IARC

“IDENTIFICACIÓN AUTOMÁTICA DE RESUMEN EN CANCIONES”

PROYECTO DE FIN DE CARRERA

FACULTAD DE INGENIERÍA
UNIVERSIDAD DE LA REPÚBLICA

10 de Marzo del 2006

Gabriela Saráchaga

Virginia Sartori

Laura Vignoli

Tutores

Alvaro Pardo Federico Lecumberry

Ernesto López Martín Rocamora

CONTENIDO

1 Introducción

1.1 Planteo del problema

Desde hace ya algún tiempo, la cantidad de música que está a nuestra disposición ha aumentado y sigue aumentando considerablemente. Esto se debe a las mejoras tecnológicas disponibles actualmente: disminución de los costos de almacenamiento, aumento del ancho de banda, servicios punto a punto para compartir archivos, etc.

Sumado a lo señalado está la aparición constante de decenas de discos nuevos de distintos artistas, que es imposible conocer en su totalidad. Por lo tanto se hace indispensable contar con herramientas que permitan agilizar la identificación, localización y ordenamiento de las bases de datos musicales. La posibilidad de acceder automáticamente a un resumen que represente la esencia de la canción, permitiría rápidamente tener una noción de la misma, facilitando la tarea.

Esta identificación resultaría útil tanto para organizar las bases existentes como para elegir un nuevo tema musical. Un ejemplo práctico es el servicio que brindan las casas de venta de discos en la búsqueda de nuevas canciones por parte del cliente. Actualmente, al buscar discos en una tienda, se tiene la opción de escuchar las canciones. Al hacer esto, la gente suele acelerar manualmente hasta encontrar una parte "interesante" de la misma, ya que reproducir todas las canciones completas de un disco implicaría demasiado tiempo. Con la identificación automática este proceso se agiliza, obteniendo en un menor tiempo el tramo de la canción deseado de manera directa.

En muchos equipos de audio, existe el modo de reproducción Intro. Este modo reproduce los primeros 20 a 30 segundos de la canción elegida. Para ciertos temas puede resultar interesante este tramo, pero para muchos otros esta parte no es representativa pudiéndose usar el resumen en su lugar.

Otra situación en la que sería útil el resumen, es al descargar música de Internet, para obtener un avance de lo que se va a descargar. Si bien ya existe esta utilidad en algunas aplicaciones, la generación de los mismos se hace manualmente, implicando muchas horas de trabajo de mano de obra especializada.

El resumen también podría ser utilizado como una buena síntesis para ser procesada por algoritmos de análisis de características musicales, en lugar de procesar el tema musical completo. De esta forma, se facilitaría la búsqueda por contenido, la identificación del género musical y los intérpretes, la generación de listas de reproducción de temas similares, etc.

Para concluir este primer acercamiento al problema, cabe señalar que las aplicaciones multimedia están actualmente desarrollándose en todo el mundo, y particularmente la identificación automática de resúmenes. Muestra de ello es la inclusión en el reciente estándar de compresión MPEG-7 (Multimedia Content Description Interface) [20] de un juego de "metadatos", donde existe la posibilidad de almacenamiento de resúmenes multimedia.

1.2 Determinación del resumen

Se ha elegido analizar los géneros musicales de rock y pop para presentar la herramienta, sin perjuicio de que ésta pueda ser utilizada en otros géneros.

Entre las canciones de estos géneros se encuentra muy frecuentemente una estructura del tipo verso-estribillo: A-A-B-A-B-B, donde A representa un verso de la canción y B el estribillo. En general, los versos se caracterizan por mantener una música determinada variando la letra, mientras que el estribillo se repite manteniendo estas dos características constantes. Asimismo, se diferencian en que en el estribillo suele haber un aumento en el dinamismo y aparecen nuevos instrumentos.

Las características antes mencionadas, hacen que el estribillo sea la parte memorable de la canción y aunque esto puede parecer un concepto subjetivo, la mayoría de las personas coinciden en el momento de determinarlo[1].

