos aspectos como el genero, compositor o estado de animo precisan de métodos de clasificación y reconocimiento como vectores de soporte (SVM), HMM o GMM. Son aplicaciones que necesitan de previo entrenamiento con parte de los datos antes de poder clasificar o recomendar [18]. El croma se puede utilizar para establecer relaciones entre estilos o compositor. Si no también se pueden utilizar descriptores de nivel medio que estimen la armonía o tonalidad mediante el cromagrama.

2.4 Algoritmos de extracción de cromagrama: HPCP y CRP

Los diferentes enfoques para la extracción del cromagrama aplican distintas técnicas para conseguir aproximar el contenido tonal. Este trabajo se centra en dos implementaciones: *'Harmonic Pitch Class Profile'*, HPCP[17] [12] y *'Chorma DCT Reduced Log Pitch'*, CRP[18] [14]. Ambos muestran grandes diferencias en sus planteamientos y por lo tanto presentan diferentes capacidades ante diferentes situaciones en la extracción de las características.

2.4.1 Harmonic Pitch Class Profile, HPCP

Estas características representan la distribución tonal por clases de una pieza musical y están relacionadas con el contenido en frecuencia. El HPCP fue propuesto en [13] con el siguiente diagrama:

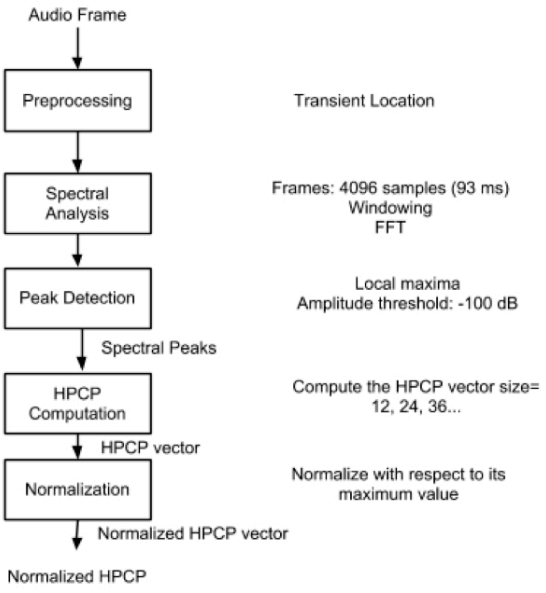


Figura 28. Esquema del proceso de extracción del HPCP. (Gómez, 2006)

[17] http://mtg.upf.edu/technologies/hpcp
[18] http://www.mpi-inf.mpg.de/resources/MIR/chromatoolbox/

2.4.1.1 Pre-procesado

En esta etapa se prepara la señal de audio para el calculo del vector de la distribución de tono. Incluye un proceso para la detección de transitorios (Bonada, J. 2010), con la finalidad de eliminar fragmentos que contienen transitorios. Este algoritmo elimina los fragmentos de audio que presentan una estructura armónica muy ruidosa. Como son fragmentos muy pequeños (100 milisegundos), no afecta al resultado del descriptor tonal.

2.4.1.2 Análisis espectral

En esta etapa se aplica el análisis espectral mediante FFT, como se muestra en la figura 21. Se compone de varias etapas: enventanado, fase cero, relleno de ceros y STFT.

En el análisis tonal que propone el HPCP es necesaria una buena resolución frecuencial para poder definir sobretodo el rango de bajas y medias frecuencias. Por esta razón se aplica un relleno de ceros que proporciona mayor resolución y se propone una tamaño de la FFT de 4096 muestras. Además se recomienda una frecuencia de muestreo de 44.1KHz. Los fragmentos consecutivos se separan por 512 muestras, lo que equivale a un $N_{FFT}/8$. Además propone utilizar una función ventana 92dB-Blackman Harris de $N_{FFT}/2$ muestras. Posteriormente al cálculo de la FFT, se aplica la detección parabólica de picos espectrales sugerida por Serra, X. en el modelo sinusoidal.

2.4.1.3 Detección de pico espectral: la detección parabólica

Este algoritmo se desarrollo en el marco '*Sinusoidal Modelling Synthesis*' (SMS) y fue propuesto por Xavier Serra, director del MTG. Principalmente se centra en la suposición de que un fragmento de audio puede representarse con N componentes en frecuencia, de esta forma el modelo asume que el espectro $X(k)$ se puede representar por un numero de componentes mucho mas pequeño. Un pico espectral se define como un máximo local del la magnitud del espectro, la única restricción que es presenta es que la magnitud del pico sea mayor que cierto umbral determinado[19]. Cada pico tiene una precisión de la mitad de un contenedor del espectro, que representa un intervalo en frecuencia de f_s/N_{FFT} *(Hz)*.

Para mejorar la precisión de la detección, se aplica una interpolación cuadrática entre las muestra mas próximas al pico que definen la nueva posición del pico. Se utilizan las tres muestras mas próximas al máximo detectado, considerando que forman una parábola.

$$y(x) = a(x-p)^2 + b \tag{2.15}$$

donde p es el centro de la parábola, a define su concavidad y b es el factor que compensa el desplazamiento *offset*.

[19] Por defecto -100 dBFS.

De esta forma se obtiene el valor del centro de la parábola p:

$$p = \frac{1}{2} \cdot \alpha - \frac{\gamma}{\alpha} - 2 \cdot \beta + \gamma \tag{2.16}$$

donde α corresponde a la magnitud en decibelios de *y(-1)*, suponiendo que *y(0)* corresponde a un pico local, γ a la magnitud de *y(1)* y β a la magnitud del pico local. El contenedor que corresponde a cada pico espectral se define mediante:

$$k^* = k_\beta + p \tag{2.17}$$

La magnitud del pico se determina de la siguiente forma:

$$y(p) = 20 \cdot log10(|X(k^*)|) = \beta - (\alpha - \gamma) \cdot p / 4 \tag{2.18}$$

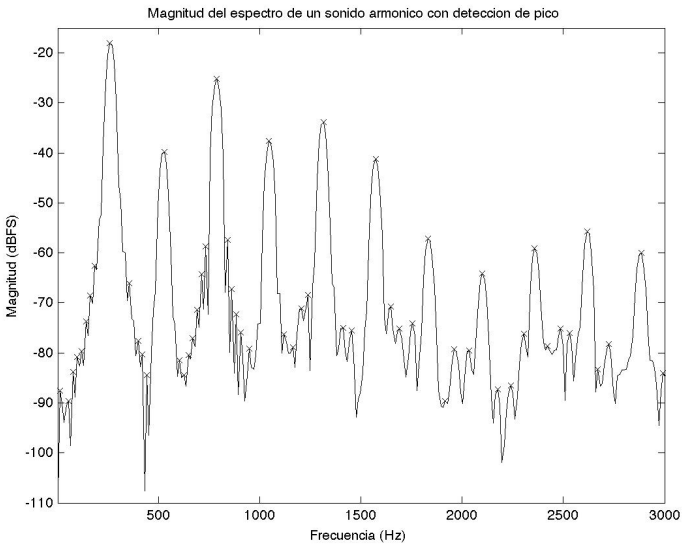


Figura 29. Detección de picos espectrales sobre la magnitud del espectro.

A continuación, el siguiente paso es realizar una selección entre los picos detectados según su frecuencia. Únicamente se consideran los picos contenidos en el intervalo de 100 a 5000Hz, donde se concentra la información tonal. Como resultado de este proceso se obtiene un conjunto de pico $\{a_i, f_i\}$, $i = 1,\ldots,n_{peaks}$ donde a_i indica la amplitud de cada pico y f_i su frecuencia.

2.4.1.4 Cálculo del HPCP

Los perfiles de tono según su clase (PCP) consisten en relacionar la frecuencia y con cada clase de tono según la escala temperada y haciendo uso de una frecuencia de referencia determinada. El problema es que las piezas musicales no siempre están perfectamente afinadas o utilizan frecuencias de referencia diferentes al estándar, 440Hz. Para conseguir un descriptor tonal independiente a la frecuencia de afinación es necesario estimar esta frecuencia para el análisis tonal.

Con las ultimas modificaciones en la implementación del HPCP en formato Vamp plugin, realizadas por Jordi Bonada, este parámetro lo define el usuario pero en la versión original se estima analizando la desviación de los picos espectrales respecto a la frecuencia estándar.

Como se dispone de una frecuencia de referencia global se procede al cálculo del HPCP . Este descriptor tonal mide la intensidad en cada uno de los 12 semitonos de la escala temperada y se obtiene mapeando cada frecuencia a su tono correspondiente. El HPCP introduce algunas modificaciones respecto al calculo del PCP propuesto por Fujishima, 1999 [15]. El vector HPCP se define con la siguiente expresión:

$$HPCP(n) = \sum_{i=1}^{nPeaks} w(n, f_i) \cdot a_i^2 \qquad n=1,...,size(HPCP); \tag{2.19}$$

a_i y f_i son la magnitud lineal y la frecuencia del pico i, n es el numero de índice del HPCP, *size(HPCP)* es el tamaño del vector HPCP (de 12 a 120) y $w(n,f_i)$ corresponde al peso de la frecuencia f_i cuando se considera el índice del HPCP numero n. La función de pesos introducida considera la presencia de armónicos y su contribución a diferentes tonos. Se utiliza para lograr una mayor resolución en el cálculo del HPCP. El peso de cada pico depende de su distancia con la frecuencia central del contenedor de tono. La frecuencia central se define de la siguiente forma:

$$f_n = f_{ref} \cdot 2^{\frac{n}{size}} \tag{2.20}$$

La distancia en semitonos entre el pico f_i y la frecuencia central del contenedor n, f_n:

$$d = 12 \cdot log_2(f_i / f_n) + 12 \cdot m \tag{2.21}$$

m es un numero entero que minimiza el modulo de la distancia d. De esta forma, definimos la función de pesos:

$$w(n, f) = cos^2(\frac{\pi}{2} \cdot \frac{d}{0.5 \cdot l}) \qquad si\ |\ d\ | \leq 0.5 \cdot l \tag{2.22}$$

l hace referencia a la longitud de la función de pesos que para este algoritmo se ha considerado de 4/3 de semitono. La función de pesos minimiza el error estimado que resulta de la inarmonicidad y diferencias de tono que se pueden encontrar en una grabación musical.

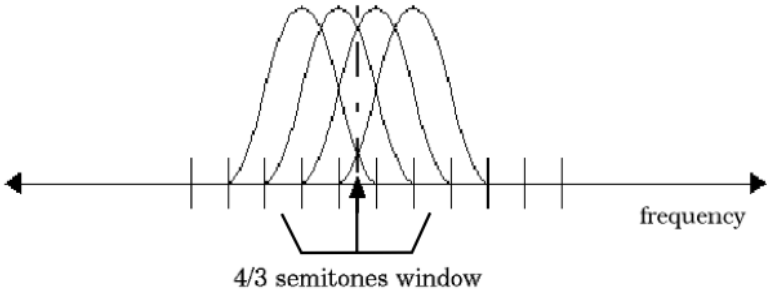


Figura 30. Función de pesos del HPCP.

El espectro de una nota se compone de varias frecuencias, su fundamental y sus armónico que acostumbran a ser múltiples de la frecuencia fundamental (*f, 2 f, 3 f, 4 f,* …). Este hecho afecta a los valores del HPCP que incrementa el valor de los diferentes componentes tonales. Por eso, para que los armónicos contribuyan mayormente a la clase tonal de la frecuencia fundamental, se introduce una nueva función que reduce la intensidad de los armónicos a medida que crece su frecuencia.

$$w_{harm}(n) = s^{n-1} \tag{2.23}$$

donde s es menor que 1, para simular que el espectro decrece con la frecuencia. El valor debería depender del timbre de los instrumentos. El valor escogido en el HPCP es 0.6.

Finalmente se aplica una función de blanqueado espectral. Esta función se adapta a la envolvente espectral. Para conseguir ecualizar los picos espectrales se aplica una técnica que normaliza los picos espectrales. Esta normalización tímbrica corrige la energía entre octavas pero no entre tonos con lo que no acaba de ecualizar el timbre. Es una forma de evitar influencias por los procesos de ecualización o coloración que puedan aparecer en el contenido musical.

2.4.1.5 Normalización

Finalmente en cada fragmento analizado, el valor del HPCP se normaliza entre 0 y 1 con la siguiente expresión:

$$HPCP_{normalized}(n) = \frac{HPCP(n)}{max_n(HPCP(n))} \qquad n = 1,...size \tag{2.24}$$

El proceso de normalización provee de independencia a la dinámica y el volumen. Las diferentes etapas de procesado son diseñados para alcanzar los principales requisitos de las características de croma. La etapa de detección de transitorios elimina los sonidos ruidosos o fragmentos con espectro ruidoso que puede introducir distorsiones en el cromagrama. El blanqueador espectral y la función de pesos aplicados en la etapa de post procesado, añade una considerable robusteza a el cambio de tono influenciado por diferentes ecualizaciones procedimientos.

Además, el paso de normalización añade robusteza al los cambios dinámicos. En general, el HPCP parece ser representativo del contenido tonal de señales monofónicas y polifónicas, pero a parte de estos procesos no añade ningún proceso para mejorar la robusteza a los cambios de timbre o de instrumentación, los cuáles serian necesarios para cumplir todos los requisitos de un descriptor tonal [12]. Por esta razón queremos evaluar su varianza al timbre y mejorar su método de extracción de acuerdo a este requisito esencial.

2.4.2 Chroma DCT Reduced Log Pitch, CRP

Por otro lado, la propuesta de Müller y Ewert, el descriptor tonal CRP, se basa en la escala temperada musical y en el procesado de los coeficientes cepstrum más bajos de la escala tonal (Pitch Frequency Cepstrum Coefficient, PFCC) [14]. La versión del CRP utilizada en este estudio se encuentra disponible en la Müller Chroma Toolbox.[20] La extracción del CRP se basa en suposiciones que ofrecen una robusteza considerable a los cambios de timbre y de instrumentación. En muchos trabajos se ha demostrado que los coeficientes PFCC mas bajos están fuertemente relacionados al timbre, por ello este método los descarta. De esta forma se evita que los resultados del cromagrama estén influenciados por posibles cambios de timbre. Al igual que el HPCP, este nuevo enfoque se puede descomponer en varias etapas:

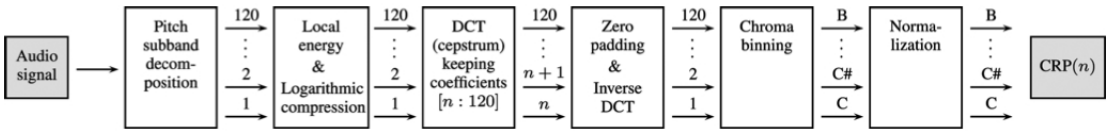


Figura 31. Esquema de la extracción del CRP.

2.4.2.1 Descomposición en bandas tonales

En la primera etapa, se descompone la señal de audio en 88 bandas aplicando un banco de filtros elípticos con múltiple resolución frecuencial que relaciona la magnitud de la frecuencia a la escala musical para obtener una descomposición por semitono musical. La frecuencia central de las 88 bandas corresponden a las notas MIDI de la 21 a la 108. Con este análisis espectral se pretende evitar la baja resolución en las bajas frecuencias de la DFT.

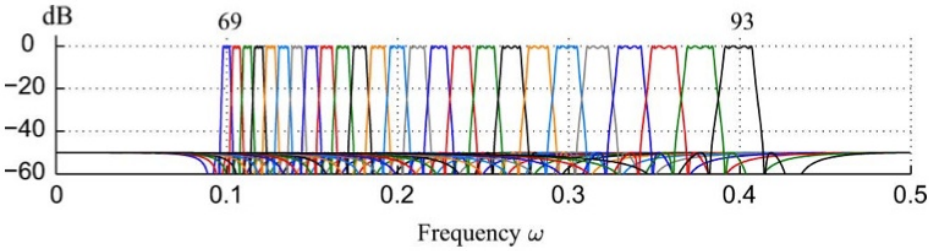


Figura 32. Magnitud en decibelios del banco de filtros para algunos de los filtros [14].

El banco de filtros emplea tres diferentes frecuencias de muestreo: 882Hz para las bandas más bajas (de la 21 a la 59), 4410Hz para las bandas medias (*p=60,…,95*) y 22050Hz para el resto de bandas[21].

[20] http://www.mpi-inf.mpg.de/resources/MIR/chromatoolbox/
[21] Notas MIDI C7-C8

2.4.2.2 Energía local y compresión logarítmica

La siguiente etapa calcula la energía de cada banda y aplica una compresión logarítmica sobre la energía de cada de ellas para proporcionar la sensación de intensidad sonora, pues nuestra percepción del sonido responde de forma logarítmica.

$$plog(n) = log(C \cdot e + 1) \tag{2.25}$$

donde *C* es una constante positiva (*C=100*) y *e* es la energía de cada una de las bandas.

2.4.2.3 Cepstrum DCT

En el paso siguiente, se calculan los coeficientes cepstrum del tono, PFCC, mediante la DCT sobre el vector de la energía de las bandas tonales. Esta etapa es la que consigue aumentar la invarianza al timbre de la extracción del CRP, puesto que se eliminan las oscilaciones de la envolvente espectral relacionadas con el timbre.

$$pred(n) = DCT(plog(n)) \tag{2.26}$$

2.4.2.4 Rellenado de ceros y la inversa de la DCT

El siguiente proceso es el que proporciona una considerable invarianza al timbre puesto que ecualiza las variaciones tímbricas. Se basa en sustituir por ceros los 55 primeros coeficientes PFFC, donde residen las oscilaciones más bajas de espectro, relacionadas directamente con el timbre.

$$npred(n) = \begin{cases} 0 & si \quad n < 55 \\ pred(n) & si \quad 55 \le n \le 120 \end{cases} \tag{2.27}$$

Los siguientes procesos no tienen misterio, se calcula la inversa de la DCT para recuperar el contenido tonal y proyectarlo sobre la escala cromática en una única octava.

$$prevCRP(n) = iDCT(npred(n)) \tag{2.28}$$

2.4.2.5 Correspondencia de tono a croma

El vector resultante se compone de 12 coeficientes que representan las notas musicales.

$$CRP(i) = \sum_{i=0}^{i=mod(n,12)} prevCRP(n) \tag{2.29}$$

donde n es el número de nota MIDI de 1 a 127 y i= 1…12 refiriéndose a las 12 notas.

2.4.2.6 Normalización euclídea

Al igual que en el HPCP, el proceso final es una normalización pero con la diferencia de que en el CRP se divide por el modulo del vector[22] para limitar su rango entre -1 y 1. En el esquema de la figura 24, no se hace referencia, pero en la distribución del Chroma Toolbox para MATLAB, se añade un ultimo proceso para eliminar algunas distorsiones del CRP. Es un proceso de decimado (downsampling) que reduce el numero de vectores croma a uno de cada dos, puesto que mantiene un ratio de 2Hz.

2.5 Objetivos y resultados esperados

Considerando que el CRP es un proceso innovador y que parece obtener considerable robusteza a los cambios de timbre, lo evaluaremos y analizaremos junto al HPCP para conocer sus capacidades en la detección de contenido tonal y sus diferencias. En principio, por su enfoque debería mostrar una varianza al timbre baja y nos sirve como referencia. También es interesante analizar su grado de eficiencia en la estimación de la clase acorde y conocer si mantiene alguna relación con su varianza al timbre.

2.5.1 Evaluación de la extracción de cromagrama

Para evaluar ambos métodos, podemos extraer las características de croma de una colección de audio que contenga diferentes combinaciones de notas con cambios de instrumentación y de octava. Aunque cada una de las versiones presentan diferencias es importante seleccionar parámetros espectrales lo más semejantes posibles para poder evaluar con cierto criterio. En la evaluación que presentamos calculamos dos tipos de medidas para los descriptores tonales, la primera se basa en la comparación de los cromagramas de cada clase acorde con su detección ideal[23] .Este análisis nos permite conocer el grado de eficiencia de la extracción de las características de croma y cuanto de aproximada es su estimación tonal a la estimación ideal. Se mide mediante el coeficiente de correlación y la distancia del coseno resultante de los cromagramas y el perfil acorde. La segunda, se basa en la comparación entre los cromagramas de diferentes clases acorde y nos permite conocer el grado de varianza al timbre de una extracción de croma. Esta medida la introducen Meinard Müller y Sebastian Ewert en su trabajo [16].

2.5.2 Mejoras y explotación en un contexto

Con los resultados de la evaluación podemos darnos cuenta que cabe la posibilidad de experimentar con nuevos métodos y conseguir mejorar los requisitos más importantes de un descriptor tonal en el HPCP, como el grado de varianza al timbre.

---

[22] Esta es una de las características del CRP y que lo diferencia del resto de implementaciones que normalmente presentan un croma con un rango de 0 a 1.
[23] Perfil acorde o perfil clase acorde

Por ello, proponemos diferentes ideas con diferentes finalidades. La primera propuesta es mejorar el problema de la variación tímbrica, para ello proponemos aplicar el procesado de coeficientes cepstrum en la extracción del HPCP. El planteamiento del HPCP, no permite realizar este proceso del mismo modo que en el CRP, puesto que no hay un mapeado frecuencia a tono, pues se mapea directamente al cromagrama. De todos modos, para observar sus efectos sobre el HPCP, proponemos emplearlo sobre la frecuencia, en la etapa de análisis espectral, justo antes que la detección de picos.

La segunda propuesta, trata de analizar la estimación del tono mediante el calculo de la frecuencia instantánea, lo que nos puede permitir una mayor tolerancia al ruido y trabajar únicamente con las componentes en frecuencia reales y dejando a un lado las componentes de los lóbulos secundarios que introduce la función ventana en el dominio en frecuencia.

En caso que la implementación y evaluación de estos métodos en el HPCP tenga éxito podremos mostrar una técnica de mejora con la presentación de la nueva versión del descriptor tonal. Este hecho puede repercutir en aplicaciones de clasificación y recomendación de descriptores de alto y medio nivel que se basan en descriptores tonales.

2.5.3 Implementación de MATLAB a C++

El proceso de evaluación de los descriptores tonales se realizara en MATLAB haciendo uso de funciones propias del lenguaje y funciones que creamos para nuestro objetivo. En principio los cálculos que debemos computar no son pesados para una computadora pero si que precisamos manejar una gran cantidad de archivos de audio y de archivos de texto *.csv* o cromagramas, lo que realizar un numero elevado de operaciones sobre todo para calcular el grado de varianza al timbre. Para evitar saturar el sistema durante la evaluación hemos diseñado una estructura para la evaluación que no requiere manejar mucha información en memoria, puesto que la extraemos y la guardamos como un archivo *.mat* al que podemos acceder en cualquier momento sin tener que cargar toda la información. Además, cuando añadimos los métodos de mejora en el HPCP necesitaremos evaluar como afectan al cromagrama. El hecho de almacenar los resultados de cada evaluación nos permite unirlos en una sola figura y mostrar su evolución.

Los experimentos y prototipos que realizamos en MATLAB para los métodos de mejora permiten visualizar y analizar con rapidez las propiedades de cada una de las propuestas. Posteriormente, transponemos estos enfoques al lenguaje C++, en el desarrollo de un Vamp plugin, uno de los formatos en el que esta implementado el HPCP. Dominar la creación de este tipo de aplicaciones proporciona nuevas herramientas de análisis de grabaciones musicales que puede ser útil para la comunidad y el desarrollo de nuevas aplicaciones basadas en descriptores de nivel medio y alto.

La implementación de la librería dinámica[24] en C++ es multiplataforma con lo que requiere un conocimiento mínimo de programación en todas las plataformas para enfrentarnos con soltura a problemas de incompatibilidades que acostumbran a surgir. Para el desarrollo de un Vamp plugin existe una librería de C++ gratuita, la Vamp plugin SDK[25]. Esta librería define diferentes funciones que facilitan la extracción de características y la estructura de la programación.

2.5.4 Retos transversales

Hay que destacar que en la evaluación de los descriptores tonales hemos de trabajar muchas veces con comandos en la consola o terminal. La implementación del HPCP permite extraer resultados con Sonic Visualizer pero tiene que ser archivo a archivo. Nuestra colección de archivos de audio, contiene 298x24 archivos (7152 archivos), lo que tardaríamos días en analizarlos. Para estos casos podemos utilizar una aplicación que se llama '*Sonic Annotator*[26]', esta aplicación permite extraer los descriptores desde la consola mediante comandos. Es una herramienta capaz de analizar todos los archivos de audio contenidos en un directorio y extraer los resultados en formato *.csv* con un solo comando en consola. Esta herramienta nos permite extraer las características de un descriptor implementado en una librería dinámica en C++ y introducir sus resultados en MATLAB para su posterior análisis.

[24] En Windows su extensión es .dll, en Macintosh .dylib y en Linux .so
[25] http://code.soundsoftware.ac.uk/embedded/vamp-plugin-sdk/index.html
[26] http://www.omras2.org/sonicannotator

3. Metodología

En este capitulo nos centramos en la metodología de evaluación de los descriptores tonales y de los métodos de mejora (Müller Toolbox, CRP y HPCP Vamp plugin). Explicamos con detalle la composición especial de nuestra base de datos para obtener las medidas oportunas. Los resultados que obtenemos de cada uno de los dos descriptores muestran donde destacan cada uno de ellos. Como ya hemos comentado la evaluación se centra en dos medidas, el grado de eficacia de la detección y el grado de varianza al timbre de un descriptor tonal. Lo importante de esta evaluación es la riqueza en combinaciones de notas y de cambios de instrumentación contenidos en la colección de archivos de audio que utilizamos en la evaluación.

3.1 Comparativa de aproximaciones para el cálculo de croma

3.1.1 Colección de sonidos

La colección de archivos de audio que empleamos es exactamente la misma que la del trabajo de Müller [20] con la que evalúan los descriptores tonales. Gracias a su colaboración y consideración de sus autores hemos podido contar con esta colección para poder llevar a cabo esta investigación. Esta colección se compone de diferentes combinaciones de notas grabadas con diferentes instrumentos MIDI y en diferentes octavas. Se compone de 24 archivos de audio monofónicos en formato mp3 con frecuencia de muestreo 22050Hz y 16 bits de profundidad. Cada uno de ellos contiene todas las combinaciones posibles con una (monocordes), dos(duetos) y tres notas (triadas). Por lo tanto, se dispone de 12 monocordes (C, C#, D, D#...), 66 duetos, $\left[\begin{smallmatrix} 12 \\ 2 \end{smallmatrix} \right]$ y 220 triadas , $\left[\begin{smallmatrix} 12 \\ 3 \end{smallmatrix} \right]$. La colección contiene 8 instrumentos: fagot, clarinete, flauta, guitarra, arpa, piano saxo y violín; y cada instrumento esta registrado en 3 octavas. Por lo tanto tenemos 24 archivos que contienen 12 combinaciones posibles en una sola nota, 66 duetos y 220 triadas, en total si los sumamos 298 combinaciones de notas distintas[27]. Cada una de estas combinaciones las denominaremos clase acorde por lo tanto disponemos de 298 clases y cada clase se compone de 24 instancias. Al tratarse de sonidos sintéticos obtendremos resultados que no son definitivos, puesto que lo ideal seria evaluarlos de la misma manera pero con instrumentos de verdad. Sin embargo, muestran un valor aproximado.

Para realizar la evaluación disponemos de dos opciones, una opción seria segmentar los 24 archivos de audio en MATLAB de forma automática y exportarlos a un nuevo directorio, para mantener cierto orden y control sobre los datos para extraer el HPCP con '*Sonic Annotator*'. La segunda opción seria realizar la evaluación a lo bruto sobre todo el archivo de audio. Esta opción no permite tanto control sobre las clases acordes y no permite analizar por instrumentos o clase acorde. Finalmente nos hemos decantado por la primera opción.

[27] Su duración total es de 15 minutos.

Previamente a la evaluación, debemos considerar que el HPCP y el CRP tienen diferente rango de salida en sus resultados. El rango del HPCP es de 0 a 1, mientras que el del CRP es de -1 a 1. Además, el HPCP esta normalizado respecto al valor máximo de cada fragmento, mientras que el CRP esta normalizado respecto al modulo de cada fragmento de croma. Para adaptar los dos descriptores, hemos normalizado el HPCP con su módulo. De esta forma cada uno de los fragmentos de ambas aproximaciones suman uno. El CRP y el HPCP tienen diferente rango, este hecho puede afectar a los resultados de la evaluación, por eso inicialmente propusimos calcular la energía del cromagrama de cada versión de croma. Sin embargo, el cálculo de la energía, empeora los resultados del CRP, puesto que según sus suposiciones la parte negativa tiene su propio significado. Por eso evitamos este paso en la evaluación.

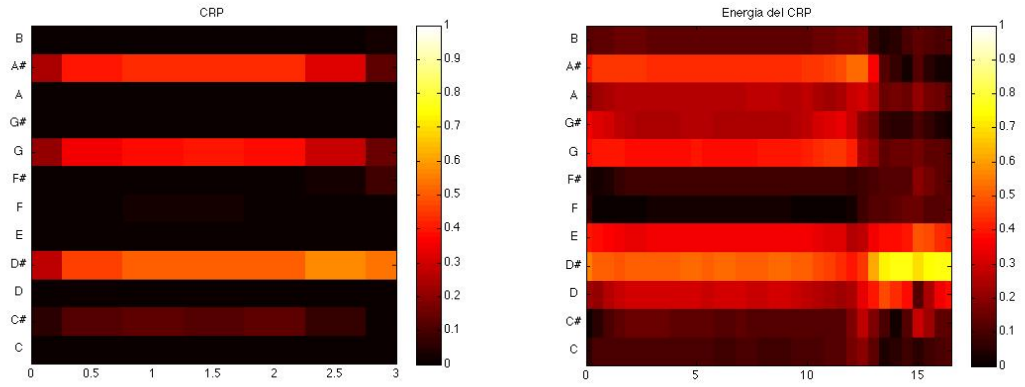


Figura 33. La energía local del CRP. Los valores negativos del CRP esconden cierta energía que pueden introducir errores en los resultados de su evaluación.

Para evaluar la eficiencia y la varianza al timbre utilizamos medidas como la distancia del coseno, puesto que sus resultados no empeoran por los valores negativos del CRP, y el coeficiente de correlación.

Otro aspecto muy importante para una correcta evaluación es garantizar que los dos descriptores tonales realizan las extracción con parámetros espectrales lo más parecido posibles. Para ello hemos tenido que modificar los parámetros por defecto. Los parámetros que hemos utilizado en nuestra evaluación son los siguientes:

```
Parámetros HPCP:
                        Frecuencia de muestreo: 22050
                        Tamaño de la FFT: 4096
                        Función ventana: @hanning
                        Factor de solapamiento: fftWindowLength/2
                        Número de contenedores: 12
```

El CRP utiliza un banco de filtros elípticos con multiresolución en frecuencia y es complicado escoger el mismo parámetro para el tamaño de la ventana o el factor de solapamiento. Finalmente hemos decidido escoger los parámetros por defecto del CRP y adaptar los parámetros del HPCP, generando un nuevo archivo .n3[28] con Sonic Annotator.

[28] http://www.omras2.org/sonicannotator

```
Parámetros CRP =
                            Frecuencia de muestreo: 22050
                            Tamaño de la FFT: 4410
                            Función ventana: @hanning
                            Factor de solapamiento: fftWindowLength/2
                            Número de contenedores: 12
```

La evaluación que proponemos sigue el siguiente esquema:

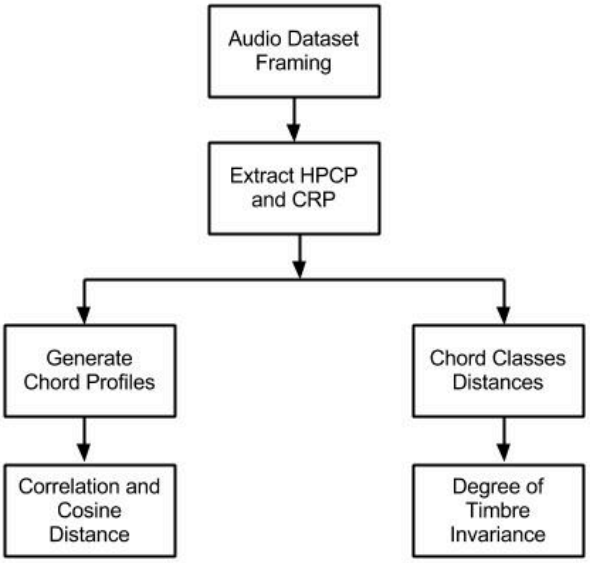


Figura 34. Esquema de la evaluación de los descriptores tonales.

En la siguiente tabla mostramos las funciones implementada para cada tarea:

Etapa del proceso de evaluación	**Funciones MATLAB**
Los archivos de audio se segmentan por acordes.	audiofragmentation.m
Etapa de procesamiento para adaptar los cromagramas.	loadTonalDescStruct.m
Genera el perfil de todas las clases acorde.	createMasks.m
Medidas de evaluación del grado de eficiencia.	evaluationWithMask.m
Medida de evaluación del grado de varianza al timbre.	featuresExtraction.m
Extrae estadísticas de la evaluación de la eficiencia.	displayStats.m
Calcula el grado de variación al timbre	evaluationWithClass.m
Traduce el numero de la clase acorde a tonos	transalateMask2pitch.m

Tabla 1. En esta tabla mostramos la relación entre las tareas y la función MATLAB.

3.1.2 Medidas de evaluación

3.1.2.1 Grado de varianza al timbre

Si suponemos que dos vectores de croma de la misma clase acorde son similares, y a su vez dos vectores de croma de diferente clase acorde deberían presentar grandes diferencias.

Basándonos en esta idea, el grado de varianza al timbre de un descriptor tonal se puede estimar calculando las distancias entre cualquier vector de croma de la misma clase acorde. Las $298 \cdot \binom{24}{2}$ distancias nos permiten conocer la media, μ_I y la desviación estándar σ_I. En el caso que μ_I sea pequeño, indicara un alto grado de invarianza al timbre. Por otro lado, calculamos las distancias entre cualquier vector croma de diferentes clases acorde $24 \cdot \binom{298}{2}$, obtenemos μ_O y σ_O. Para este caso, cuanto mayor sea μ_O indica la alta discriminación de energía entre clases acorde. Finalmente, obtenemos el grado de varianza expresado por las distancias entre las mismas clases y las distancias entre diferentes clases.

$$\delta = \frac{\mu_I}{\mu_O} \tag{3.1}$$

En nuestra evaluación deseamos que este valor de δ sea lo más pequeño posible pues indica el grado de varianza al timbre de un descriptor tonal. Durante la evaluación cuando nos referimos a distancias, hablamos de la distancia del coseno, que se puede calcular para dos vectores de la siguiente manera:

$$dis = 1 - \langle A, B \rangle \tag{3.2}$$

donde A y B son dos vectores croma cualquiera. Para optimizar el calculo sobre una matriz, proponemos calcular la distancia del coseno con la media de la diagonal de la matriz resultante del producto escalar de A y B. Este modo de cálculo reduce el coste computacional y es directo.

$$dis = \frac{1}{N} \sum_{n=1}^{N} (1 - diag(\langle A, B \rangle)) \tag{3.3}$$

3.1.2.2 Grado de eficiencia

Para medir el grado de eficiencia utilizamos medidas como el coeficiente de correlación y la distancia del coseno. Esta medida se basa en la comparación entre un vector de croma de una clase acorde y el perfil de esa clase acorde (detección ideal). El hecho de conocer las notas contenidas en cada clase acorde de la colección de archivos de audio, nos permite generar el perfil teórico para cada clase acorde. Cada perfil acorde esta normalizado con la norma euclídea.

0	0	0	1	0	0	0	1	0	0	1	0
A	A#	B	C	C#	D	D#	E	F	F#	G	G#

Figura 35. El perfil acorde del acorde C sin normalizar.

Mediante un sencillo algoritmo, podemos generar un perfil acorde para cada una de las clases acorde y calcular el coeficiente de correlación entre el perfil y el vector de croma. Este procedimiento se realiza para todas las clases acordes de la colección de archivos de audio (24x298 operaciones). De esta manera podemos calcular la media y cuanto mayor sea indicara mejor eficacia en la detección. El coeficiente de correlación lo hemos calculado de la siguiente forma:

$$\rho_{x,y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \frac{E\left[(X-\mu_x)(X-\mu_y)\right]}{\sigma_x \sigma_y} \tag{3.4}$$

donde X y Y son dos vectores de N dimensiones donde es N es un numero real, σ_x o σ_y es la desviación estándar y μ_x o μ_y la media de cada uno de los vectores. De la forma que hemos estructurado la colección de archivos de audio podemos separar esta medida por instrumentos y analizar cada una de las clases acordes independientemente con un diagrama de cajas.

La otra medida que utilizamos para evaluar el grado de eficiencia es la distancia del coseno. Su cálculo lo hacemos de la misma forma que en la ecuación 3.3, lo único que en vez de utilizar dos vectores de croma utilizamos un vector de croma y su respectivo perfil de la clase acorde.

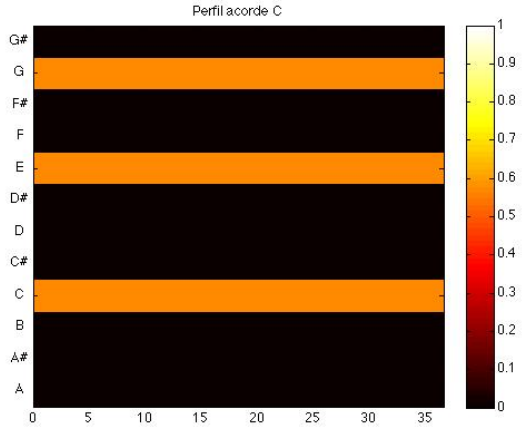


Figura 36. Perfil tonal del acorde C mayor

3.2 Métodos propuestos

Las mejoras propuestas se centran en dos ideas claras: en la primera proponemos mejorar el problema de la variación tímbrica, para ello analizaremos y evaluaremos el el filtrado de coeficientes cepstrum como blanqueador y la segunda, se centra en la estimación del tono mediante el cálculo de la frecuencia instantánea, lo que permite trabajar únicamente con las componentes en frecuencia reales y dejando a un lado las componentes de los artefactos introducidos por el análisis espectral como lóbulos secundarios que introduce la función ventana en el dominio en frecuencia.

3.2.1 Filtrado de coeficientes cepstrum

Como ya hemos mostrado en el capitulo anterior, los coeficientes cepstrum permiten un análisis de la envolvente espectral y nos permite relacionarlo con la frecuencia. El hecho de filtrar o eliminar los primeros coeficientes cepstrum ejerce un cambio sobre la envolvente espectral que tiende a ecualizar la contribución de cada uno de los picos espectrales. Eliminar los coeficientes mas bajos equivale a eliminar las oscilaciones más suaves de la envolvente espectral. Los dos coeficientes más bajos corresponden a la energía y no tienen relación con el timbre. Sin embargo los siguientes coeficientes si que modifican las variaciones de la envolvente espectral que definen el timbre. El problema de este procedimiento es que no es lineal y que depende de la señal que analicemos, por esta razón creemos que es difícil definir un parámetro establecido para cualquier análisis.

Sin embargo, podemos dejar que el usuario defina este parámetro o desarrollar un método inteligente que minimice la distancia entre la media de los picos espectrales, su mínimo y su máximo, con tal de encontrar el mejor filtro en cada uno de los fragmentos de audio analizados.

En este documento, proponemos dos tipos de procesado sobre los coeficientes cepstrum: el relleno de ceros y el filtrado de los coeficientes mediante la función gaussiana. Ambos muestran un comportamiento similar pero con resultados distintos. La sustitución por zeros en los coeficientes cepstrum (Zeroing) se conoce como cepstral liftering y se define con la siguiente ecuación.

$$\overline{\overline{cepstrum(n)}} = \begin{cases} 0 & si \quad n < L \\ cepstrum(n) & si \quad L < n \le N \end{cases} \tag{3.5}$$

Donde L es el parámetro que indica hasta que coeficiente va a ser sustituido por un cero, donde n es el índice de coeficiente cepstrum donde N es el numero total de coeficientes cepstrum. Puesto que Zeroing es un filtro un poco agresivo, valoramos la posibilidad de aplicar un filtro que no eliminara totalmente los primeros coeficientes si no que atenuará su aportación. En el filtrado de cepstrum mediante la función gaussiana, primero necesitamos definir la forma de la gaussiana mediante el parámetro alfa:

$$w_G(n) = e^{-\frac{1}{2}(\alpha\frac{2\cdot n}{M})} \tag{3.6}$$

Donde M indica el numero total de muestras de la función, y donde $-\frac{M-1}{2} \le n \le \frac{M-1}{2}$

El valor alfa es inversamente proporcional a la a la desviación estándar de la función gaussiana, a partir de este valor definimos el ancho de banda del filtro pasa altos.

$$\sigma = \frac{M}{2\alpha} \tag{3.7}$$

Cuando alfa vale uno se genera una gaussiana con desviación estándar igual a la mitad del numero de coeficientes cepstrum. De esta forma podemos definir el filtrado de coeficientes cepstrum con un solo parámetro.

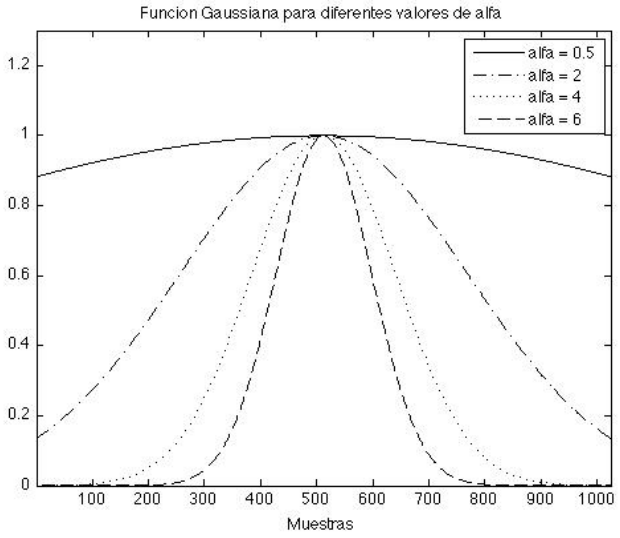


Figura 37. Función gaussiana definida por el parámetro alfa.

Para aplicarlo como filtro pasa altos definimos una función gaussiana de $M = 2 \cdot N$ muestras y conservamos la mitad de la función, de 1 a N muestras, y lo multiplicamos por los N coeficientes cepstrum. Una vez generado el filtro podemos aplicarlo sobre los coeficientes calculando el producto entre los dos vectores.

$$G(n) = w_G(n) \qquad n = 1...N \tag{3.8}$$

$$\overline{\overline{cepstrum}}(n) = G(n) \cdot cepstrum(n) \qquad n = 1...N \tag{3.9}$$

Es importante mencionar que una vez realizadas cualquier de los procesos, se debe calcular la inversa de la FFT para recuperar el espectro modificado y continuar con la extracción de características mediante la siguiente expresión:

$$\tilde{X}(n) = 10^{FFT(\overline{\overline{cepstrum}}(n))} \tag{3.10}$$

donde $\tilde{X}(n)$ es el nuevo espectro donde se realiza la detección de picos espectrales.

3.2.2 Estimación del tono mediante la frecuencia instantánea

La estimación de tono mediante la frecuencia instantánea se realiza en el análisis espectral. Es un método propuesto en 1966, por Flanagan, J. [21]. Además algunas investigaciones proponen la estimación de la frecuencia fundamental mediante la frecuencia instantánea o la dominancia del espectro [22].

La frecuencia instantánea se define como la derivada direccional de la fase y permite una estimación precisa del tono en presencia de ruido o de otras distorsiones espectrales.

$$\phi(k) = \frac{\partial}{\partial k} tan^{-1}(X(k)) \tag{3.11}$$

donde X es el espectro complejo y $k = 1 \ldots N_{frame} - 1$, indica cada una de las funciones bases del dominio en frecuencia. El cálculo de la derivada de la fase, lo resolvemos como propone Flanagan en *'Phase Vocoder'* [21]. Este procedimiento actúa en el análisis espectral (figura 21), paralelamente se calcula análisis espectral del mismo fragmento ponderado con la derivada de la función ventana utilizada en el análisis.

$$X(k) = \sum_{n=-\frac{N_{FFT}}{2}}^{\frac{N_{FFT}}{2}-1} w(n) \cdot x(n) \cdot e^{-j2\pi nk/N_{FFT}} \tag{3.12}$$

$$\phi(k) = f + \frac{a(k) \cdot \partial/\partial k\, b(k) - b(k) \cdot \partial/\partial k\, a(k)}{a(k)^2 + b(k)^2} \tag{3.13}$$

donde $k = 1 \ldots N_{frame}-1$, $\phi(k)$ es la frecuencia instantánea, $a(k)$ y $b(k)$ corresponden a la parte real y imaginaria del espectro y f es la frecuencia normalizada entre 0 y π. De esta forma conseguimos mapear la frecuencia central de cada componente frecuencial de la STFT a su frecuencia instantánea correspondiente.

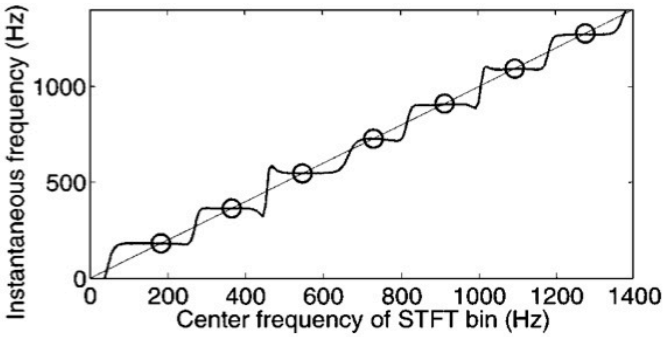


Figura 38. Frecuencias vs. frecuencia instantánea de una señal de voz.