Se decidió entonces que el resumen a devolver será el estribillo, en el caso de que la canción lo tenga. En caso contrario, será algún fragmento de la canción que aunque no se repita, pueda caracterizarla. Éste se denominará fragmento representativo.

En canciones que presentan estribillo, se establecen las siguientes hipótesis para identificar el resumen:

- Su duración oscila entre 8 y 45 segundos.
- Se repite al menos una vez tanto instrumental como verbalmente con un máximo de 6 veces.

1.3 Objetivo

El objetivo principal de este trabajo es la generación de resúmenes de canciones.

Como objetivo adicional se identificarán en el caso de que la canción tenga estribillo, sus repeticiones, para tener una noción de la estructura.

1.4 Descripción del trabajo

La herramienta desarrollada identifica el resumen de canciones a partir de un archivo de audio[2]. El procedimiento utilizado en la generación del resumen consta de dos partes.

La primera consiste en extraer características en frecuencia a intervalos cortos de tiempo y compararlos para cada pareja de intervalos. Los distintos métodos estudiados para la extracción de características fueron:

[1] Ver apéndice B: Encuesta sobre detección de estribillo.
[2] Se analiza el archivo de audio en formato WAV (Waveform Audio Format), monofónico.

- Coeficientes Cepstrales de Frecuencia Mel (*MFCC*)
- Vectores de Cromas (VC)
- Transformada de Constante Q (CQT)

MFCC contiene información relativa al timbre de música polifónica[3], mientras que VC y CQT caracterizan los intervalos según la línea melódica.

Las características halladas se comparan para cada pareja de intervalos de tiempo empleando una medida de distancia elegida de acuerdo al método utilizado. Esta comparación resulta en una matriz que llamaremos *Matriz de Similitud.* En ella se puede estudiar "visualmente" la canción como una imagen en escala de grises, donde los píxeles claros representan más similitud que los oscuros.

En la segunda parte, se estudia la matriz para determinar el resumen. El método principal aplicado identifica al resumen mediante la búsqueda de repeticiones, por lo tanto, bajo ciertas hipótesis, el resumen detectado es el estribillo. Llamaremos a este método, *Identificación de Estribillo (IE).*

Cuando la canción no presenta estribillo, la búsqueda de repeticiones no es adecuada para identificar el resumen. También existe la posibilidad de que la canción posea estribillo pero la herramienta no logre detectarlo. Por estas razones surge la necesidad de buscar otra forma de identificar el resumen.

Se consideró entonces un segundo método, para los casos en que el anterior no obtenga una detección exitosa. Consiste en buscar la máxima similitud de un fragmento respecto a la totalidad de la canción, y devuelve un resumen alternativo. En adelante se llamará a este método *Identificación de Fragmento Representativo (IFR).*

En la siguiente figura se presenta un esquema del procedimiento explicado anteriormente.

[3] Este concepto se explica en la sección3.2.2.

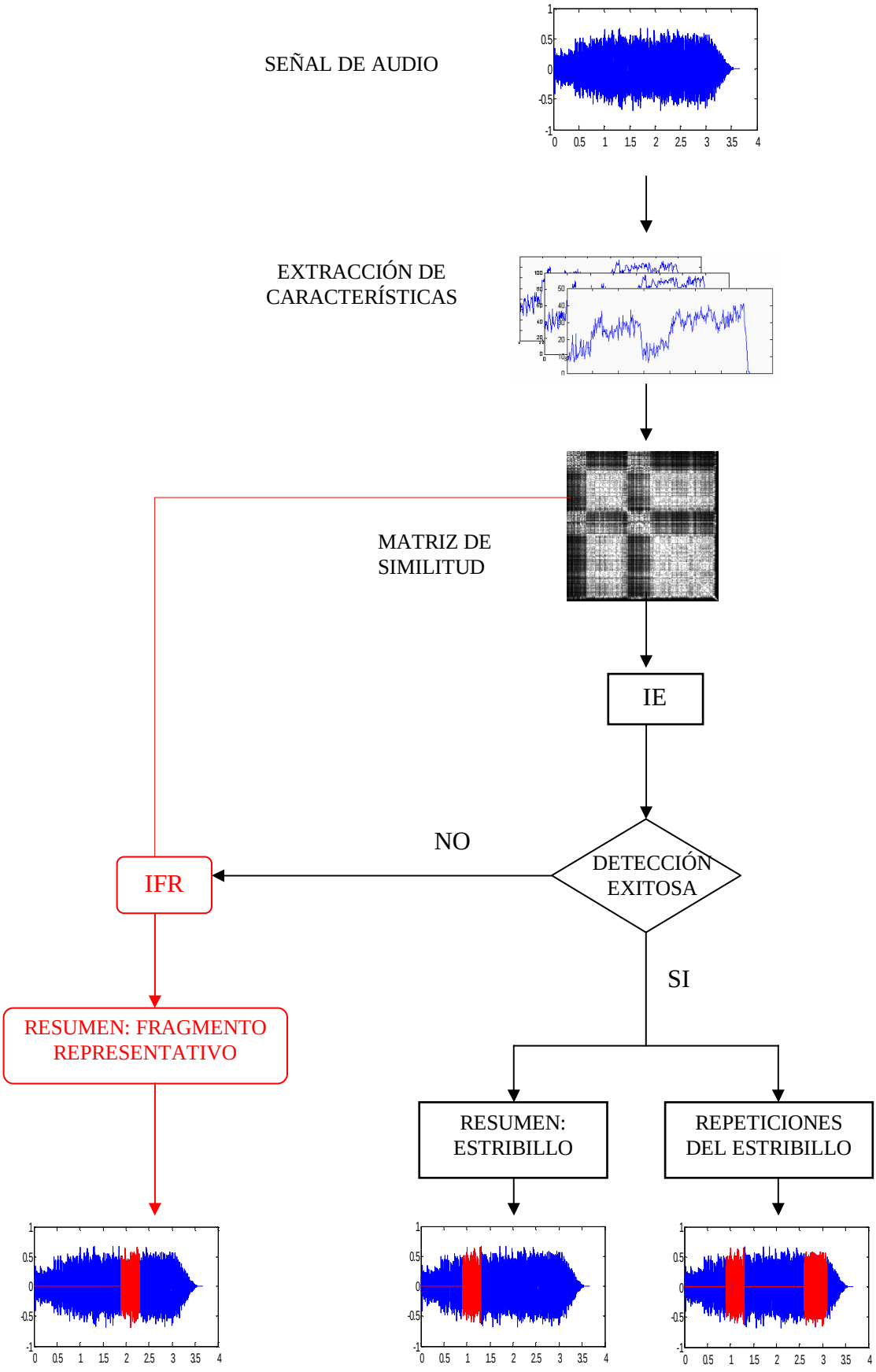


Figura 1-1: Esquema del procedimiento general de identificación de resumen

1.5 Antecedentes

El análisis estructural de canciones con el propósito de generar automáticamente un resumen, ha sido ampliamente estudiado en los últimos años.

Investigaciones recientes han presentado distintas técnicas aplicadas originalmente al reconocimiento de voz para el estudio de la música. Éstas se pueden dividir en dos ramas principales de acuerdo a qué fragmento de la canción se considera como el más representativo.

La primera y más extendida, utilizada en [6], [8], [10], [19], *et al.*, estudia patrones de repetición para determinar el segmento más representativo de una canción. La segunda presentada en [9] plantea el resumir cualquier fuente de audio mediante un fragmento de largo fijo, que sea lo más similar posible a su totalidad.

Los primeros en exponer el problema fueron Chu y Logan en [10], quienes utilizaron como caracterización del audio los coeficientes cepstrales de frecuencia Mel (*MFCC*s), desarrollados originalmente para reconocimiento de voz. Este método analiza aspectos de la música relacionados con el timbre. Esta misma línea de investigación fue seguida por Cooper y Foote [19].

En [18] Foote introduce el concepto de Matriz de Similitud, una forma novedosa de visualizar la comparación de canciones. Este acercamiento permitirá la aplicación de técnicas de tratamiento de imágenes al audio.