Para conocer $\partial/\partial k$ $(a(k))$ o $\partial/\partial k$ $(b(k))$ realizamos el siguiente cálculo:

$$\frac{\partial}{\partial k} X(k) = \sum_{n=-\frac{N_{FFT}}{2}}^{\frac{N_{FFT}}{2}-1} \frac{\partial}{\partial n} w(n) \cdot x(n) \cdot e^{-j2\pi nk/N_{FFT}} \tag{3.14}$$

donde $\partial/\partial n(w(n))$ es la derivada de la función ventana normalizada con el área de la función ventana $w(n)$.

En la figura 38, observamos que existen diferentes intervalos de frecuencias que apuntan hacia la frecuencia central de cada función base de la STFT, indicando que su contribución en la magnitud del espectro realmente es fruto de la frecuencia central y que sus diferencias se pueden contrarrestar con la derivada de la fase.

En nuestro trabajo analizaremos la influencia de los parámetros espectrales como: la función ventana y su tamaño, el tamaño de la FFT y el factor de solapamiento entre fragmentos. Para ello analizamos los resultados de señales armónicas, señales de voz, señales monofónicas y polifónicas, donde la correspondencia entre frecuencia central y frecuencia instantánea, no es tan marcada y evidente como en la figura anterior, puesto que el contenido armónico es mucho mayor y la detección se complica.

El cálculo de la IF, nos permite encontrar la frecuencia central de cada una de las componentes en frecuencia de la STFT y sobre ellas podemos estimar la frecuencia fundamental o estimar su tono musical. La detección de la frecuencia instantánea se puede llevara a cabo mediante el calculo de los cruces que hay entre la recta de frecuencias y la función de la frecuencia instantánea.

$$CruceIF(n) = \begin{cases} 1 & si \quad \phi(n) < f(n) \quad \wedge \quad \phi(n-1) > f(n-1) \\ 0 & otros \end{cases} \tag{3.15}$$

En algunos casos se pueden dar discontinuidades o distorsiones que introducen cruces en el mapeado que no corresponde a un pico. Para evitar estas falsas detecciones aplicamos una condición geométrica que únicamente considere un cruce como una frecuencia instantánea si el ángulo que forma con la siguiente muestra esta dentro del rango 15-75º.

$$\angle = cos^{-1}\left(\frac{A \cdot B}{\| A \| \cdot \| B \|}\right) = cos^{-1}\left(\frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}\right) \tag{3.16}$$

donde *A* corresponde al mapeado de las frecuencias instantáneas (función escalera de la figura 38)y *B* a la frecuencia central de cada función base de la STFT (diagonal de la figura 38).

Para el análisis de las frecuencias instantáneas, hacemos uso de la función STFT, que nos permite escoger cualquier parámetro espectral (tamaño de la FFT, factor de solapamiento y función ventana) y analizar sus resultados. Además hemos tenido que desarrollar diferentes funciones en MATLAB que nos permiten: generar cualquier función ventana y su derivada, la detección de frecuencias instantáneas en el mapeado mediante el cruce de ceros y el cálculo del ángulo de similaridad.

3.3 Mejora de invarianza al timbre

Tras el análisis de estos métodos, la idea es incorporarlos al descriptor tonal HPCP para mejorar su invarianza al timbre y crear una nueva versión. Una vez dispongamos de la nueva versión del HPCP, creemos que es de gran interés conocer la influencia de estos métodos en la extracción del vector de croma. Por ello, finalmente estimaremos el grado de mejora de la nueva versión de HPCP aprovechando las medidas utilizadas en la comparativa de aproximaciones de croma.

4. Resultados

4.1 Resultados de la evaluación de los descriptores tonales: HPCP y CRP.

En la evaluación de los descriptores tonales HPCP y CRP, extraemos dos medidas de sus resultados con la colección de archivos de audio: el grado de varianza al timbre, descrito por δ, y el grado de eficiencia de la estimación de acordes, descrito por la distancia del coseno y el coeficiente de correlación.

4.1.1 Estimación del grado de varianza al timbre

Como hemos comentado en el capitulo anterior, el grado de varianza al timbre se calcula mediante el cociente de μ_I y μ_o. En la tabla siguiente mostramos los resultados obtenidos para el HPCP y el CRP estimando su grado de varianza al timbre:

HPCP	0.3456	0.0528	0.5516	0.0837	0.6265
CRP	0.3391	0.0865	0.9923	0.0864	0.3417

Tabla 2. Resultados de la estimación del grado de varianza al timbre

Como esperábamos, el grado de invarianza al timbre del HPCP es mayor que el del CRP. Podemos observar que la diferencia entre los grados de varianza de cada uno de los descriptores, son fruto del valor μ_o, puesto que μ_I es similar en ambos casos y el grado de varianza se define por su cociente. Esta diferencia indica que los vectores croma del HPCP de diferentes clases acordes presentan cierta similitud, si no μ_o seria mucho mayor. El CRP consigue discriminar y penalizar las diferencias entre las diferentes clase acorde. Los resultados que presenta el autor del CRP en su publicación no coinciden con los obtenidos en nuestra evaluación. En todo momento hemos seguido los mismos pasos que proponen pero parece ser que su evaluación se ha realizado únicamente sobre el rango positivo del CRP o no han calculado correctamente las diferencias entre clases. Estamos conversando para estudiar estas diferencias.

4.1.2 Estimación del grado de eficiencia

El grado de eficiencia se puede estimar para toda la colección para cada clase acorde o para cada instrumento. Las dos medidas utilizadas muestran que más o menos son complementarias, cuando la distancia del coseno aumenta, la correlación disminuye y viceversa.

			Mediana	Max	Min
HPCP	0.7040	0.0415	0.7080	0.7968	0.5884
CRP	0.6620	0.1204	0.6887	0.87722	0.2826

Tabla 3. Resultados de la estimación del grado de eficiencia a partir del coeficiente de correlación ρ.

			Mediana	Max	Min
HPCP	0.2339	0.0303	0.2359	0.3064	0.1631
CRP	0.5040	0.0915	0.4848	0.7936	0.3534

Tabla 4. Resultados de la estimación del grado de eficiencia a partir de la distancia coseno.

En el caso de ρ queremos que el resultado sea lo más alto posible y en el caso de la distancia del coseno, lo más bajo posible. Los resultados de la evaluación por instrumentos, los presentamos mediante un gráfico de cajas:

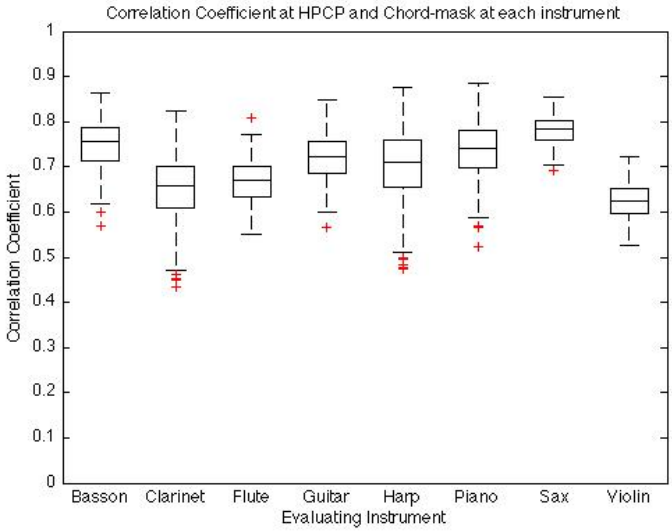


Figura 39. Resultados de la evaluación del HPCP por instrumentos

El grado de eficiencia del HPCP cambia para cada instrumento. El violín muestra la mayor diferencia con los perfiles acorde. El clarinete también destaca por su bajo valor y la amplitud de sus cuartiles. De todas formas, ningún instrumento esta por debajo de 0.5 lo que indica que aunque hay diferencias, el grado de eficiencia es bastante alto en todos los instrumentos. El saxo destaca por su eficacia en la detección y por presentar una distribución de las medidas de correlación de sus perfiles acorde bastante concentrada. Esto se debe a que el sonido de saxo es muy armónico y su desviación armónica es muy baja, más tratándose de un sonido MIDI. Si se tratará de una colección de sonidos de instrumentos reales seguramente el resultado cambiaria.

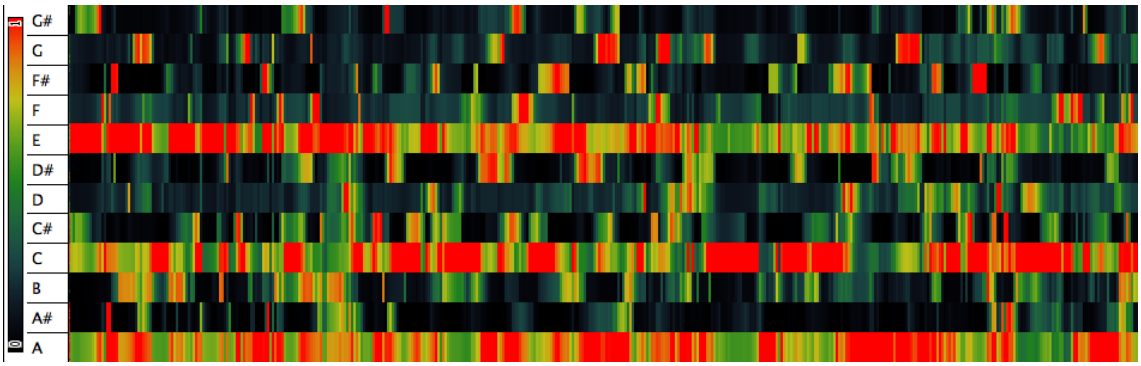


Figura 40. Resultados del HPCP con un acorde A_m de un sonido MIDI de saxo (clase 110).

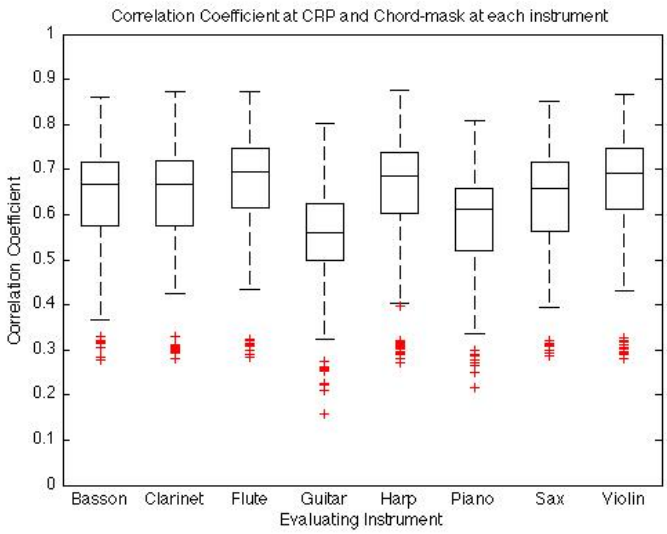


Figura 41. Resultados de la evaluación del CRP por instrumentos

Los resultados de la evaluación del CRP por instrumentos muestran su bajo grado de varianza al timbre y como podemos observar en la figura 41 las distribuciones para cada instrumento son muy parecidas. Además, la mediana de los resultados de cada instrumento es similar, exceptuando la guitarra y el piano.

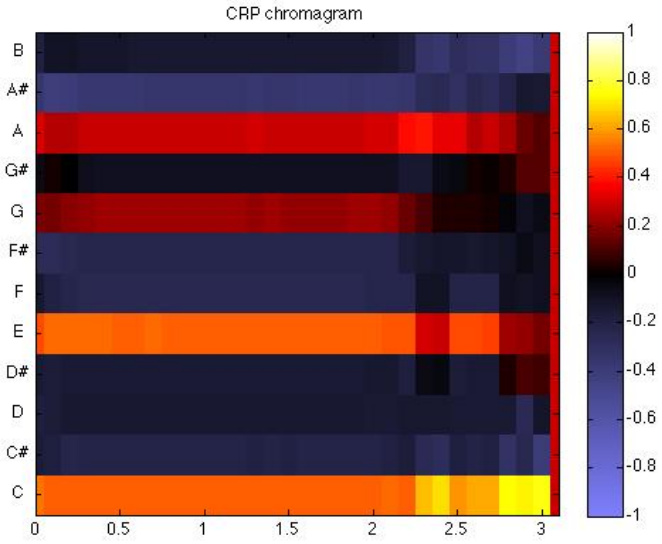


Figura 42. Resultados del CRP con un acorde Am de un sonido MIDI de saxo (clase 110).

Si nos fijamos en las barras de colores, vemos que el HPCP extrae una mayor intensidad relativa de las notas contenidas en el acorde de A_m (A-C-E), alcanzando el valor máximo en varias ocasiones. Esto es debido a que el HPCP esta normalizado con el máximo. Mientras en la figura 42, vemos que las notas extraídas por el CRP se mantienen en un rango de 0.5, debido a la normalización euclideana del vector croma, que hace la energía quede repartida y que la suma de esta sea igual a 1. Por otro lado, la detección del CRP ha añadido una nota (G) que no compone el acorde, seguramente introducida por armónicos que contribuyen en las notas del acorde.

Este es un caso muy concreto pero si dependiera de este aspecto en concreto, podríamos afirmar que el HPCP extrae una mejor aproximación del croma que el CRP, puesto que no introduce energía en notas que no formas el acorde de forma continuada.

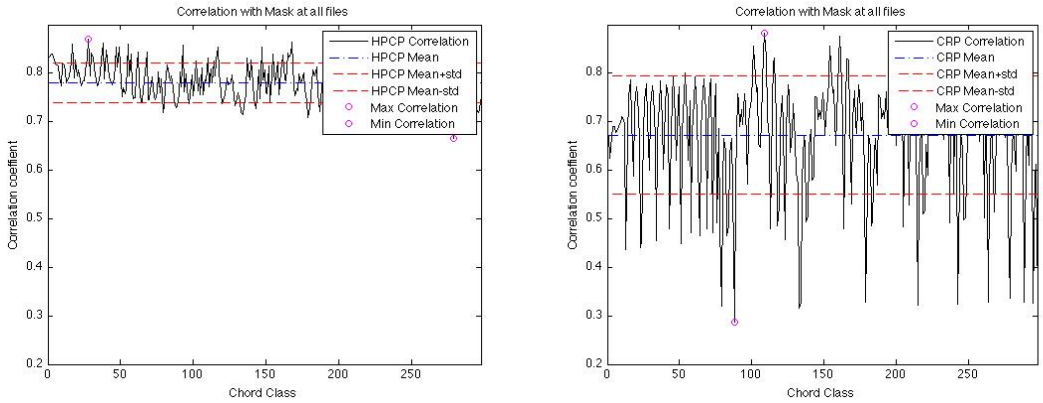


Figura 43. Resultados de la evaluación por clase acorde del HPCP y CRP (I).

En la figura anterior vemos que el resultado del CRP, muestra que la media del coeficiente de correlación por clase acorde oscila según los intervalos de notas que contiene cada clase. Por lo tanto, tiene una baja varianza al timbre pero una alta varianza al cambio de intervalos musicales. Observemos los resultados de la distancia del coseno.

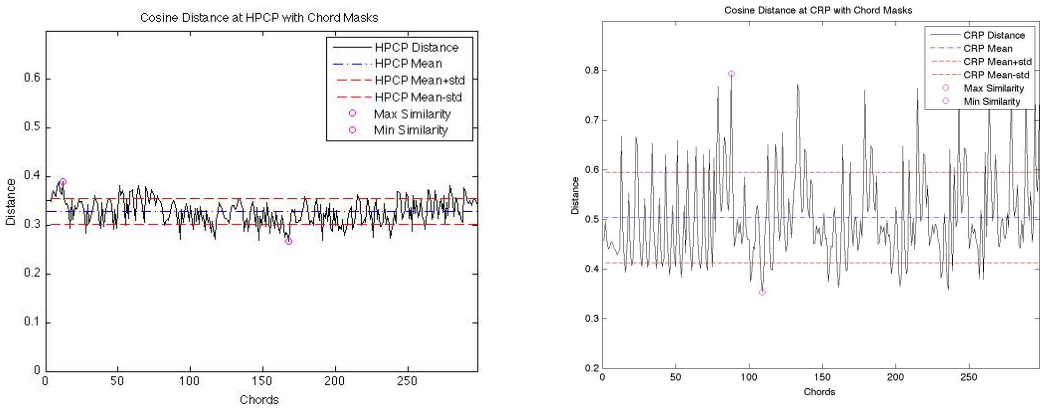


Figura 44. Resultados de la evaluación por clase acorde del HPCP y CRP (II).

Con la distancia coseno muestran resultados muy parecidos, casi podemos considerarlos inversamente proporcional. Aunque los máximos y mínimo no siempre coinciden. El CRP sigue mostrando un patrón que se va acortando cuando cambia la primera nota de los acordes. El HPCP no muestra un patrón definido, más bien parece aleatorio o al menos no se parecía una repetición tan clara. A continuación, presentamos los mismos resultados con un gráfico de cajas, que nos permite analizar valores atípicos y la distribución de las 298 clases acorde.

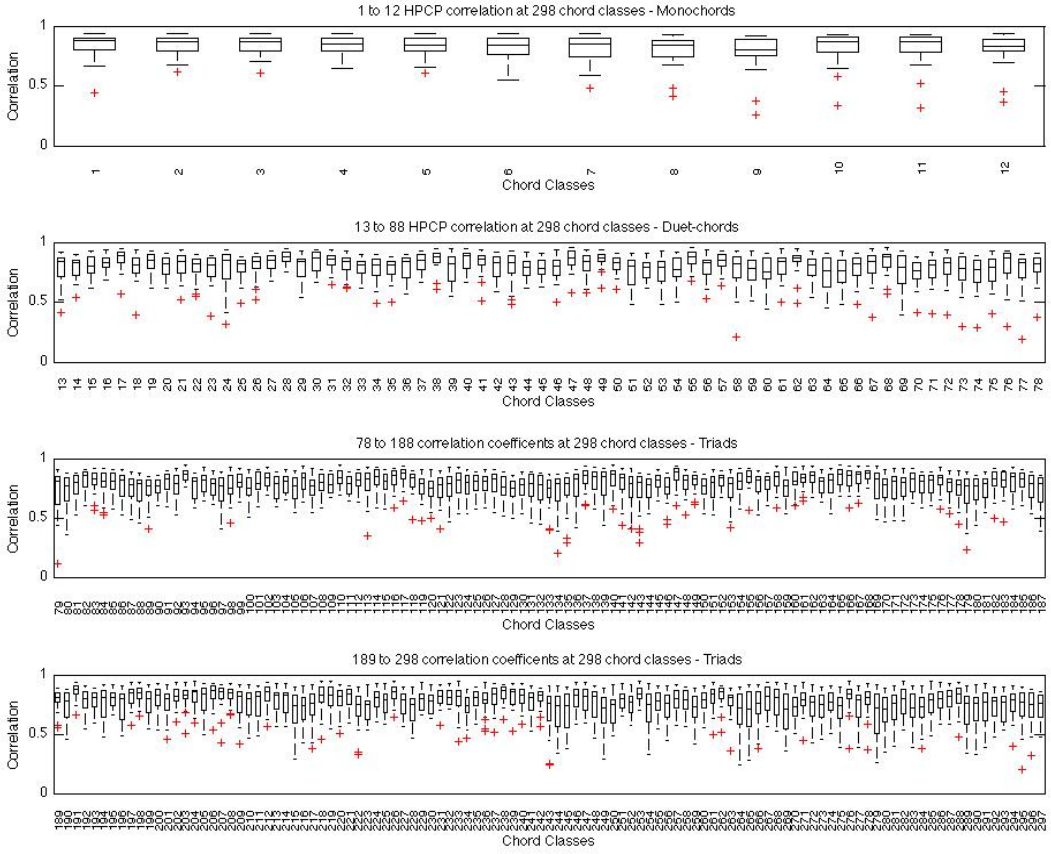


Figura 45. Gráfico de cajas de la correlación de las clases acorde del HPCP.

Un grafico de cajas nos permite analizar cuales son los intervalos de notas más polémicos o los que influyen a la baja el coeficiente de correlación. Si tenemos en cuenta que las clases acorde de una sola nota muestran una distribución de los resultados muy concentrada. En general, se mantiene la media aunque según la combinación de notas la distribución relacionada a cada clase acorde cambia. Por otro lado, una clase acorde no supera los dos valores atípicos, menos las clases 143 y 238 que presentan 3 valores atípicos. Éstos pueden estar introducidos por instrumentos como: la flauta y el violín, que son para los que el HPCP obtiene peores resultados (Figura 39). Podríamos analizar la composición de notas y los intervalos de las clases acorde que mejor o peor resuelven, pero no hemos encontrado ninguna relación aparente. Por lo general en el HPCP la mediana de cada clase acorde es muy parecida, sin embargo, lo que difiere en cada clase son sus distribuciones. Si comparamos las distribuciones de las clases acorde del HPCP con las del CRP (figura 46), vemos que en las del CRP oscila la mediana pero las distribuciones son más parecidas. Esto indica que el HPCP presenta una peor invarianza al timbre, puesto que la distribución de las clases acordes cambian según el instrumento. Los resultados de la distancia coseno en un grafico de cajas muestra el mismo patrón, ya que como hemos visto casi es inversamente proporcional.