En la misma dirección realiza su estudio Masataka Goto [8], introduciendo los VC como método de caracterización para detectar melodías a partir de las notas musicales.

Tal vez el enfoque más novedoso, es el planteado por Lu, Wang y Zhang [6]. En este caso se utiliza la Transformada de Constante Q (CQT), para buscar líneas melódicas similares, al igual que los VC. La innovación viene dada por una medida de distancia, que enfatiza la similitud melódica y suprime el peso de la información del timbre.

En contraste con estos estudios donde se analizan patrones de repetición, Foote y Cooper presentan en [9] un punto de vista que reduce el análisis. Éste se basa en encontrar un resumen, encontrando el fragmento de largo fijo con la máxima similitud respecto al resto de la canción.

En los trabajos anteriores se descomprime el audio de forma de trabajar con la señal PCM. Shao, Xu, Wang y Kanhanhalli [17] extraen las características de la señal directamente del archivo en MP3, disminuyendo así el tiempo de procesamiento.

1.6 Organización

A continuación se presenta cómo se desarrollarán los siguientes capítulos.

Capítulos 2 y 3	*Procesamiento de Señales digitales de Audio y Fundamentos de la música*: conceptos y términos necesarios para la comprensión de los capítulos posteriores.
Capítulos 4 y 5	*Métodos de extracción de características y Matriz de Similitud*: descripción de las técnicas empleadas en la caracterización del audio y herramienta para la comparación de las mismas.
Capítulos 6 y 7	*Identificación de estribillo e Identificación de fragmento representativo*: métodos de análisis de la matriz de similitud para la obtención del resumen.
Capítulos 8 y 9	*Evaluación y elección de métodos para el sistema* e *Implementación*: descripción de la metodología de validación y determinación de los parámetros a utilizar en la implementación de la herramienta.
Capítulos 10 y 11	*Validación* y *Conclusiones*: resultados, conclusiones y trabajos futuros.

2 Procesamiento de Señales digitales de Audio

2.1 Resumen

Las señales de audio que se van a analizar están representadas en formato digital, por lo tanto se explicarán algunos conceptos básicos del procesamiento de señales digitales que luego se utilizarán en el resto de los capítulos.

2.2 Señales digitales

Las señales en tiempo discreto son aquellas que se representan matemáticamente como una secuencia de números. Además del carácter de estar definidas en tiempo discreto, la amplitud de la señal puede ser también discreta. Las señales digitales son aquellas que son discretas tanto en el tiempo como en la amplitud [1].

2.2.1 Muestreo

En la mayoría de los casos las señales en tiempo discreto surgen de tomar muestras de una señal analógica. De esta forma, el valor numérico del n-ésimo número de la secuencia es igual al valor de la señal analógica $x_a(t)$, en el instante temporal nT_s, es decir,

$$\hat{x}(n) = x_a(nT_s), \quad -\infty < n < \infty$$

La cantidad T_s se denomina período de muestreo, y su inversa es la frecuencia de muestreo f_s.

El teorema de Nyquist garantiza que para poder reconstruir una señal a partir de sus muestras, se debe utilizar una frecuencia $f_s \geq 2 f_N$, o sea al menos el doble de f_N. Siendo f_N la componente de más alta frecuencia de la señal.

El espectro de frecuencias del sonido audible por los humanos, es aproximadamente de 20 Hz a 20 kHz. Por esto, las señales de audio se muestrean generalmente a 44100 Hz, o sea más del doble de la máxima frecuencia audible. Éste es el caso del CD de audio.

El contenido en frecuencia de las señales de voz puede abarcar hasta 15 khz o más, pero la voz es altamente inteligible incluso con bandas de frecuencia limitadas a unos 4 kHz. Ese es el caso de los sistemas telefónicos comerciales donde la frecuencia de muestreo estándar utilizada para la voz es de 8 kHz.

En la etapa de muestreo se obtiene una señal en tiempo discreto cuyas amplitudes $\hat{x}(n)$ son valores continuos. Para digitalizar la señal resta discretizar esos valores (cuantizarlos).

2.2.2 Cuantización

El propósito del cuantizador es transformar la muestra de entrada $\hat{x}(n)$ en un valor $x(n)$ de un conjunto finito de valores preestablecidos. Esto se realiza redondeando los valores de las muestras hasta el nivel de cuantización más próximo.

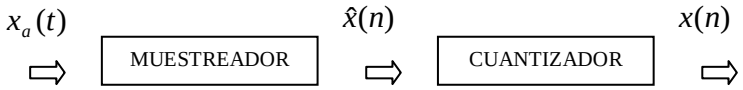


Figura 2-1: Representación conceptual de la digitalización de una señal analógica

La precisión de los datos dependerá del número de bits con que se codifiquen los niveles de cuantización [2]. Por tanto, se introduce un ruido de cuantización que se asume como ruido blanco.

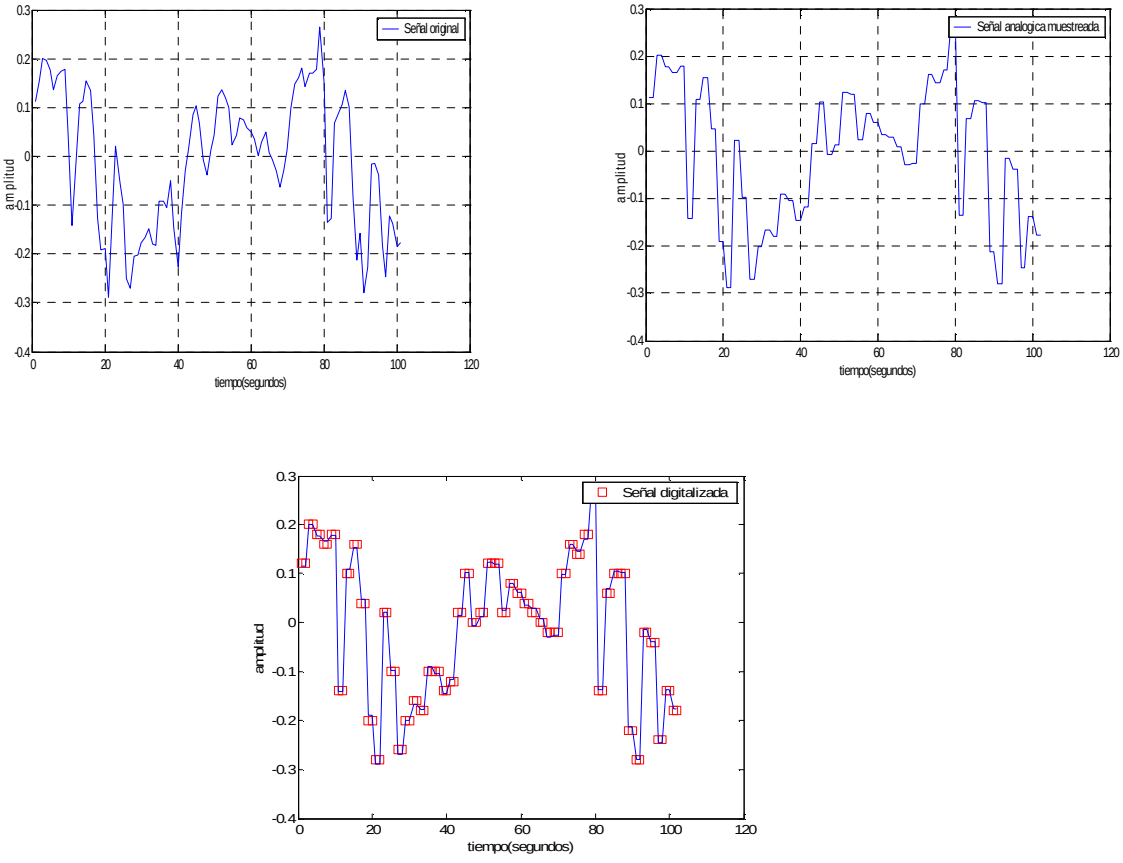


Figura 2-2: Etapas de la digitalización. a) Señal original, b) Señal muestreada con amplitudes analógicas y c) Señal digital.