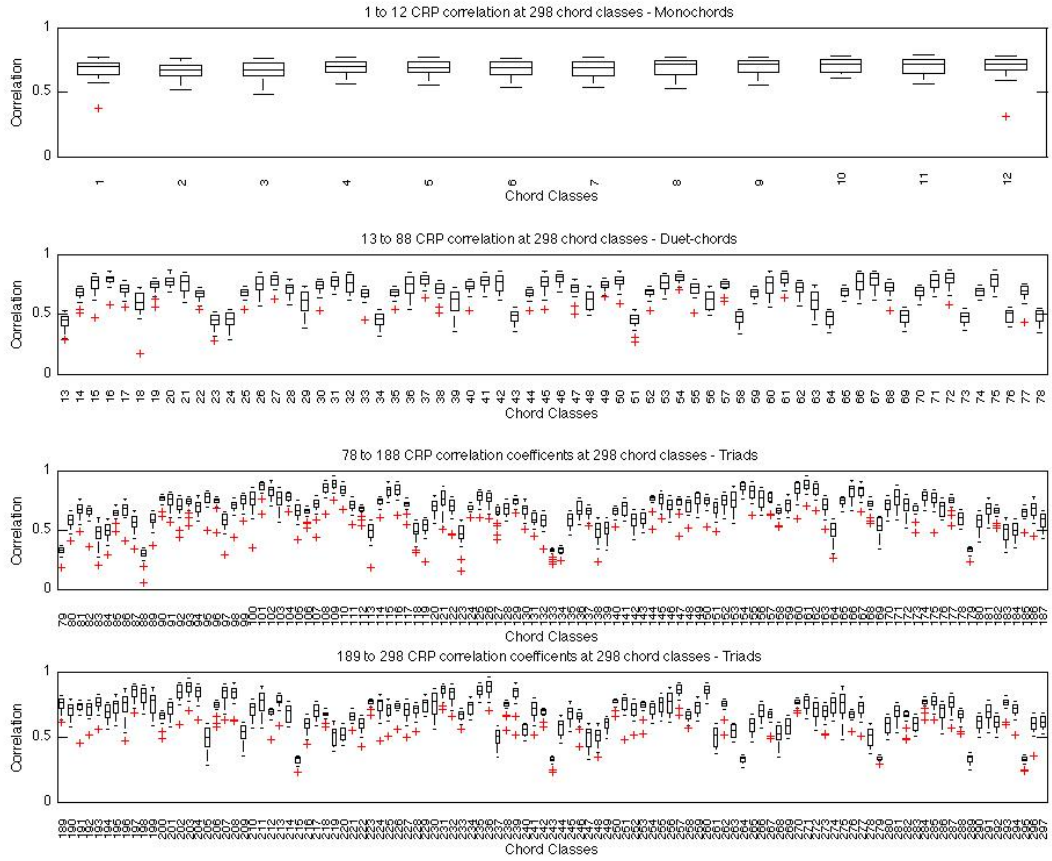


Figura 46. Gráfico de cajas de la correlación de las clases acorde del CRP.

En el grafico de cajas por clases del CRP, nos damos cuenta que su resultado depende de la clase acorde y que presenta un patrón irregular cuando aparecen algunos intervalos tonales concretos. Para conocer esos intervalos hemos de realizar un análisis de intervalos musicales en cada uno los valores atípicos. Para ello, clasificaremos los tonos musicales cualitativamente.

Distancia en semitonos	Intervalo ascendente	Intervalo descendente	Relación entre intervalos
1	Segunda menor	Séptima mayor	2m/7M
2	Segunda mayor	Séptima menor	2M/7m
3	Tercera menor	Sexta mayor	3m/6M
4	Tercera mayor	Sexta menor	3M/6m
5	Cuarta justa	Quinta perfecta	4j/5j
6	Cuarta aumentada	Quinta disminuida	5j/4j
7	Quinta perfecta	Cuarta perfecta	#4/5dim
8	Sexta menor	Tercera mayor	6m/3M
9	Sexta mayor	Tercera menor	6M/3m
10	Séptima menor	Segunda mayor	7m/2M
11	Séptima mayor	Segunda menor	7M/2m
12	Octava superior	Octava inferior	

Tabla 5. Clasificación cualitativa de las distancia en semitonos.

Si vamos a la figura 46 y nos fijamos al comienzo de los duetos de las clases acorde (clase 13) vemos que aparece el primer valor atípico que corresponde a un intervalo de segunda menor (C-C#) que presenta un coeficiente medio de correlación de 0.4443.

La clase 16, corresponde a un intervalo de tercera mayor, C-E y consigue un alto rendimiento. La clase 18 muestra un pico intermedio y corresponde a un intervalo de cuarta aumentada. El siguiente pico aparece en la clase 23 y la 24, ambos intervalos corresponden a un intervalo de segunda menor, en el primer caso descendente y en el segundo ascendente.

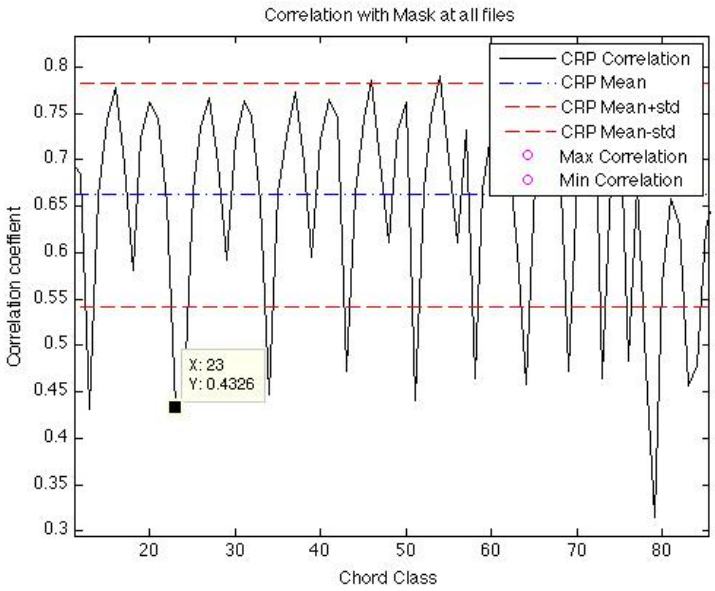


Figura 46. Ampliación de la vista la correlación de las clases acorde (duetos).

Si ampliamos la vista sobre esta zona de la media de la correlación del CRP, observamos que este patrón se repite a lo largo de todas las clases. Se va reduciendo cada vez que la primera nota del intervalo cambia pues siempre hay un intervalo menos, si no lo repetiríamos la misma combinación de notas varias veces.

En el caso de las clases triadas, el patrón se repite. Aquellos acordes que contengan un intervalo de segunda menor (descendente o ascendente) o cuarta aumentada resultan en un pico de baja correlación. Estos intervalos son los intervalos musicales que presentan menor consonancia. Por estos resultados podemos decir que el CRP es variante a los intervalos musicales de mayor disonancia.

La clase 78 coincide con la primera clase acorde de los triadas y esta formado por dos intervalos de segunda menor (C-C#-D), lo que corrobora nuestra suposición. Cuantos más intervalos disonantes compongan el acorde mayor error introduce en la detección.

Por otro lado, la clase con el resultado mínimo es la clase 88. Ésta se compone de las notas C-C#-B, lo que es lo mismo dos intervalos de segunda menor (ascendente y descendente) o el cromatismo entre B-C-C#.

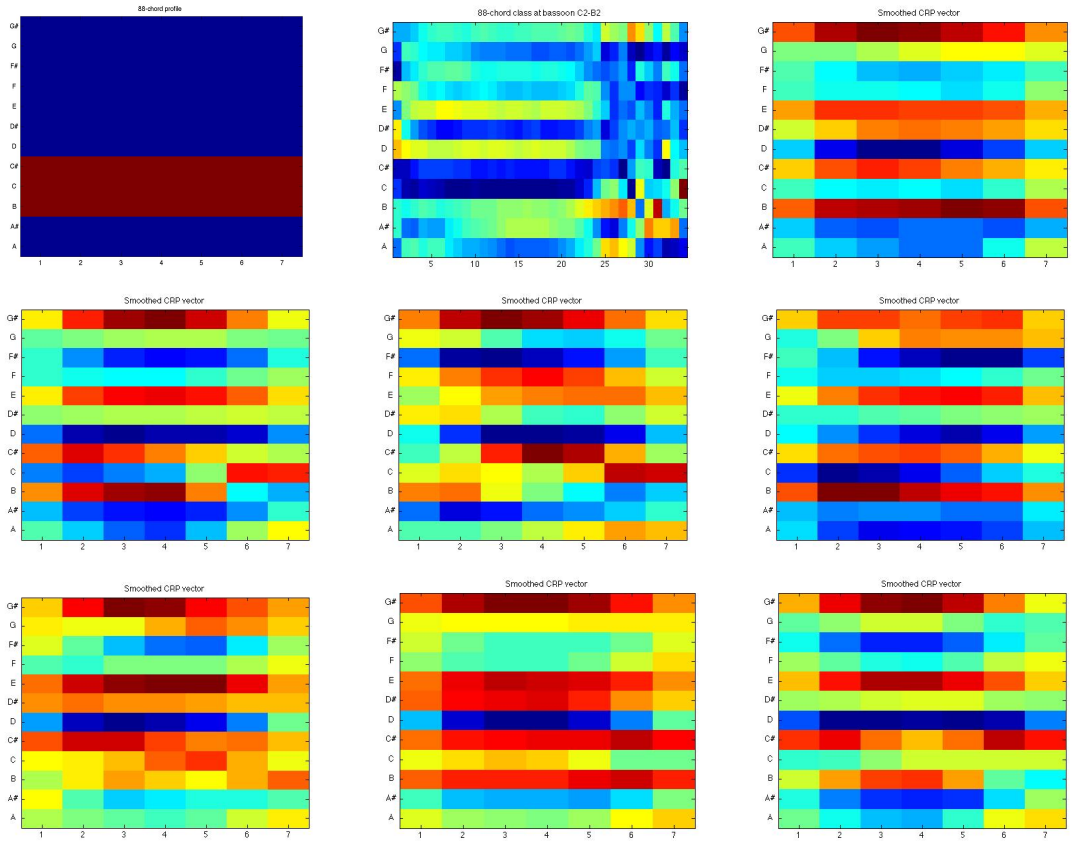


Figura 48. Resultados del CRP para la clase acorde 88 con diferentes instancias.

En la primera figura, podemos ver el perfil de la clase acorde 88 (B-C-C#), el resto osn extracciones del CRP de diferentes instrumentos. Aunque se trata de una combinación de notas difícil de encontrar en un contexto musical por su alta disonancia, los resultados del CRP extraen notas que no tienen nada que ver con la realidad y presenta una muy baja eficiencia. En estos casos, la barra de colores va del azul (0) al rojo (1).

4.2 Resultados de los métodos propuestos

4.2.1 Experimentos en MATLAB

En este apartado comentaremos los experimentos que hemos realizado tanto en MATLAB como en C++ (SDK Vamp plugin) y los resultados que hemos obtenido.

4.2.1.1 Experimentos con filtrado de coeficientes cepstrum

Durante el desarrollo inicial de este método, la idea era implementar lo que se conoce como *'cepstral liftering'*. Esta técnica se centra en sustituir los coeficientes cepstrum más bajos por ceros. Es la misma técnica que emplea el CRP en su extracción (Apartado 2.4.2.4).

Este proceso influye sobre el espectro, alterando directamente la magnitud de cada uno de los picos espectrales, según el coeficiente que eliminemos y su oscilación de la envolvente espectral que le corresponda. Por ello, analizamos este proceso y observamos la variación del espectro con el relleno de ceros, además de comentar sus ventajas y sus inconvenientes para extraer una conclusión final.

La ventaja más importante que ofrece, como acabamos de comentar, es que permite realizar una ecualización tímbrica, y su mayor inconveniente, es que no tenemos un control directo sobre sus efectos como blanqueador espectral, además de que según que coeficientes eliminemos la magnitud del espectro puede cambiar drásticamente su rango.

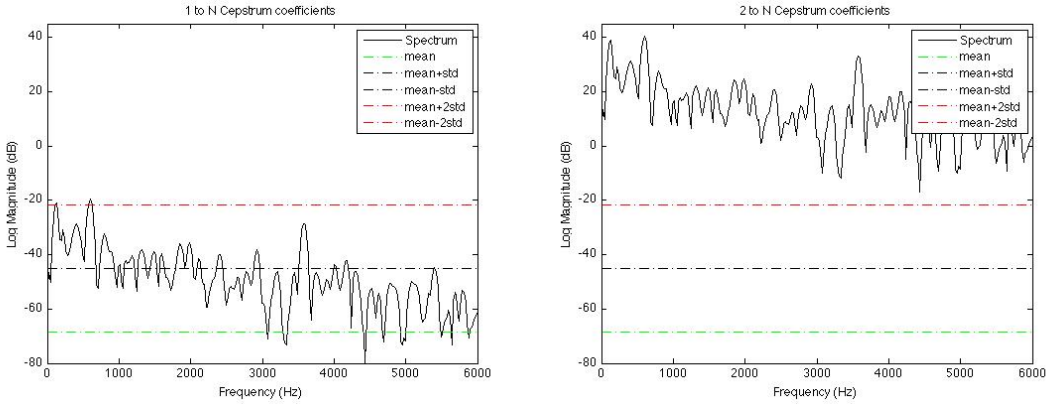


Figura 49. Diferencias: espectro original y procesado cepstrum de un sonido polifónico. A la derecha el espectro después de eliminar el primer coeficiente cepstrum, a la izquierda el espectro original.

Los dos primeros coeficientes contienen información sobre la energía y no ecualizan el timbre. En la figura 49, observamos que al eliminar el primer coeficiente cepstrum del fragmento de un sonido polifónico, la magnitud del espectro se amplifica situándose por encima de 0dB, sin aplicar ninguna ecualización de los picos. Pasar de una magnitud positiva a una negativa puede generar errores en la detección de picos espectrales del HPCP. En cambio si conservamos los dos primeros coeficientes y eliminamos los siguientes 56 coeficientes, conseguimos el efecto deseado, similar al de un blanqueador espectral.

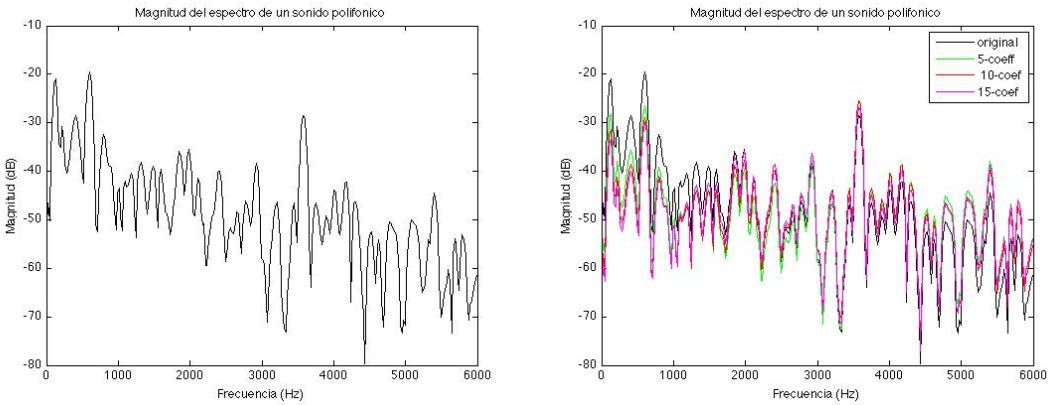


Figura 50. Sustitución de los coeficientes cepstrum por zeros como blanqueador espectral.

Para poder visualizar estos cambios, hemos añadido tres nuevas etapas de procesado, posteriores al análisis espectral, que nos permite eliminar coeficientes cepstrum y recuperar el espectro ecualizado mediante la inversa de la transformada de Fourier.

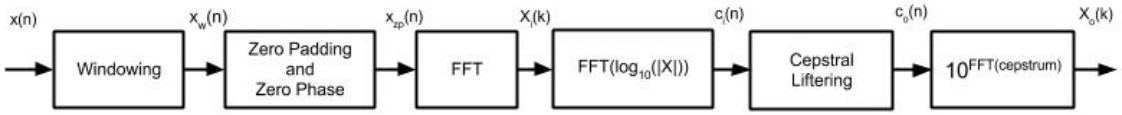


Figura 51. Esquema de la sustitución por ceros de los coeficientes cepstrum más bajos.

Cuando se aplica este proceso eliminando menos de 20 coeficientes, conseguimos un efecto sobre el espectro similar al de un filtro pasa-altos de preénfasis, filtro muy utilizado en la etapa inicial de muchas aplicaciones de audio para compensar el contenido en frecuencia. En el anexo I, se encuentran los experimentos de esta técnica con sonidos monofónicos (guitarra y piano) y polifónicos.

Posteriormente a este experimento, barajamos la idea de aplicar un filtro pasa altos sobre los coeficientes cepstrum mediante la función gaussiana. El filtrado de coeficientes cepstrum consigue un efecto semejante al del relleno de ceros, pero con algunas diferencias.

Como la mayor parte de la estructura la hemos implementado para el relleno de ceros, nos centraremos en generar una función gaussiana a partir de su desviación estándar. Esta función permite filtrar los coeficientes cepstrum aplicando una ponderación con la función gaussiana, donde los coeficientes mas bajos se ven modificados. El ancho de banda del filtro lo definimos mediante el parámetro alfa que es inversamente proporcional a la desviación estándar. En el anexo I, se pueden consultar los resultados obtenidos con una señal polifónica, donde definimos que el mejor rango de alfa esta entre 0.5 y 2.5.

Por los resultados obtenidos en los experimentos de filtrado de coeficientes cepstrum (figura 50), vemos que puede ser una buena opción incluirlo en el HPCP, aplicándolas sobre el espectro justo antes de la etapa de detección de picos espectrales.

4.2.1.2 Experimentos con la estimación de tono mediante las frecuencias instantáneas

Tras elaborar el siguiente esquema (figura 52), hemos practicado diferentes experimentos que nos permiten visualizar el efecto de cada una de las funciones ventana y los parámetros espectrales.

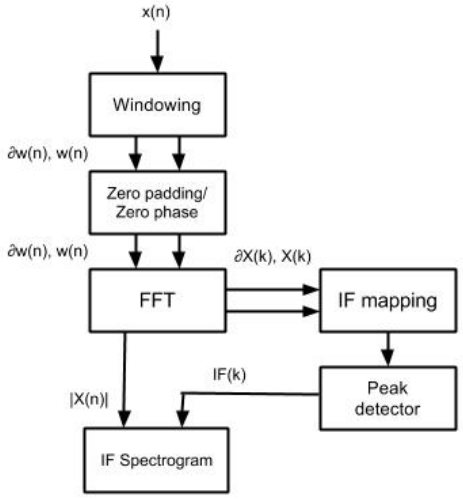


Figura 52. Esquema para la el cálculo de la frecuencia instantánea

Durante los experimentos con las frecuencias instantáneas, hemos analizado los efectos de diferentes funciones ventanas y parámetros espectrales. Para ello hemos implementado todas las funciones ventana y su derivada direccional (Anexo II) y hemos encontrado que algunas de las funciones ventana que mejor funcionan son: Hann, Hamming, Blackman, Blackman Harris y Nutall.

Los parámetros espectrales que permiten detectar los picos en cada fragmento sin introducir distorsiones son: un tamaño de la FFT de 2048 muestra, una ventana de 1024 muestras y un factor de solapamiento de 256 muestras. Con tamaños de FFT superiores, la detección introduce frecuencias instantáneas o picos espectrales que no existen y no conseguimos obtener una buena estimación del tono. Un tamaño de la FFT de 2048 muestras para archivos de audio con frecuencia de muestreo estándar (44.1KHz) equivale a una resolución en frecuencia de 44100/2048 = 21.53Hz. En altas y medias frecuencias supone una buena resolución pero en las bajas frecuencias es una resolución demasiado baja que puede introducir errores en la estimación de tono. Además, en ninguno de los experimentos planteados hemos conseguido que la detección de la IF sea estable cuando nos encontramos en frecuencias superiores a los 3KHz. Este efecto puede ser contradictorio, pero como la mayor parte de la energía que contribuye al tono la encontramos en este rango de frecuencias inferior a 3KHz, no creemos que este aspecto sea contra producente para un descriptor tonal.

Sin embargo, durante el trabajo hemos estimado que por su dependencia en los parámetros espectrales y su coste computacional (el cálculo de dos FFT), no es relevante para nuestras mejoras, puesto que lo que ganemos por un lado se puede perder por el otro (Anexo II). Aunque con las herramientas que hemos desarrollado no descartamos la opción de añadir y evaluar el proceso en trabajos futuros. De todos modos hemos implementado el proceso en C++ en un Vamp plugin independiente al HPCP que permita a la comunidad analizar y visualizar las frecuencias instantáneas en un espectrograma, como se propone algunas investigaciones [23]. Esta aplicación puede facilitar añadir el proceso al HPCP en un futuro.

4.2.2 Implementación de Vamp plugins en C++

Para el desarrollo de un Vamp plugin disponemos de una librería Open Source, SDK Vamp plugin y pequeñas aplicaciones que sirven como herramientas para el desarrollo[29]. Además en el repositorio de Vamp plugins[30] encontramos instrucciones para implementar nuestro propio plugin. También ponen a nuestra disposición un tutorial que explica todo lo que debemos saber para hacer uso de las funciones de la SDK Vamp plugin. La implementación de los Vamp plugins se puede hacer en C++, Python o Java, pero el código del HPCP esta en C++, el desarrollo se ha hecho en el mismo lenguaje. Es importante tener en cuenta que un Vamp plugin es una librería dinámica que necesita de aplicaciones como Sonic Visualizer, Sonic Annotator o Audacity[31] para poder ejecutarlas. No funcionan como aplicación independiente. Además no aplica ningún cambio sobre el archivo de audio, sino que extrae características. En el tutorial de desarrollo de un Vamp plugin, vienen instrucciones muy claras (línea a línea) para compilar la librería y crear una nueva librería dinámica [25, 26]. Principalmente, la implementación del código la hemos realizado en Windows XP con Visual Studio, los retoques finales los he hecho en Mac OSX. En ambas plataformas se requiere el uso del terminal, puesto que la aplicación de testeo, vamp-plugin-tester, funciona con comandos de consola. Como no son aplicaciones que podamos ejecutar y no podemos estar comprobando cada cambio en Sonic Visualizer, la aplicación, vamp-simple-host, nos permite visualizar los resultados por consola y resulta muy útil durante el desarrollo de los plugins. Los dos Vamp plugins implementados se entregan en el anexo III (CD).

4.2.2.1 HPCP 3.0

La versión beta de la nuevo HPCP Vamp plugin incorpora el filtrado de coeficiente cepstrum mediante el relleno de ceros o la función gaussiana. Cada uno de esto procesos se conocen como cepstral liftering y cepstral filtering. Esta aplicación esta desarrollada en el dominio frecuencial.

El relleno de ceros se controla mediante el parámetro (cepsCoeff) que indica el numero de los coeficientes cepstrum más bajos que son sustituidos por ceros. Este parámetro por defecto es 0, es decir, el proceso esta desactivado por defecto y es el usuario quien debe definirlo, puesto que el filtro siempre dependerá del contenido del archivo de audio o de la aplicación para la que se quiera utilizar el vector de croma. El filtrado de coeficientes cepstrum mediante la función gaussiana, se controla mediante el parámetro alfa, que hemos definido en los experimentos anteriores. En un futuro se podría desarrollar la elección del cepstral zeros automáticamente, pero esto requiere de un mecanismo que evalúe cual es el mejor filtrado en cada fragmento de audio. Este proceso lo incorporamos en la cadena de procesos del HPCP antes de la de detección de picos. La implementación esta basada en el siguiente pseudocódigo:

[29] vamp -plugin tester y vamp-simple-host
[30] http://www.vamp-plugins.org/develop.html
[31] http://audacity.sourceforge.net/

Entrada: buffer
Salida: newSpec

Librerías: **fftsg_h.c**[32]

Parámetros:

- frameSize
- cepsCoeff

Funciones:

- computeSpectrum(): Esta función calcula la magnitud del espectro

Entradas: frameSize, buffer
Salidas: mX (lineal)

- computeCepstrum(): Esta función calcula la FFT del logaritmo de mX

Entradas: mX
Salidas: Cepstrum

- discardCepsCoeff(): Esta función elimina coeficientes cepstrum.

Entradas: Cepstrum, cepsCoeff
Salidas: newCepstrumVector

- computeNewSpectrum(): Esta función calcula la inversa del logaritmo de la inversa de la FFT.

Entradas: newCepstrumVector
Salidas: newSpec

Programar en C++ no permite extraer gráficos con tanta facilidad como en MATLAB, por eso hemos tenido que crearnos nuestra propias herramientas para testear esta aplicación. Para cerciorarnos de que el proceso se realizaba con éxito desarrollamos un Vamp plugin, que simplemente muestra el espectro con la posibilidad de configurar el filtrado de coeficientes cepstrum. En las siguientes figuras podemos su efecto como función de blanqueador espectral.

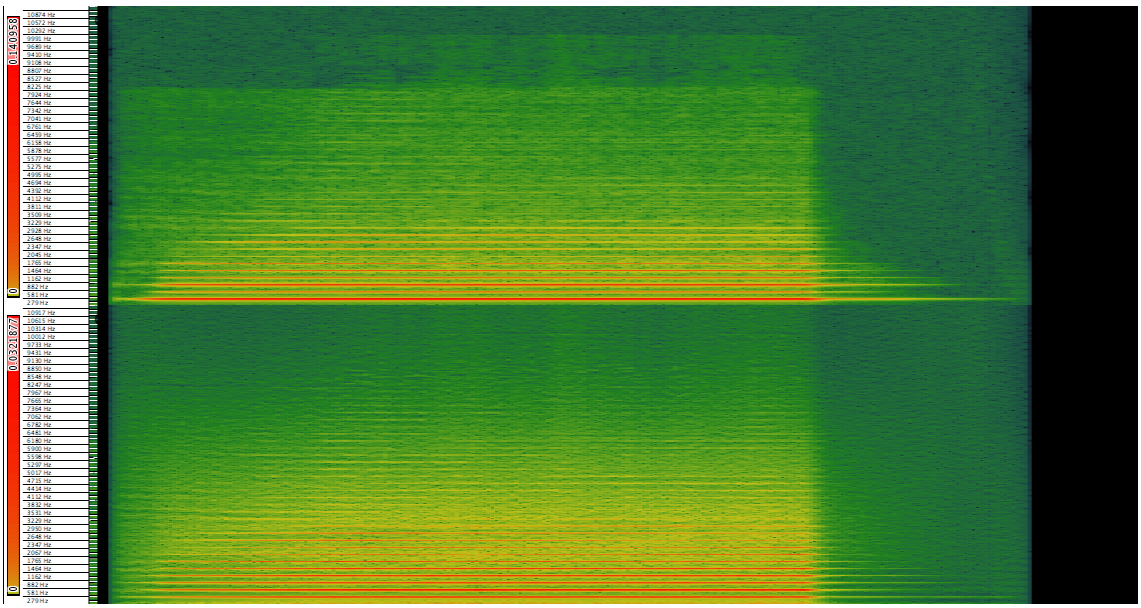


Figura 53. Comparación del espectro original y el resultado del relleno de ceros cepstral. En la parte superior, el espectro original de un sonido armónico y en la parte inferior su espectro después de aplicar le relleno de ceros de los primeros 50 coeficientes cepstrum.

[32] http://www.kurims.kyoto-u.ac.jp/~ooura/fft.html

Esta implementación es una versión beta que nos ha servido para visualizar los resultados del proceso. No se creó con la finalidad de publicarla sino como herramienta. A continuación, mostramos el panel de configuración del nuevo HPCP y los parámetros de control que hemos añadido.

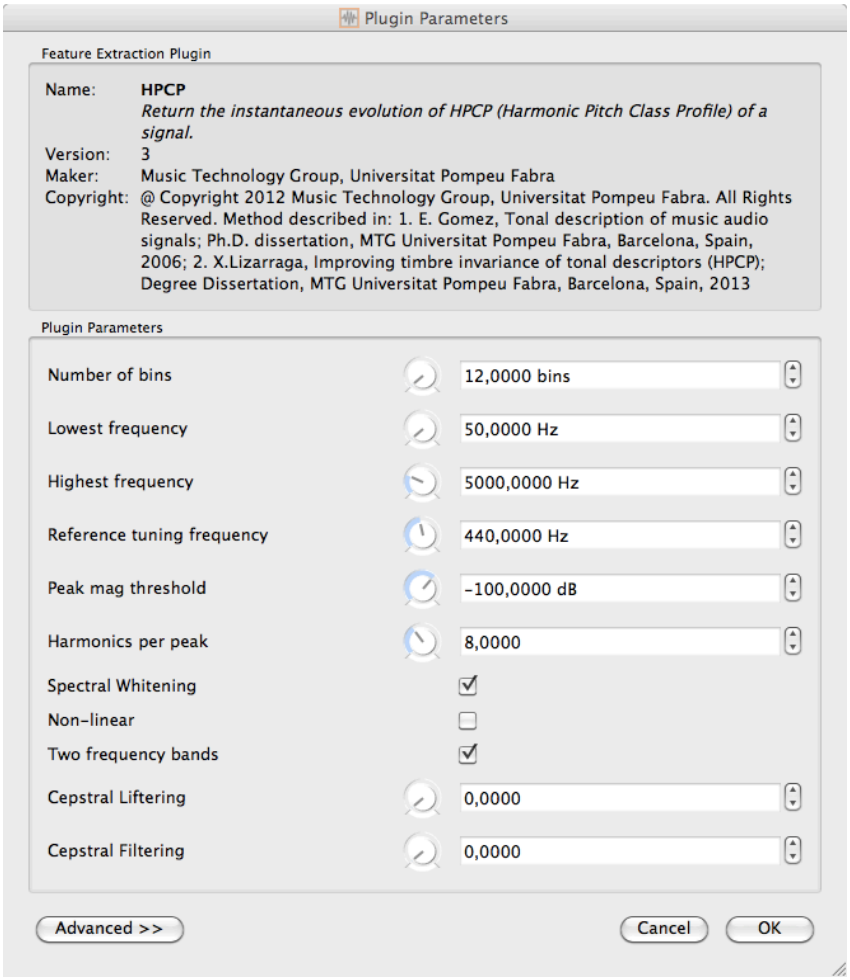


Figura 54. Configuración del HPCP 3.0 en Sonic Visualizer

Finalmente tras la implementación en el HPCP vemos que el proceso afecta al resultado del croma, dependiendo del valor de alfa o del coeficiente ceptrum que indiquemos.

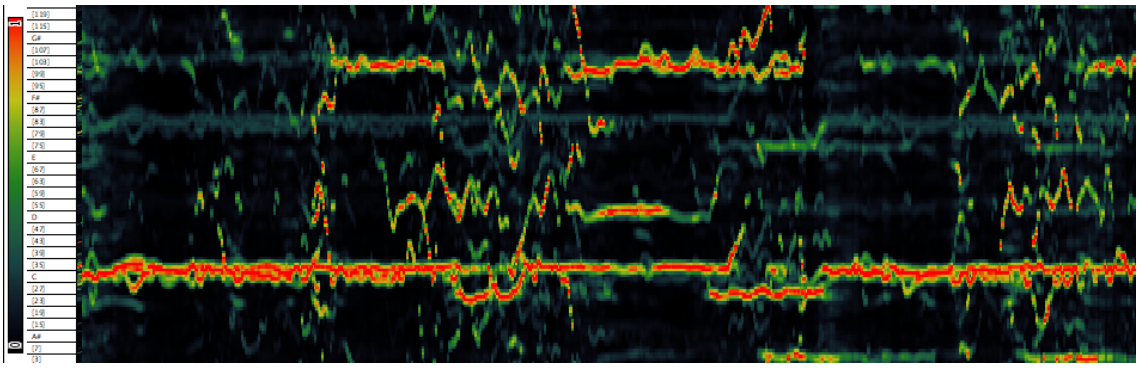


Figura 55. Resultados del croma (120 bins) de la primera melodía del Bolero de Ravel (Ravel, M. 1928) con el HPCP utilizando el blanqueador espectral por defecto y sin activar el filtrado cepstrum.

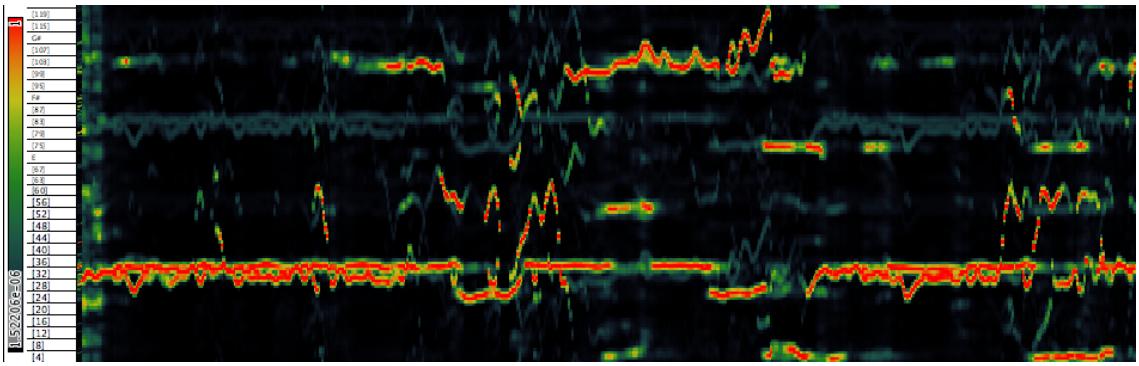


Figura 56. Mostramos el resultado del mismo fragmento del Bolero de Ravel, aplicando un relleno de ceros en los primeros 80 coeficientes cepstrum.

En estas dos figuras donde cada semitono esta dividido en decimas de semitono, observamos que el procesado sobre los coeficientes cepstrum consigue un resultado con muchas menos interferencias y consigue extraer un cromagrama más definido, aunque no define la melodía del mismo modo y podemos perder partes de ésta. El spectarl whitenning puede utilizarse para extraer melodías y el cepstral liftering para identificar perfiles o describir contenido polifónico. Observamos las diferencias añadiendo el procesado de cepstrum mediante la función gaussiana con alfa igual a 1.5

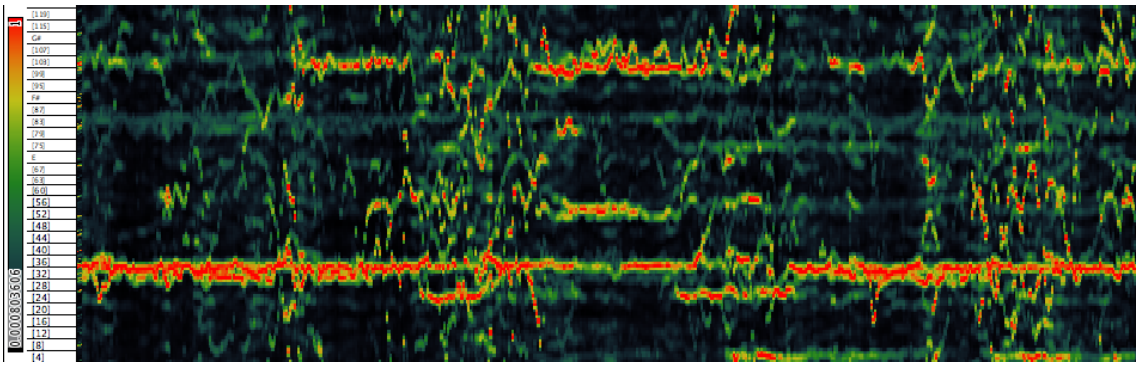


Figura 57. En esta caso aplicamos el filtrado de cepstrum mediante la función gaussiana (alfa = 1.5).

Nos damos cuenta que el filtro gaussiano sobre los coeficientes cepstrum no ofrece una visualización definida del contenido del croma. Introduce demasiado contenido musical que inicialmente no existe. Esto puede ser debido al aumento de la magnitud que provoca este proceso y que en algunos casos puede situar el ruido por encima del umbral de detección de picos espectrales, mostrado en el anexo II.

4.2.2.2 Espectrograma IF

Para desarrollar el espectrograma de las frecuencias instantáneas hemos tenido que desarrollar toda la estructura del análisis espectral, puesto que la SDK Vamp plugin no permite acceder a la ventana y esto no nos permite calcular la derivada de la función ventana. Debido a ello, este plugin lo desarrollamos en el domino temporal y para calcular la transformada reutilizamos la librería Ooura FFT, de licencia Open Source. El desarrollo de este plugin se ha centrado en el siguiente pseudocódigo.

Diseño del Vamp plugin IF - Spectrogram

Esta aplicación se desarrolla en dominio temporal.

Input: muestras de audio

Output: El espectrograma IF es un vector de tamaño frame size/2 en cada fragmento analizado. Contiene la magnitud de las frecuencias instantáneas detectadas.

Librerías: **fftsg_h.c**

Parámetros:

- frameSize size_t
- hopSize size_t
- windowType string o enum

Funciones:

- generateWindow(): Esta función genera la función ventana y su derivada, ambas normalizadas con la área de la ventana.

Entradas: frameSize, WindowType
Salidas: win, dwin

- windowedFrame(): Esta función multiplica cada fragmento por la ventana y su derivada.

Entradas: samples, win, dwin.
Salidas: xw, dxw

- zeroPhase(): Esta función aplica el cero fas en cada fragmento.

Entradas: frameSize, xw, dxw
Salidas: xwz, dxwz

- computeFFTandZPadding(): Calcula la FFT con el relleno de ceros.

Entradas: xwz, dxwz
Salidas: X, dX

- computeIF(): Calcula las frecuencias instantáneas

Entradas: X, dX
Salidas: IF

- computePeaks(): Detecta los picos de las frecuencias instantáneas del mapeado con la función de cruce de ceros y la similaridad. El espectrograma de IF relaciona estas frecuencias a su magnitud.

Entradas: IF, X, frameSize
Salidas: IF-Spectrogram

Como mostramos en el panel de configuración, hemos acabado añadiendo un parámetro más que permite indicar un umbral de magnitud para eliminar las frecuencias instantáneas con baja magnitud, que pueden despistar. Este parámetro se indica en decibelios (dB).

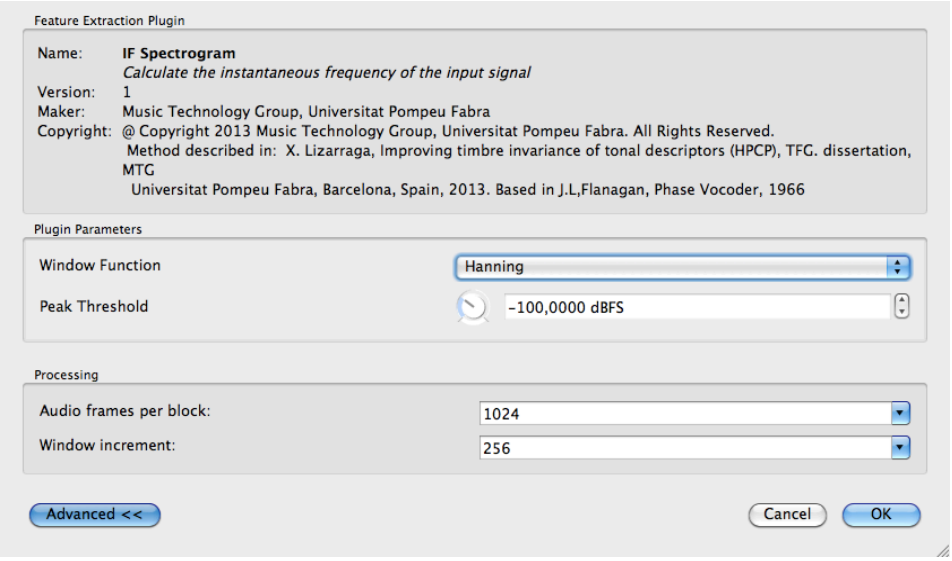


Figura 58. Configuración del Espectrograma IF en Sonic Visualizer.

En las siguientes figuras observamos el efecto que produce el umbral sobre el espectrograma IF y la dependencia de los parámetros espectrales de esta aplicación.

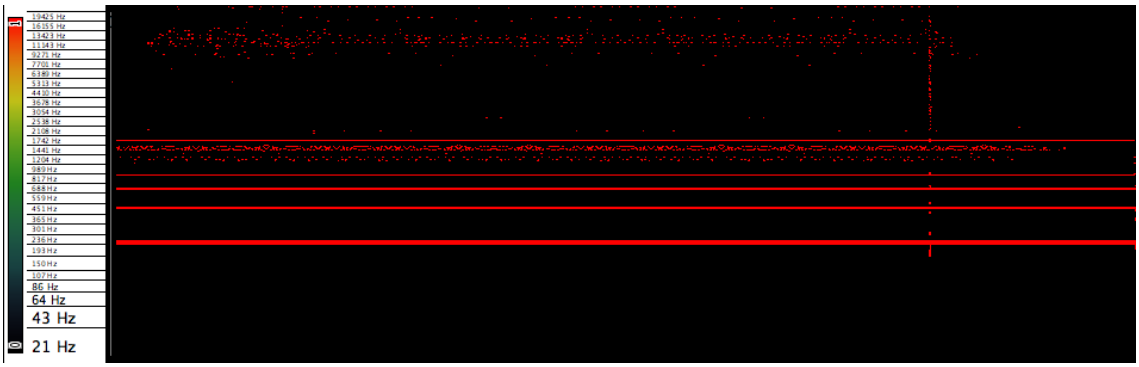


Figura 59. Muestra la detección de las frecuencias instantáneas de una señal compuesta por 5 armónicos con una frecuencia fundamental de 200Hz, aplicando los parámetros por defecto, una función ventana Hanning y un umbral de -100dB.

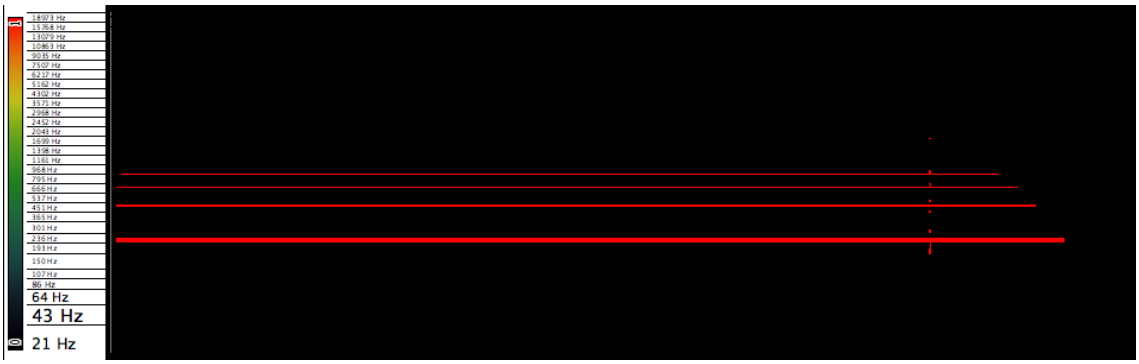


Figura 60. Mostramos como con los mismos parámetros pero con un umbral de -60 dB. Conseguimos una detección de la frecuencia instantánea mucho más limpia pero se pierde uno de los armónicos.

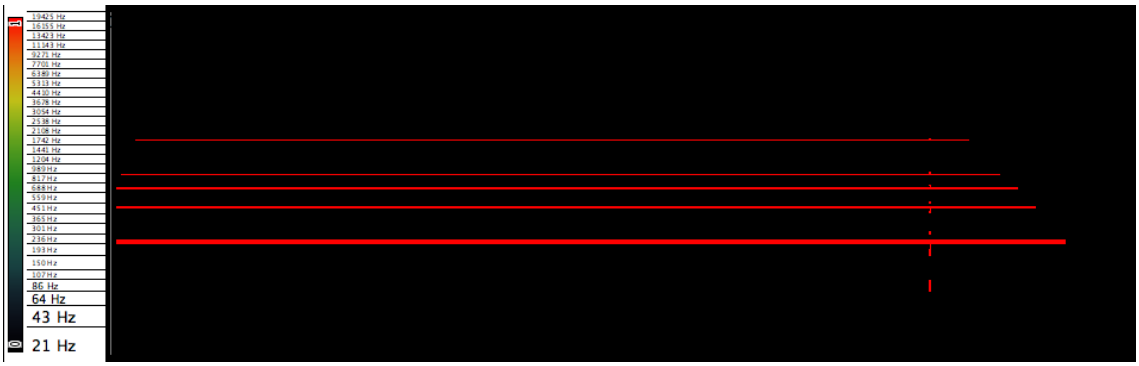


Figura 61. El espectrograma IF con los mismo parámetros y con ventana Barlet.

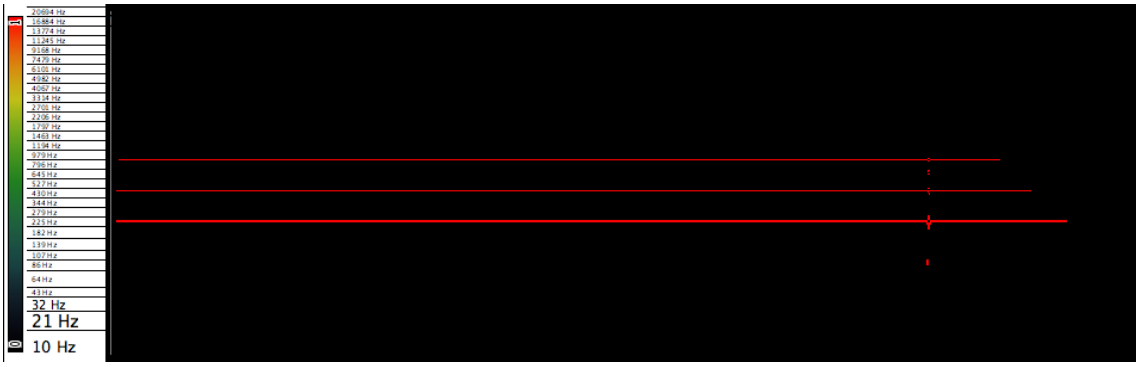


Figura 62. Espectrograma IF con tamaño de la FFT de 4096 y ventana Hanning.

En las siguientes figuras mostramos los resultados con un sonido polifónico, la figura 63 corresponde a un tamaño de la FFT de 1024 y 256 muestras de factor de salto.