2.3 Representación en el dominio de la frecuencia de señales en tiempo discreto

2.3.1 Transformada Discreta de Fourier

Los sistemas lineales e invariantes en el tiempo, cumplen ciertas propiedades que permiten la representación de las señales en frecuencia [1].

Una de las propiedades es que la respuesta a secuencias sinusoidales es también sinusoidal, de igual frecuencia y con amplitud y fase determinadas por el sistema. Esta propiedad hace que las representaciones de las señales mediante sinusoides o exponenciales complejas (es decir, las representaciones de Fourier) sean muy útiles.

Para las secuencias de duración finita, se utiliza la Transformada Discreta de Fourier (DFT). Se llaman secuencias base a las exponenciales complejas que se utilizan para representar la señal.

Dada una señal en tiempo discreto $x(n)$ con N muestras, su transformada $X(k)$ está dada por:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}}$$

En la práctica, el costo computacional del cálculo de la DFT se reduce utilizando la transformada rápida de Fourier (FFT)[4].

2.3.2 Transformada de Fourier de tiempo corto (STFT)[5]

A partir de la transformada discreta de Fourier, se realiza un análisis del contenido en frecuencia de las señales. Pero en la práctica, para aplicaciones de audio por ejemplo, las propiedades de la señal no son estacionarias y una sola DFT no es suficiente para describir el comportamiento de esas señales.

La transformada de Fourier en tiempo corto soluciona este problema calculando la DFT a intervalos de la señal. Este proceso se llama “enventanado”.

Enventanado

El enventanado consiste en agrupar las muestras de la señal $x(n)$ en bloques de N elementos, y multiplicarlas por una ventana $w(n)$ (Figura 2-3).

[4] FFT proviene de la expresión en inglés, “Fast Fourier Transform”
[5] STFT proviene de la expresión en inglés, “Short Time Fourier Transform”

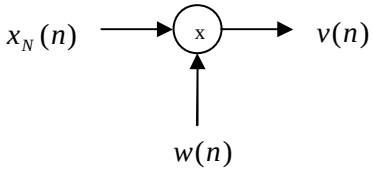


Figura 2-3: Enventanado

El principal propósito de la ventana en la transformada de Fourier de tiempo corto es limitar la extensión de la secuencia que se va a transformar. Esto debe hacerse para que las características espectrales sean razonablemente estacionarias en el intervalo de duración de la ventana.

Para mantener la continuidad de la información de la señal, es muy común realizar el enventanado con bloques de muestras solapados entre sí, de esta forma no se pierden los eventos en la transición entre ventanas.

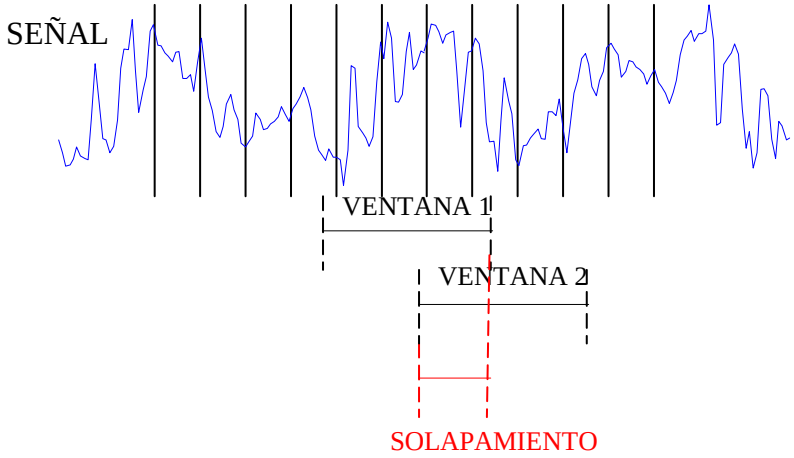


Figura 2-4: Solapamiento

Cuanto más rápidamente cambien las características de la señal, más corta deberá ser la ventana para poder detectar esos cambios en el tiempo. Por otra parte, a medida que decrece la longitud de la ventana, se reduce la resolución frecuencial, es decir, la capacidad de distinguir componentes cercanas en frecuencia.