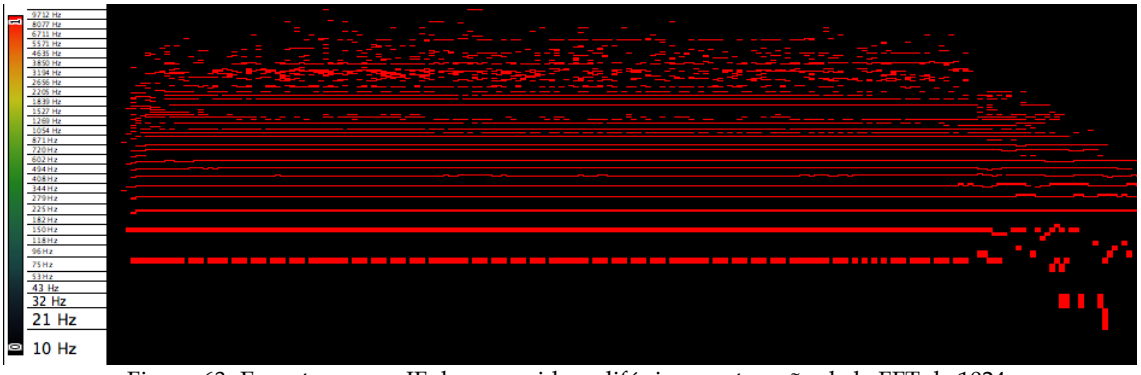


Figura 63. Espectrograma IF de un sonido polifónico con tamaño de la FFT de 1024.

La figura 64 muestra los resultados para una FFT de 4096 y 256 muestras de factor de salto, los mismos parámetros espectrales que se recomiendan para el HPCP. Comparándolo con la figura anterior podemos observar que se pierden algunos de los armónicos, puesto que solo detecta frecuencias instantáneas en el rango de 50 Hz a 1000 Hz. La dependencia a los parámetros espectrales introducen cambios en los resultados que podrían ser significativos, aunque por otro lado la detección de picos espectrales es muy exacta y su rango podría concentrarse en los primeros armónicos que contienen la información de tono de más peso. Por ello no descartamos que en un futuro se añada al HPCP. Esta aplicación puede servir a la comunidad para visualizar las propiedades que ofrecen cada tipo de ventana en la estimación de las frecuencias instantáneas.

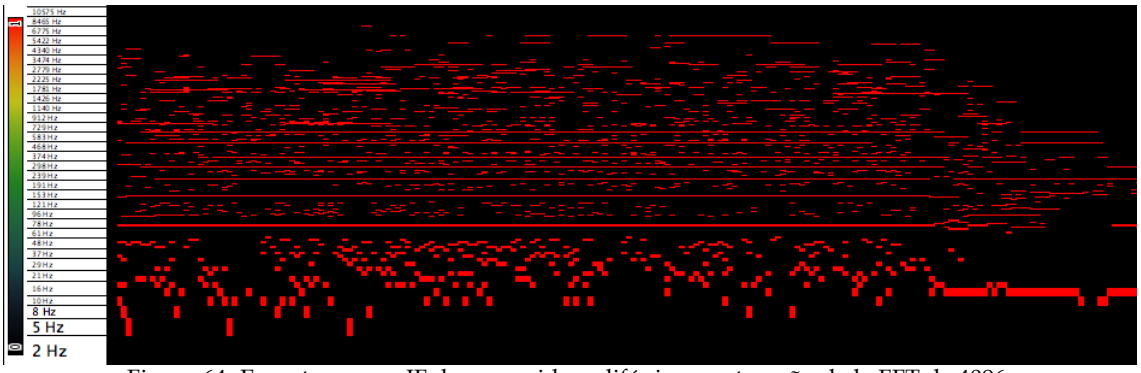


Figura 64. Espectrograma IF de un sonido polifónico con tamaño de la FFT de 4096.

4.3 Evaluación de la nueva versión del HPCP

Para evaluar el nuevo HPCP aplicamos la evaluación de los descriptores tonales que hemos definido en el capitulo anterior. En este caso, evaluamos los resultados del HPCP con diferentes valores de los parámetros de los métodos introducidos: cepstral liftering y cepstral filtering. En el caso del cepstral liftering hemos evaluado los resultados del HPCP para 12 índices de coeficiente cepstrum que indican hasta que coeficiente aplicamos el relleno de ceros: 25, 50, 75, 100, 125, 150, 175, 200, 225…, 300.

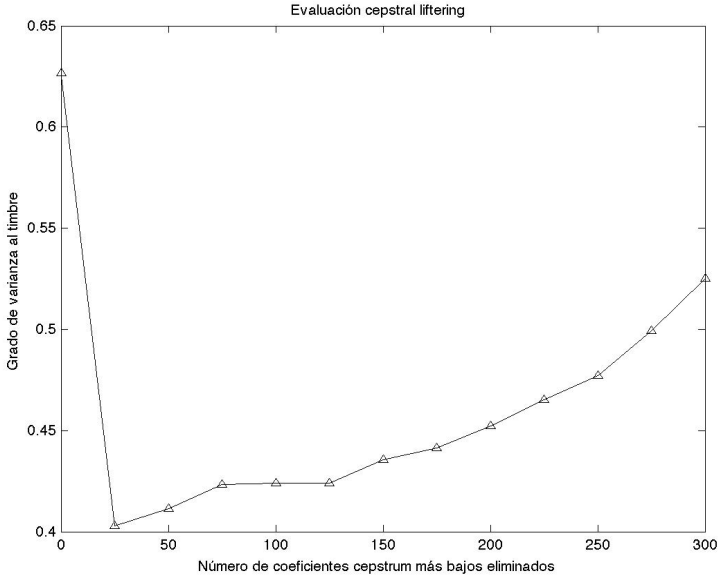


Figura 65. Varianza al timbre para el cepstral liftering

En la siguiente tabla comparamos los resultados obtenidos del HPCP original y el de la nueva versión de HPCP aplicando cepstral liftering eliminando los 25 primeros coeficientes:

HPCP	0.3456	0.0528	0.5516	0.0837	0.6265
HPCP3	0.1691	0.0125	0.4197	0.0262	0.4029

Tabla 6. Resultados de la estimación del grado de varianza al timbre HPCP 3.0

Como muestra la figura 65 a medida que eliminamos coeficientes la invarianza al timbre empeora debido a que cuantos más coeficientes eliminamos más plano es el espectro pero eso tiene un coste y es que esta igualando la contribución de los picos locales con los picos relacionados al ruido de fondo o artefactos del análisis espectral que producen diferencias en μ_l y reducen la invarianza al timbre. Veamos los resultados del grado de eficiencia.

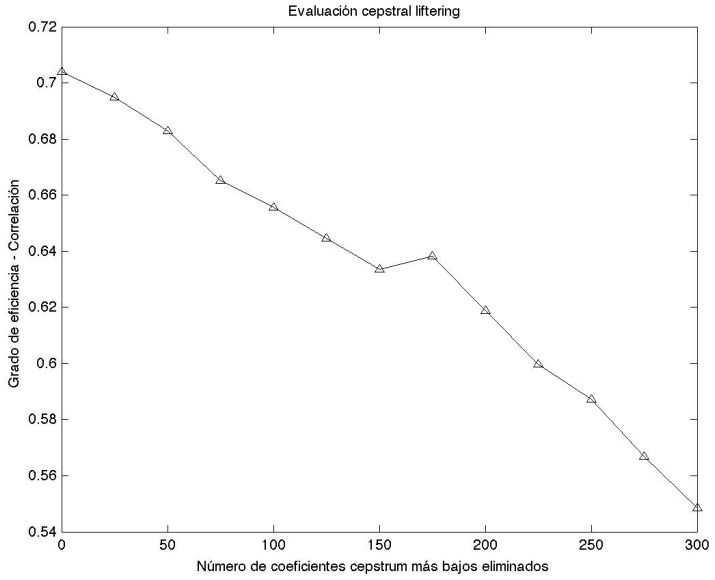


Figura 66. Grado de eficiencia estimado para el cepstral liftering (correlación)

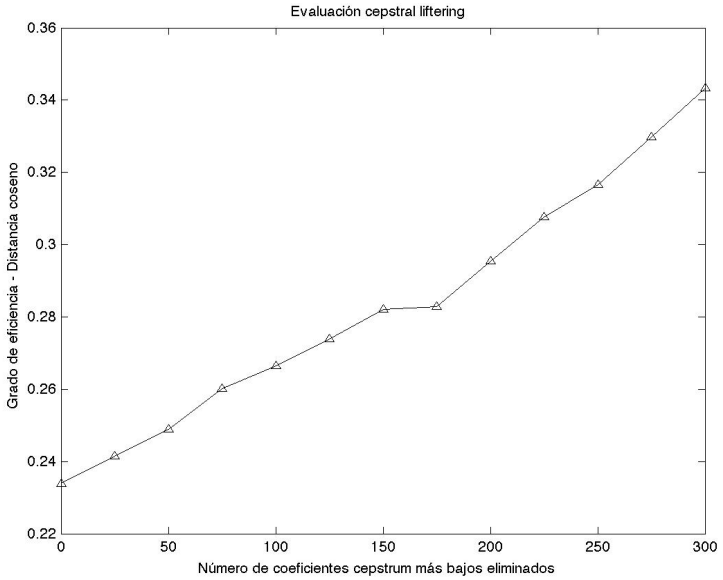


Figura 67. Grado de eficiencia para el cepstral liftering (distancia coseno)

Los resultados de la evaluación aplicando cepstral liftering en el HPCP, muestran que la variación al timbre mejora, puesto que se reduce el valor μ_l y con ello las distancias de las distintas instancias de una clase acorde. Esto quiere decir que a pesar de las variaciones tímbricas de cada instrumento, con este proceso podemos conseguimos que se asemejen.

Ante estos resultados podemos afirmar que hemos conseguido el objetivo que nos proponíamos para nuestro trabajo. Sin embargo, como vemos en las dos imágenes siguientes el grado de eficiencia disminuye drásticamente, lo que ganamos en la variación al timbre lo perdemos en eficiencia para la estimación de acordes. Por lo tanto este proceso puede ser útil según la aplicación que le demos al descriptor tonal HPCP. Por otro lado, el filtrado cepstrum mediante la función gaussiana se define por el parámetro alfa (del 0 a 5). La evaluación de este método sobre el resultado del HPCP la mostramos en las siguientes figuras:

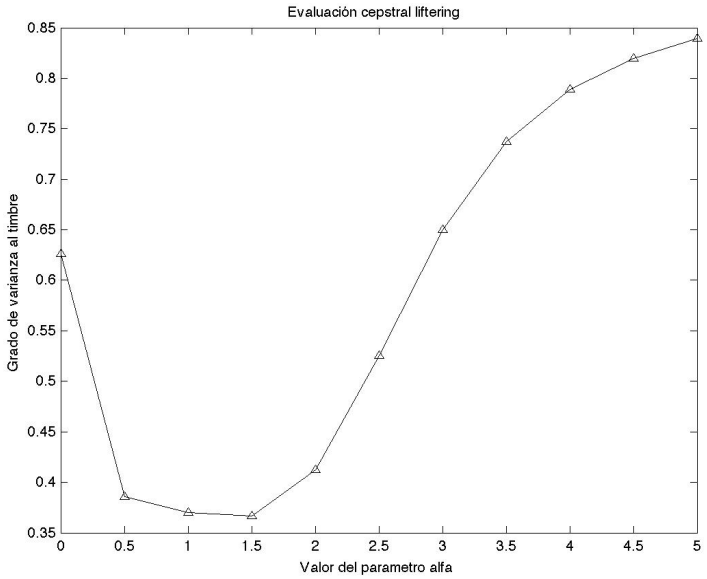


Figura 68. Varianza al timbre para el cepstral filtering

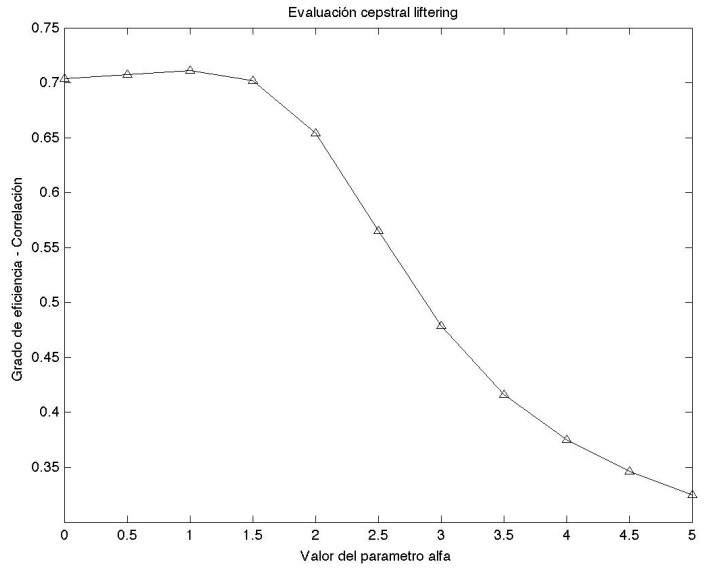


Figura 69. Grado de eficiencia estimado para el cepstral filtering (correlación)

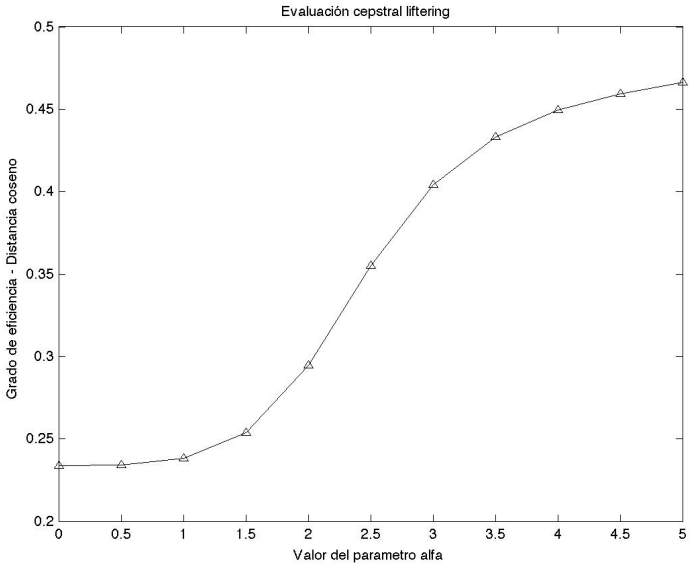


Figura 70. Grado de eficiencia estimado para el cepstral filtering (distancia coseno)

Vemos como el resultado del cepstral filtering difiere del cepstral liftering. Un detalle bastante importante es que existe un rango donde el parámetro alfa define un filtro que consigue mejorar la varianza al timbre sin afectar al grado de eficiencia. Este rango de alfa va de 0.25 a 1.5. Con una alfa de 1.5 conseguimos la varianza al timbre mínima pero el grado de eficiencia comienza a disminuir.

En el caso del Vamp plugin IF Spectrogram, no disponemos de ninguna herramienta de evaluación. Por ello simplemente compararemos con los resultados que obtuvimos en la versión implementada en MATLAB. Sabemos que no es el modo más correcto de evaluarlo pero no disponemos de ninguna herramienta para ello. Mientras escribía esta memoria me he dado cuenta que una posible evaluación para el IF-Spectrogram sería sintetizar una señal y crear una mascara binaria con los picos espectrales que sabemos que contiene o que debería detectar el algoritmo y calcular la distancia del coseno. Puede ser una herramienta a desarrollar en un futuro.

4.4 Usuarios de prueba

Usuario 1:

El usuario 1, es un compañero de clase, estudiante de ingeniería en sistemas audiovisuales, un usuario con suficiente experiencia. Ha intentado testear los dos Vamp plugin con Windows 8 pero se ha encontrado con errores que genera el mismo Sonic Visualizer con la carga de cualquier Vamp plugin. No ha podido aportar información sobre los resultados pero nos informa de estos errores que lo añadiremos al archivo readme que acompaña a cada una de las aplicaciones.

Usuario 2:

El usuario 2 al igual que el anterior usuario, es un estudiante de ingeniería en sistemas audiovisuales. Comenta que los resultados le parecen correctos y muy acertados aunque no tenia muy claro qué diferencias debía observar. Por otro lado, nos ha sugerido añadir dos botones que indiquen cuando cualquier de los dos filtros cepstrum están activados.

Usuario 3:

El usuario 3 tiene una gran experiencia en el campo del audio y de los descriptores. Le convence el resultado, pero me ha comentado algo similar al usuario anterior. Sugiere que añadamos un botón de activación para que quede más claro. También me ha sugerido que cambie algún parámetro por defecto como el factor de solapamiento o el numero de contenedores de croma.

5. Conclusiones

En la evaluación inicial de los descriptores tonales hemos podido comprobar que el CRP es muy robusto a los cambios de timbre aunque no destaca por su grado de eficiencia en la detección. El HPCP ha mostrado un comportamiento totalmente contrario, baja invarianza al timbre y un grado bastante alto de eficiencia (0.28 con la distancia coseno). El resultado del análisis de clases acordes (pag. 51-54) muestra una dependencia de la consonancia de los intervalos musicales que contiene la señal de audio. En este sentido es importante comentar que seguimos en contacto con el autor y estamos discutiendo los resultados.

Por otro lado, como hemos visto en el capitulo anterior el cepstral liftering y el cepstral filtering nos ofrecen nuevas posibilidades en la extracción de croma. No es fácil concretar cual de los dos consigue un mejor resultado pero por los resultados obtenidos en la última evaluación podemos recomendar utilizar cepstral filtering con un alfa entre 0.25 y 1.5, puesto que mejora la invarianza al timbre pero mantiene el grado de eficiencia. Esto no quiere decir que el cepstral liftering consiga malos resultados. Como hemos repetido a lo largo del documento, siempre dependerá de la aplicación que queramos darle y del contenido espectral. Otra conclusión interesante relacionada con esta técnica, es que sus resultados cambian cuando se trata de un sonido monofónico o polifónico. Un sonido polifónico permite la eliminación de muchos más coeficientes cepstrum sin introducir distorsiones, puesto que el espectro contiene muchos mas picos locales y necesitamos eliminar más coeficientes para alcanzar un espectro plano. Sin embargo en un sonido monofónico hay partes del espectro donde no hay picos locales, sin embargo si eliminamos demasiados coeficientes podemos llegar a amplificar los artefactos del análisis espectral. Por eso eliminando menos coeficientes conseguimos un resultado similar (Anexo I).

Otra conclusión o recomendación a futuras investigaciones en el mismo campo es que el procesado cepstral se puede mejorar añadiendo un proceso inteligente que minimice la distancia entre los picos espectrales en cada uno de los fragmentos procesados, analizando todos los filtrados cepstrum, con la finalidad de encontrar el mejor procesado para cada uno de los fragmentos. Se trataría de clasificar cada uno de los diferentes filtros en cada fragmento para encontrar el que más se aproxima a un espectro plano.

En cuanto al espectrograma por frecuencias instantáneas, hay que mencionar que su implementación como Vamp plugin no formaba parte del objetivo inicial, puesto que la idea era utilizarlo en el HPCP. La detección de picos espectrales utilizando las frecuencias instantáneas es más efectiva que la detección parabólica, pero su dependencia a los parámetros espectrales ha complicado su adaptación al HPCP. El HPCP obtiene sus mejores resultados con un tamaño de FFT de 4096 muestras y con este tamaño de bloque, como hemos visto anteriormente, la detección de las frecuencias instantáneas solo es eficaz en un rango de 50 Hz a 1000 Hz.

Lo que por un lado podría distorsionar los resultados del vector croma resultante, por la falta de la contribución de los armónicos superiores, pero que por otro lado, eliminar la contribución de estos armónicos, en combinación con el cepstral liftering, puede ayudar a ser más robusto a la varianza de timbre. Puesto que a partir del tercer armónico de una nota, el mapeado de tono de éstos corresponden a otras notas y si los descartamos evitaremos su contribución y obtendremos un perfil acorde más definido. Por ejemplo, los armónicos que componen la nota C se mapean a las notas de su acorde mayor C-E-G, normalmente a partir del tercer armónico encontramos los armónicos que contribuyen a otras notas distintas al perfil autentico. Si los descartamos y aplicamos la función de blanqueado basada en cepstral liftering es posible que se obtengan cromagramas mucho más robustos a la variación de instrumentos y de octava. Puede ser un tema a tratar en futuros trabajos y en el que me encantaría participar. El caso es que durante este trabajo hemos considerado que su dependencia a los parámetros espectrales han mostrado que no es eficiente con los parámetros espectrales del HPCP y por ello desestimamos añadir el proceso a la cadena de procesado. Sin embargo, contribuimos con la comunidad mediante una aplicación independiente para el análisis de picos espectrales basado en las frecuencias instantáneas (IF-Spectrogram Vamp plugin) para que puedan analizar señales de audio y su contenido mediante este método.

Bibliografía

[1] Gómez, E. (2006). *Tonal Description of Music Audio Signals,* PhD thesis, Universitat Pompeu Fabra, Barcelona.

[2] Postigo, J., & Valldares, E. (2012). *Comercio Electrónico en el Reino Unido* in *Apartado II-1.3 Contenidos Digitales,* page 11. Oficina económica y comercial de la embajada española en el Reino Unido, Londres.

[3] PwC. (2012).*Global enterteinment and Media Outlook: 2012- 2016.* PWC. Retrieved from: http://www.pwc.es/es/publicaciones/entretenimiento-y-medios/assets/global-entertainment-and-media-outlook-2012-2016.pdf

[4] Grosche, P., Mueller, M., & Serrà, J. (2012). *Mutimodal Music Process* at Chapter 9 titled*: Audio Content-Based Music Retrieval, page 158-180*. Leibniz Zentrum für Informatik GmbH, Leibniz.

[5] Cano,P,. Batlle, E., Kalker, T., & Haitsma.J. (2002). *A Review of Algortihms for Audio Fingerprints, page 2*. Universitat Pompeu Fabra and Philips Research Eindhoven, Barcelona- Eindhoven.

[6] Wang, A. (2000). *An Industrial-Strength Audio Search Algorithm*. Shazam Entertainment Ltd., California.

[7] Bello, J., Pickens, J. (2006). *A Robust Mid-level Representation for Harmonic Content in Music Signals*. Queen Mary, University of London, London.

[8] Gouyon, F. Herrera, P. Gómez, E. Cano, P. Bonada, J. Loscos, À. Amatriain, X. Serra, X. *Content processing of music audio signals*, page 88 in *Sound to sense, sense to sound:A state- of-the-art in Sound and Music Computing.* Polotti P. and Roccheso D. (eds). Logos Verlag, Berlin GmbH

[9] Peeters, G. (2004). *A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project*. IRCAM, Paris.

[10] Zölzer, U., Amatrain, X., Arfib, D., Bonada, J., & Serra, X. (1998). *DAFX-Digital Audio Effects,* page 383. John Wiley & Sons.

[11] Ricardo. G. Duran Mesz, B. (2010). *¿Por qué usamos 12 notas? De Pitágoras a Bach.*

[12] Gómez, E. (2006). *Tonal description of music audio signals, page 66.* PhD thesis. Universitat Pompeu Fabra, Barcelona.

[13] Gómez,E. (2006). *Tonal description of music audio signals.* PhD thesis. Universitat Pompeu Fabra, Barcelona.

[14] Müller, M. Ewert, S. (2010). *Towards timbre-invariant audio features for harmony-based music.* IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 3, pp. 651

[15] Rabiner, R. L., & Schaffer, R. (2011). *Theory and applications of Digital Speech Processing. Pearson Education, Ltd.*

[16] Nanzhu J., Peter G., Verena K., and Meinard M. (2011). *Analyzing chroma features types for automated chord recognition*. Saarland University and MPI Informatik, Campus E1.4, 66123 Saarbru¨cken, Germany.

[17] Serrà, J. (2011). *Identification of versions of the same musical composition by processing audio descriptions*. PhD thesis, Universitat Pompeu Fabra, Barcelona.

[18] Z., Guojun, L., Ting, K. M., Zhang, D. (2011). *A survey of audio-based music classification and annotation,* IEEE Transactions on Multimedia, 13(2), 2011.

[19] Fujishima, T. (1999). *Real-time chord recognition of musical sound: a system using Common Lisp Music.* Proceedings of International Computer Music Conference (ICMC99), Beijing.

[20] Müller, M. Ewert, S. (2010). *Towards timbre-invariant audio features for harmony-based music. IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 3, page. 653*

[21] Flanagan, J. Golden, R. (1966). *Phase Vocoder.* The Bell System Technical Journal November 1966, Manuscript received July 18, 1966

[22] Nakatani,T. Irino, T. (2004). *Robust and accurate fundamental frequency estimation based on dominant harmonic components.* Wakayama University/NTT Communication Science Laboratories, Kyoto.

[23] Abe, T., Kobayashi, T., Imai, S. (1997). *The IF spectrogram: a new spectral representation.* Precision and Intelligence Labs, Tokyo Institute of Technology, Yokohama.

[24] Amatriain, X. (2004). *An Object-Oriented Metamodel for Digital Signal Processing with a focus on Audio and Music*. Universitat Pompeu Fabra, Barcelona. Retrieved from xavier.amatriain.net/Thesis/

[25] Cannam, C., Landone, C., Sandler, M., & Bello J. (2006). *The sonic visualiser: A visualisation platform for semantic descriptors from musical signals* in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Victoria, BC, Canada.

[26] Cannam, C. (2010). *The Vamp Audio Analysis Plugin API: A Programmer's Guide.* Revision 1.2, covering the Vamp plugin SDK version 2.0. Centre for Digital Music, Queen Mary, University of London.

[27] Herrera, P. (2006). *Automatic classification of percussion sounds: from acoustic features to semantic de- scriptions*. Doctoral dissertation, in preparation, Universitat Pompeu Fabra, Barcelona.

[28] Katz, B. (2002). *Mastering Audio, The Art and the Science*. Focal Press, Inc.

[29] Knopke, I., & Cannam, C. (2008). *Sonic Annotator Tutorial*. Centre for Digital Music, Queen Mary, University of London.

[30] McClellan, J. H., Schafer, R. W. , & Yoder M.A. (1998). *Signal Processing First: a multimedia approach*. NJ:Prentice-Hall, Inc.

[31] Montuliu, R. (2003). *Herramientas básicas, Tonalidad, Melodía*. Edición 2003-2004. Escola de Música de Badalona, Badalona.

[32] Livio, M. (2006). *La proporción áurea*. Ediciones Planeta, S.A., Barcelona.

[33] Rabiner, L. R., & Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice-Hall.

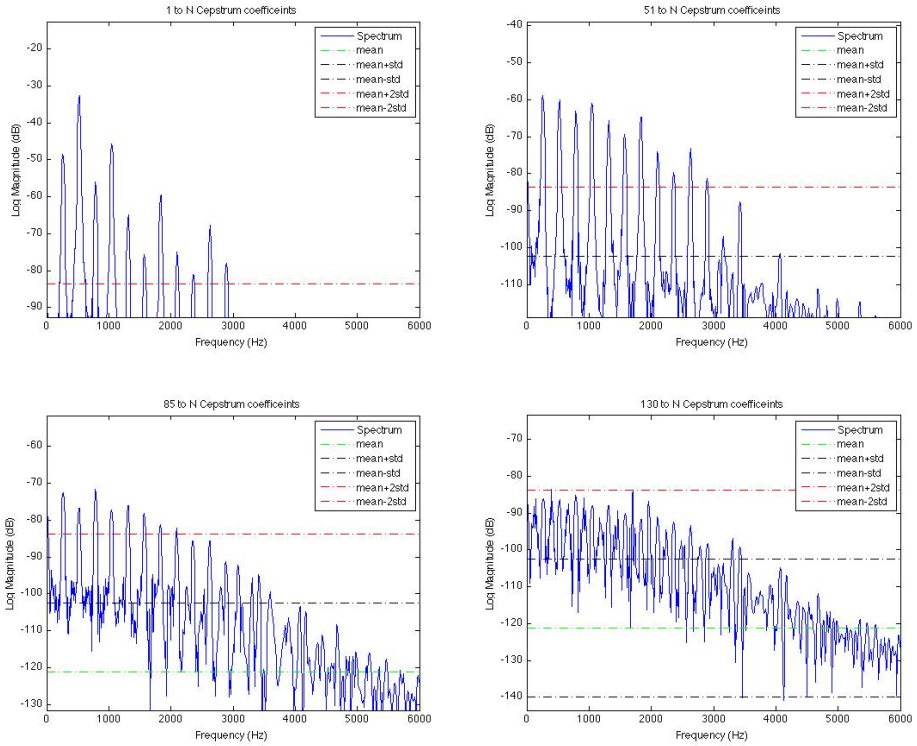
[34] Serrà, J., & Gomèz, E. (2008). *Audio cover song identification based in tonal sequence alingment*, Universitat Pompeu Fabra, Barcelona. Retrieved from files/publications/jserra_ICASSP08.pdf

[35] Watkinson, J. (1988). *The Art of Digital Audio*. 2nd Edition. Focal Press

ANEXO

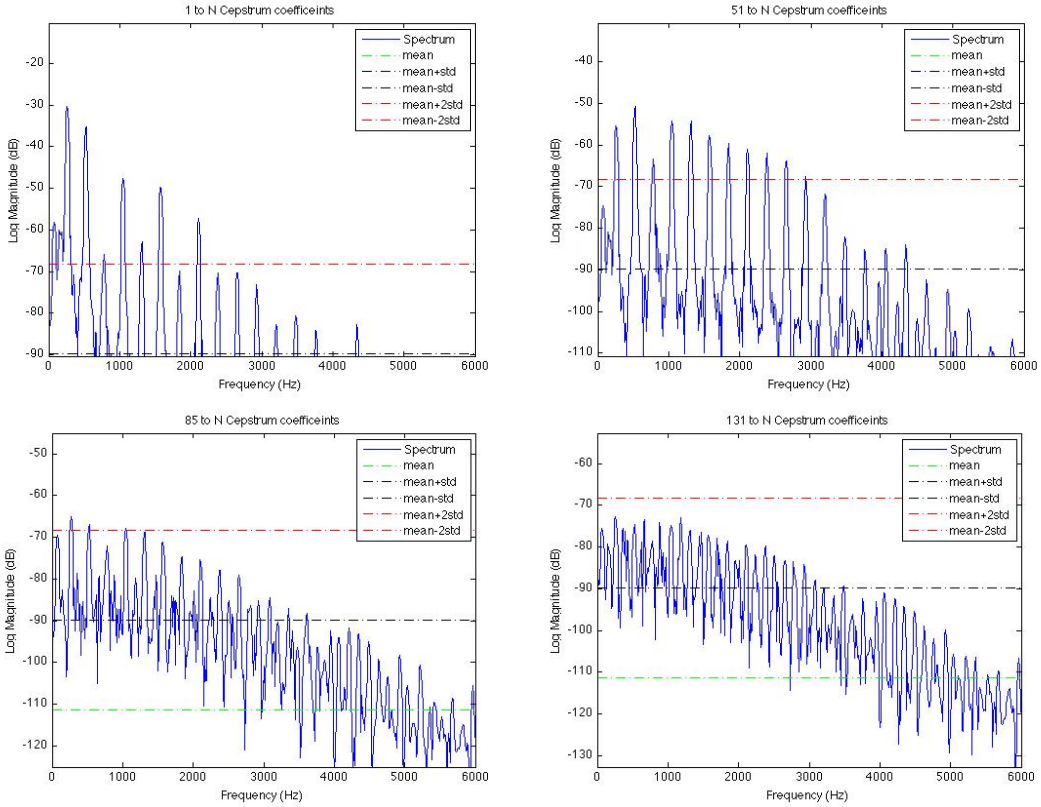
I. Experimentos con el relleno de ceros y filtrado de los coeficientes cepstrum

A lo largo de estos experimentos hemos utilizado la función STFT con un tamaño de la FFT de 2048 muestras, un factor de solapamiento de 512 muestras y una ventana Hanning de 1024 muestras. En los siguientes figuras mostramos el efecto del relleno de ceros de los coeficientes cepstrum sobre un sonido de guitarra (nota C_3) muestreado a 22050Hz.

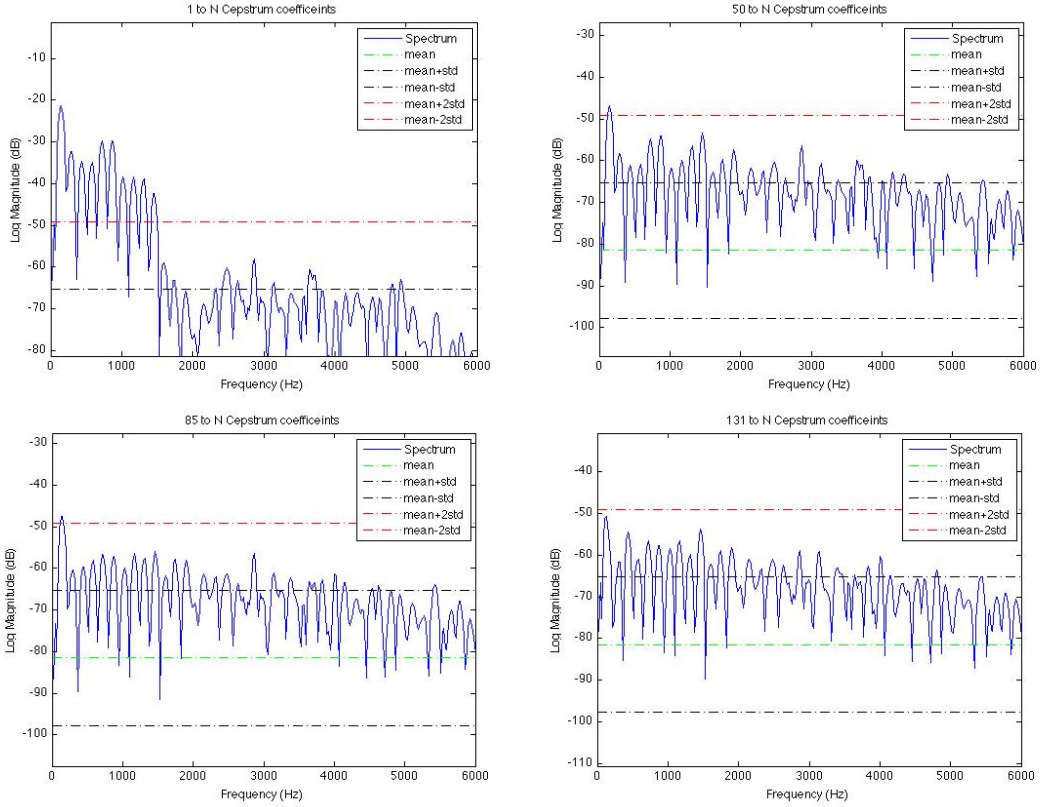


Arriba a la izquierda tenemos el espectro original del sonido de guitarra. La figura de arriba a la derecha, corresponde al espectro tras eliminar los primeros 50 coeficientes cepstrum. Su equivalente en frecuencia corresponde a 22050/50 = 441Hz. En este caso vemos su efecto de blanqueador espectral y la ecualización aplicada a los picos espectrales. Abajo a la izquierda, se han eliminado 85 coeficientes (259.4Hz) y vemos que su efecto reduce considerablemente la magnitud de los picos y puede introducir errores en la detección de picos. Arriba a le derecha vemos como el efecto del relleno de ceros es excesivo y elimina total mente el contenido inicial.

A continuación haremos el mismo experimento con un sonido de piano con la misma nota y repitiendo en el análisis espectral los mismos parámetros.



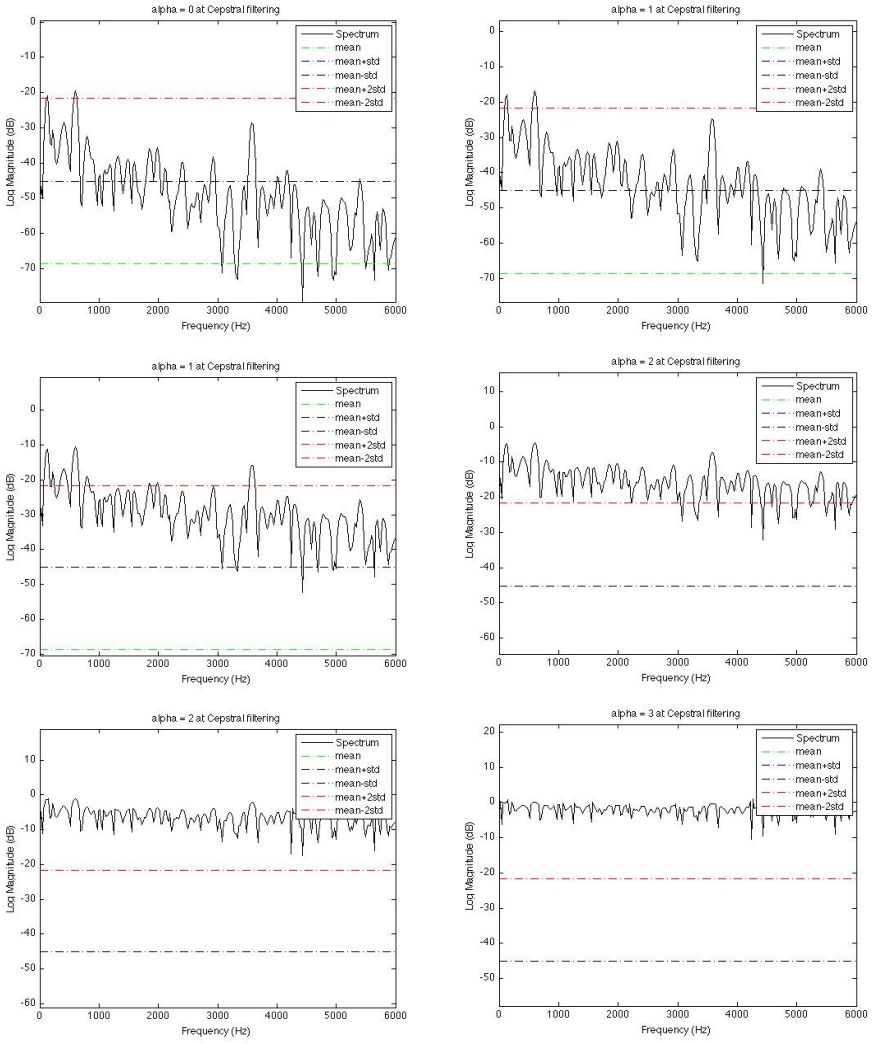
Vemos un efecto muy similar y que el mejor resultado lo obtenemos eliminando 50 coeficientes. Si comparamos con las figuras anteriores (guitarra), aunque mantienen diferencias, observamos como tanto con la guitarra como con el piano, el proceso ecualiza la aportación de sus armónicos. Veamos el efecto sobre un sonido polifónico.



Vemos que aunque el espectro es más irregular conseguimos un efecto similar al de los casos anteriores, aunque en el sonido polifónico eliminando 130 coeficientes conseguimos un blanqueado que no introduce artefactos y que permite apretar más el filtro eliminando un mayor número de coeficientes.

Por tanto, podemos afirmar que la sustitución de ceros de los primeros coeficientes cepstrum consigue ecualizar los picos espectrales y dependiendo de la señal que analicemos permite eliminar a mayor o menor número de coeficientes. Para una señal armónica monofónica, diremos que más de 75 coeficientes puede introducir artefactos, traduciéndose en un mala estimación del perfil de pitch. En cambio en señales polifónicas diremos que el margen de filtrado es mucho más amplio y podemos llegar hasta los 200 coeficientes sin introducir distorsiones. Definir estos márgenes nos ayudan a definir el rango de los parámetros de control del filtro en la nueva implementación del HPCP.

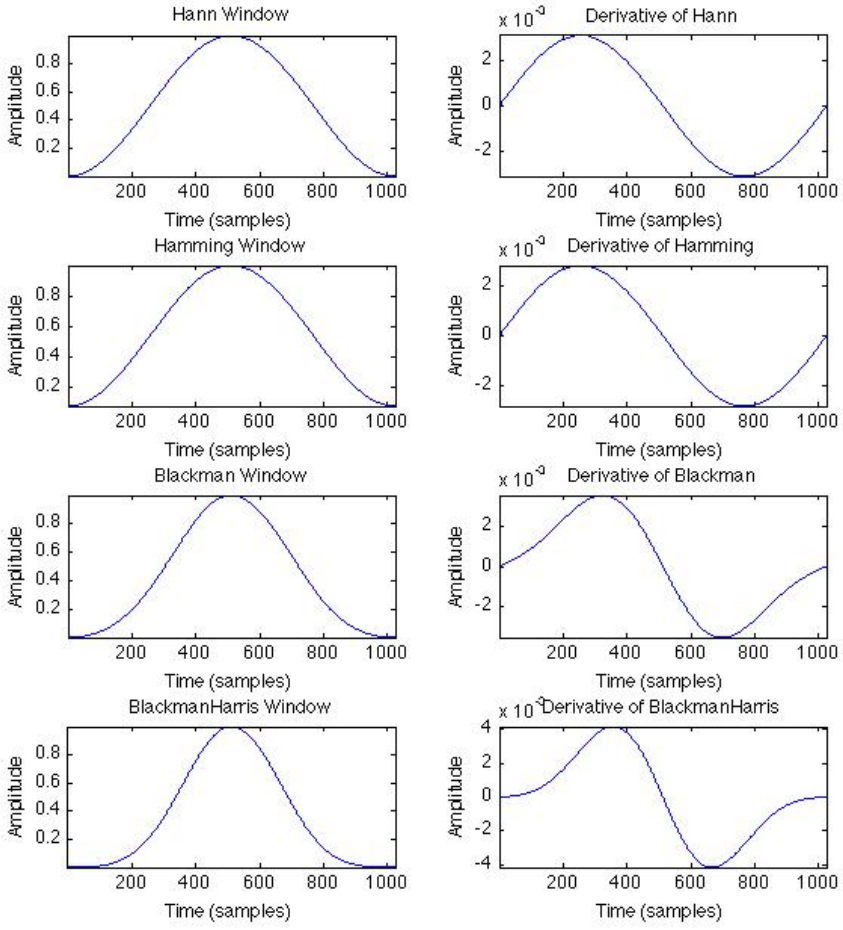
Si hacemos el mismo experimento para el filtrado de coeficientes mediante la función gaussiana podemos definir los rangos del parámetro alfa que define el filtro y analizar sus efectos.

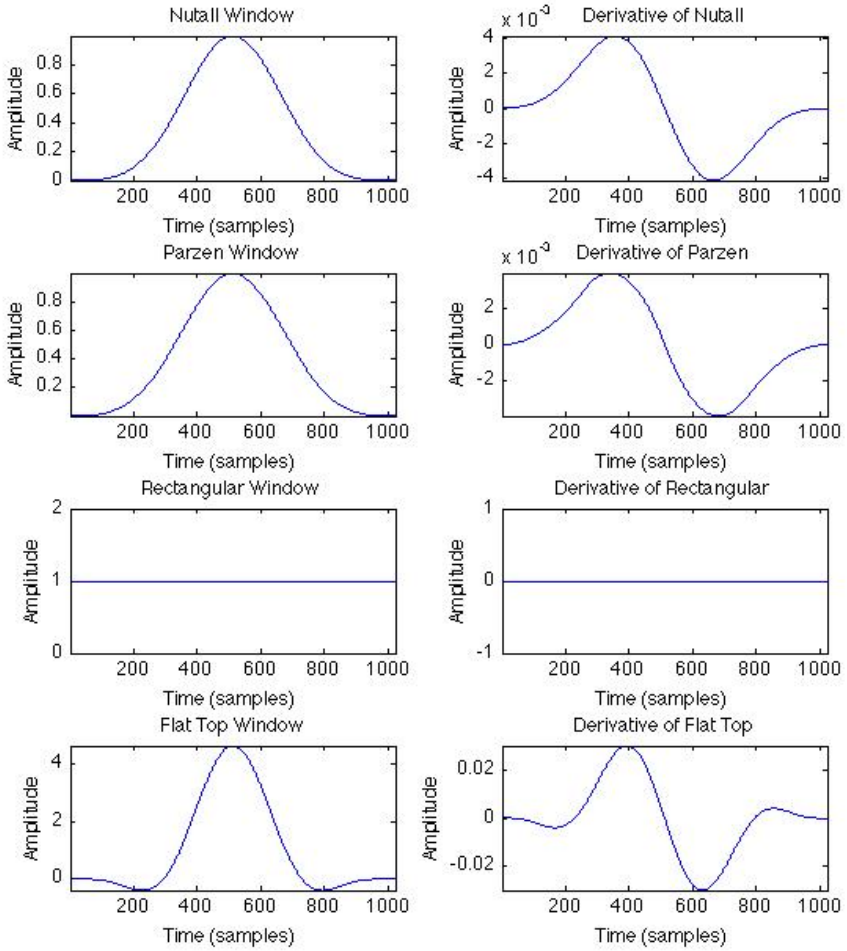


En este caso, vemos el efecto del filtrado mediante la función gaussiana, definida por el parámetro alfa igual a: 0, 0.5, 1.0 , 1.5, 2.0, 2.5 y 3.0. Ante estos resultados, podemos afirmar que el filtrado de cepstrum con un parámetro alfa mayor que 2.5 puede introducir errores en la detección de picos espectrales. Por eso hemos tenido en cuenta este rango en el desarrollo del proceso en el descriptor tonal.

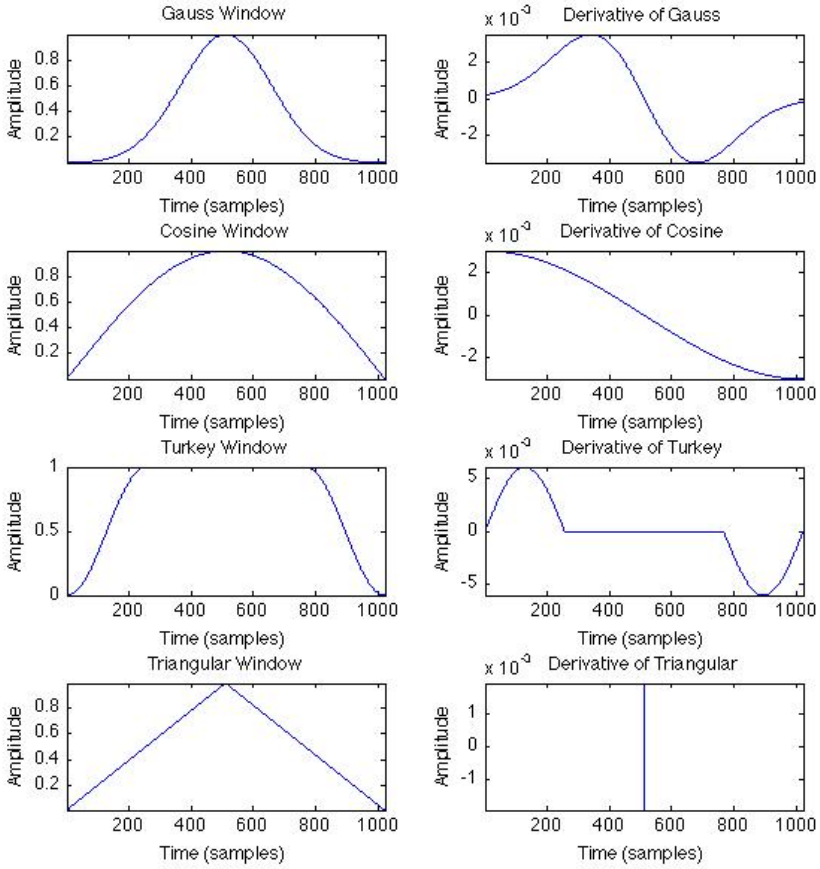
II. Experimentos para el cálculo de las frecuencias instantáneas

A continuación mostramos las funciones ventanas y sus derivadas que hemos definido para este cálculo:



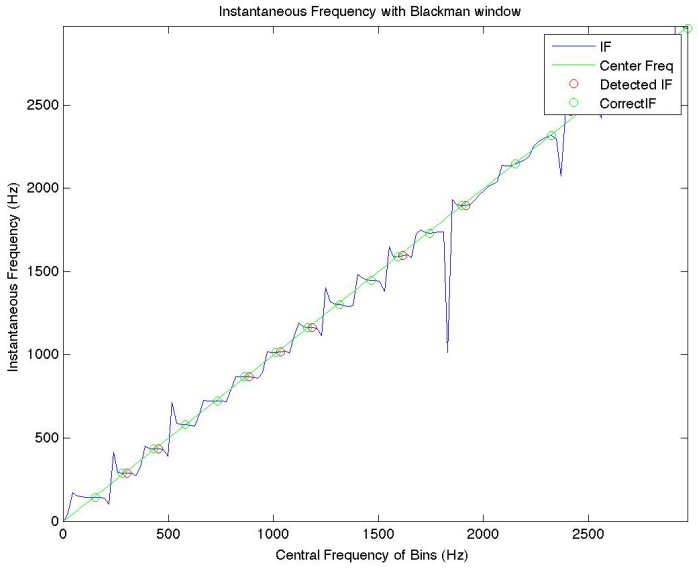


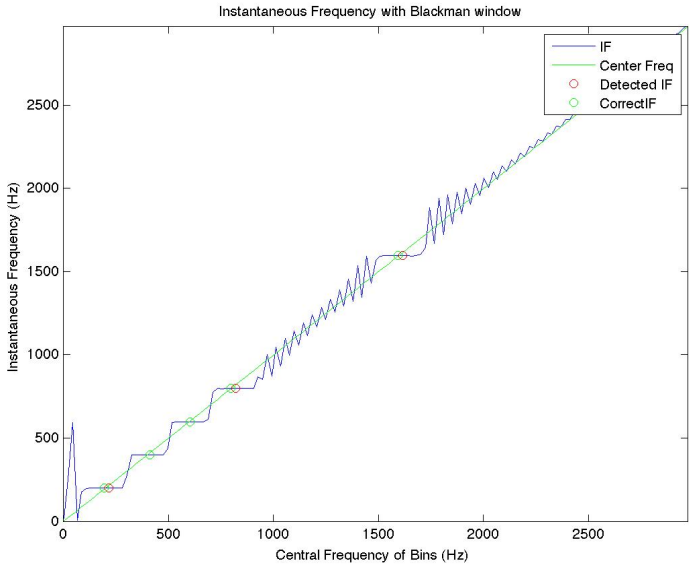
Cabe mencionar que las derivadas de las ventanas se han calculando derivando la función ventana pertinente, sin necesidad de calcular la diferencia entre muestras adyacentes. Todas las funciones ventanas y sus derivadas están incluidas en el código del Vamp plugin Espectrograma IF.



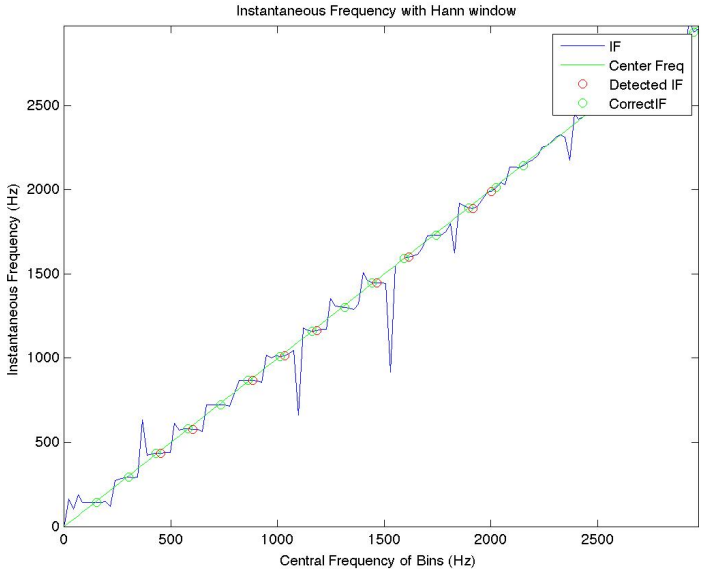
A continuación mostramos el mapeado y la detección de la IF para señales de audio armónicas monofónicas, una señal de habla y una señal armónica de prueba. Los parámetros utilizados son: 2048 muestras del bloque de FFT, una ventana de 1024 muestras y un solapamiento de 256 muestras.

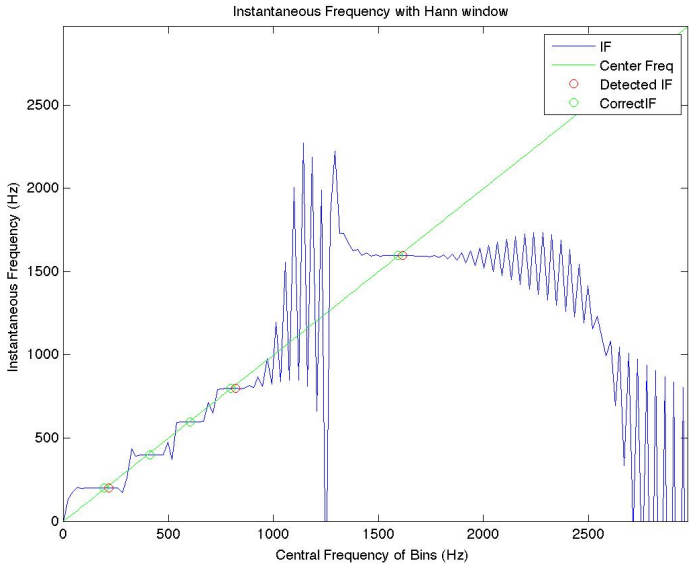
- Función Blackman



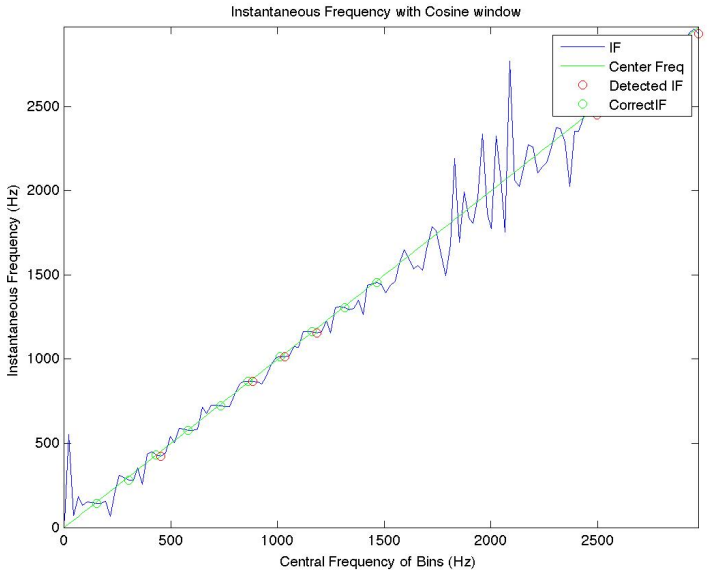


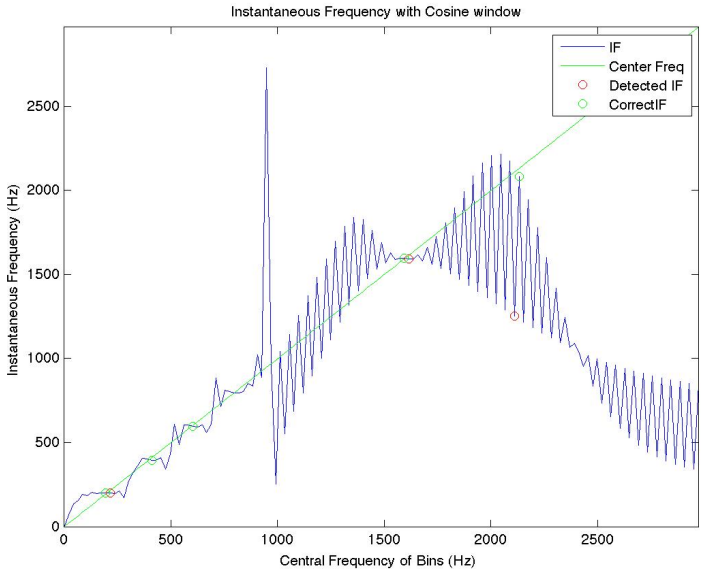
- Función Hann



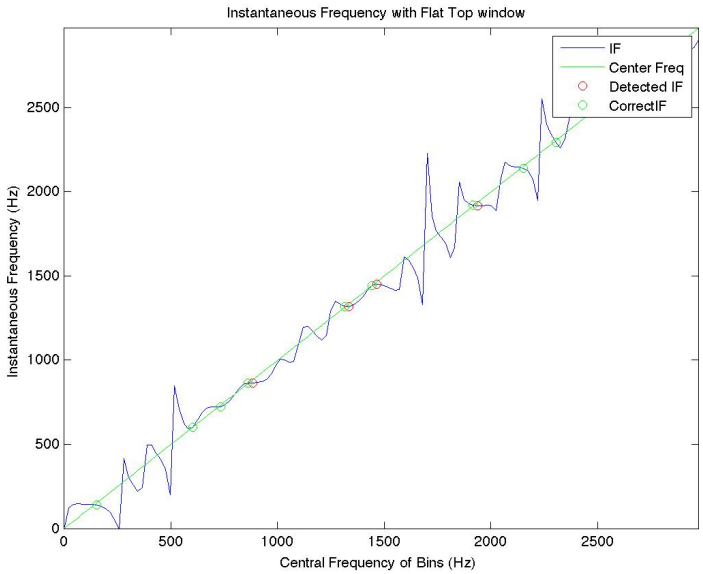


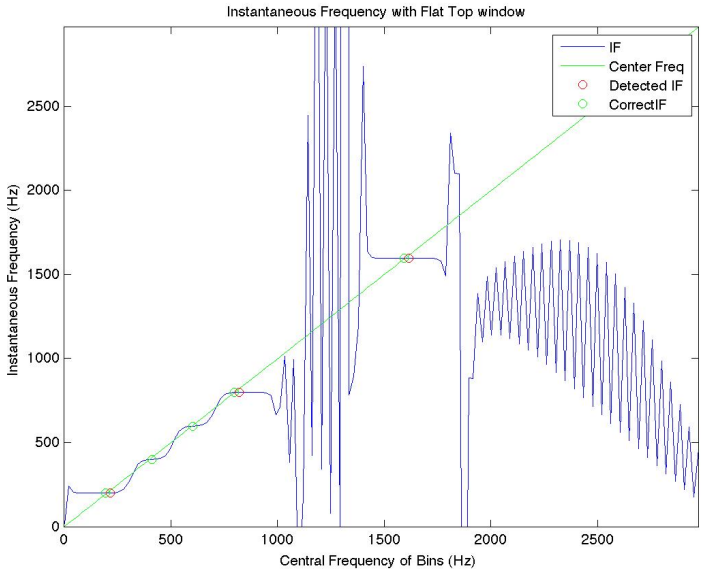
- Función Coseno





- Función Flat Top

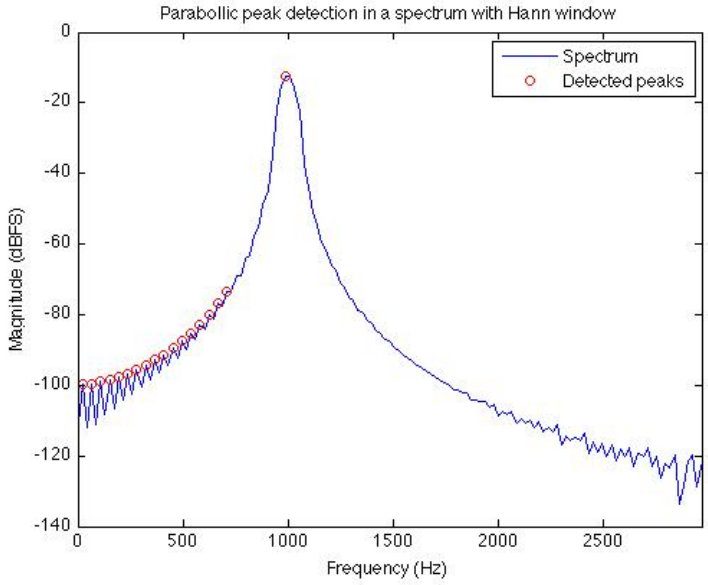




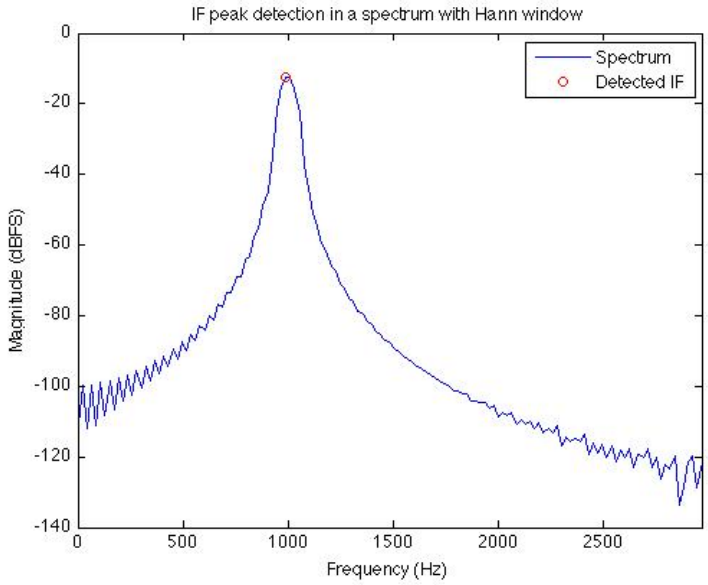
Como podemos observar, aun y escoger los mejores parámetros espectrales de para estimar las frecuencias instantáneas el método depende del tipo de función ventana que utilicemos.

En las siguientes figuras mostramos la diferencia entre la detección de pico parabólica y las frecuencias instantáneas:

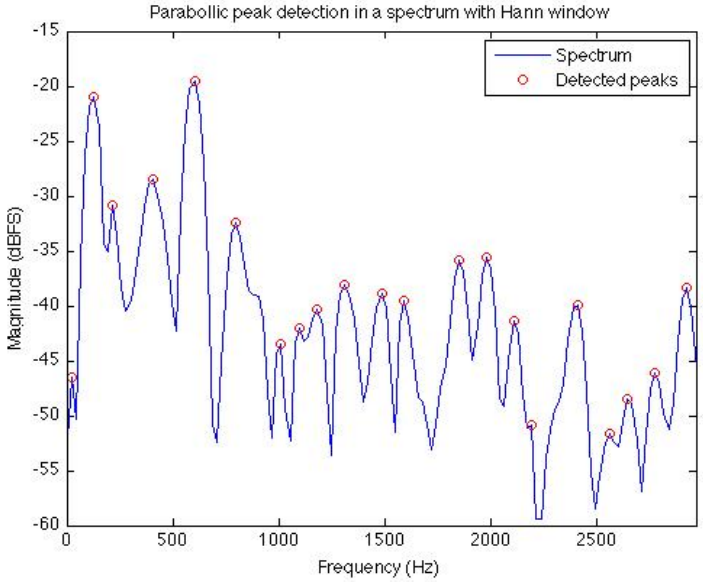
- 1KHz

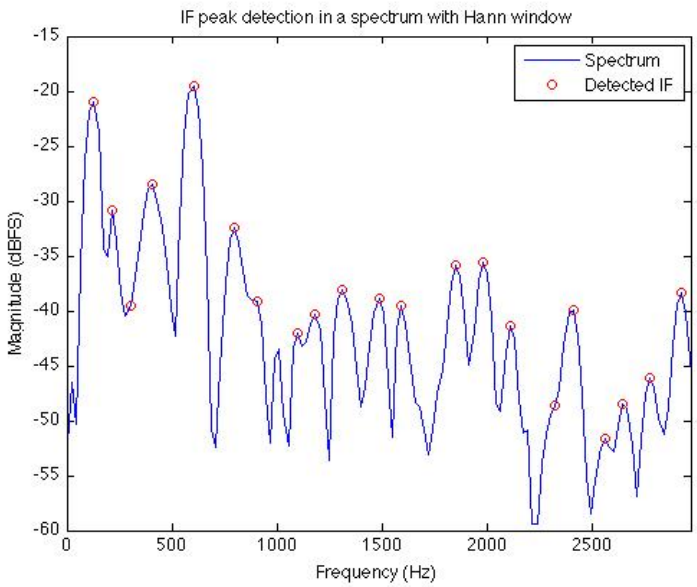


La detección parabólica de picos espectrales, detecta cualquier forma geométrica semejante a una parábola situada por encima del umbral que determinemos. No es la mejor suposición puesto que detecta picos en los lóbulos secundarios o artefactos del análisis espectral que no se consideran picos espectrales.

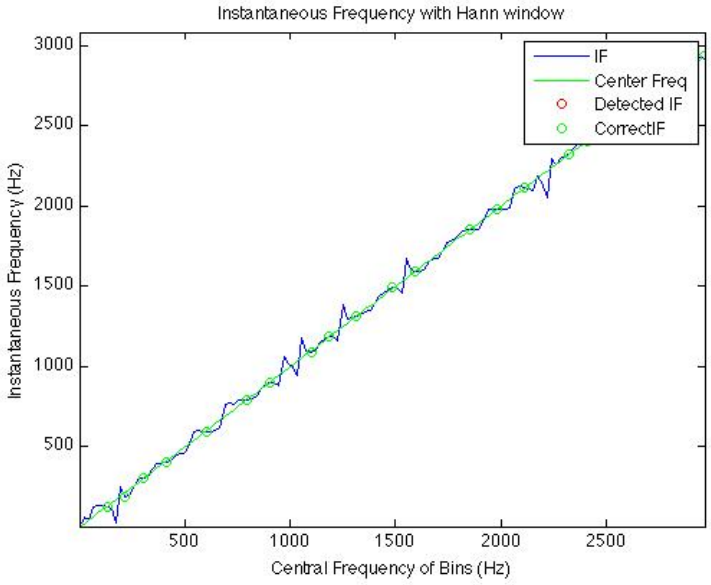


Sin embargo la frecuencia instantánea apunta únicamente al pico espectral sin introducir ninguna distorsión. La detección de las frecuencias instantáneas cuando están separadas más de 200 Hz no presenta complicaciones para la detección. Pero en este caso, nos interesaría para el análisis de señales polifónicas donde las frecuencias o picos espectrales se encuentran mucho más próximas. Veamos los resultados de la detección parabólica de picos y la detección de las frecuencias instantáneas para una señal polifónica.





En las figuras vemos que la detección de las frecuencias instantáneas ha introducido diferentes errores, confundiendo un mínimo con un máximo o algunos casos que se han pasado por alto (en la frecuencia 250Hz aprox. detecta un máximo cuando es un mínimo y en 1KHz no detecta el pico local, etc.). Hemos de considerar que algunos de los picos que no detecta pueden ser contribuciones de la suma de lóbulos secundarios o artefactos del análisis espectral. Por otro lado el mapeado entre frecuencias y las frecuencias instantáneas presenta una geometría bastante compleja que no definen los escalones como en el caso de una señal monofónica, debido al solapamiento de los picos espectrales. Las formas irregulares dificultan la detección de las frecuencias instantáneas mediante la función de picos que hemos implementado.



III. Archivos digitales

Todos los archivos digitales como: aplicaciones Vamp plugins , archivos de audio utilizados en los experimentos.

Vamp plugins implementados:

- HPCP 3.0: Versión para Mac OS X y para Windows.
- IF Spectrogram: Versión para Mac OS X y para Windows

Archivos de audio utilizados en los experimentos:

- 1g.wav
- 1p.wav
- 1c.wav
- 1b.wav
- 110.wav
- Ah.wav
- harmSig.wav
- 1KHz.wav
- CMajor8Instruments.wav
- MelodiaRavel.wav
- w_acoustics1.wav

Todos los archivos de audio que empiezan por un numero pertenecen a la colección de archivos de audio empleada en la evaluación de descriptores tonales. Los que empiezan por un corresponden a la nota C y su letra indica la inicial del instrumento (g es guitarra, p piano, etc.). En el resto de casos corresponden a combinaciones de notas que hemos utilizado en los experimentos. El resto son señales de prueba de todo tipo, señal de habla, señal armónica generada sintéticamente en MATLAB, señal sinusoidal de prueba y señales polifónicas. En el caso de las señales polifónicas, disponemos de una señal con el acorde C que registra diferentes instrumentos, este archivo esta disponible en el Chroma Toolbox CRT (Müller et al.), la primera melodía del Bolero de Ravel, para analizar un fragmento de audio con contribuciones de diferentes instrumentos.