Por tanto, aparece un compromiso en la selección de la longitud de la ventana entre la resolución en tiempo y en frecuencia.

Además de la longitud se debe elegir la forma de la ventana, o más específicamente, el tipo de suavizado que se requiere en los extremos de la misma [1].

Las ventanas más utilizadas se definen para N muestras como:

- Rectangular: $w(n) = 1$
- Hanning: $w(n) = \frac{1}{2} - \frac{1}{2} \cdot \cos\left(\frac{2\pi n}{N}\right)$

- Hamming: $w(n) = \frac{27}{50} - \frac{23}{50} \cdot \cos\left(\frac{2\pi n}{N}\right)$
- Bartlett: $\begin{cases} \frac{2n}{N} & 0 < n < \frac{N}{2} \\ 2 - \frac{2n}{N} & \frac{N}{2} < n < N \end{cases}$
- Blackman: $w(n) = \frac{21}{50} - \frac{1}{2} \cdot \cos\left(\frac{2\pi n}{N}\right) + \frac{2}{25} \cdot \cos\left(\frac{4\pi n}{N}\right)$

Cada ventana se caracteriza por la forma de sus lóbulos central y laterales en frecuencia. Se requiere de una ventana, que su lóbulo central sea lo más angosto posible y que los lóbulos laterales sean pequeños para tener una buena resolución en frecuencia.

La ventana rectangular posee el lóbulo central con menor ancho de banda de todos, pero sus lóbulos laterales decaen muy lentamente. Estos hacen aparecer el efecto de 'ripple' (fenómeno de Gibbs), no deseado por la distorsión armónica que generan. El resto de las ventanas tiene cada una distintas propiedades, que según la aplicación podrán ser de un modo u otro ventajosas.

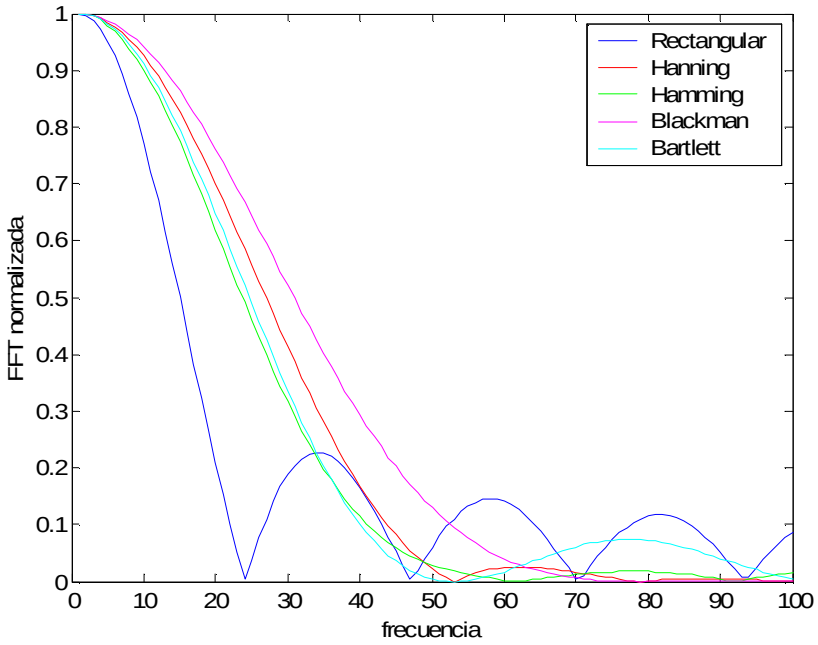


Figura 2-5: Espectros de las distintas ventanas

Dado el compromiso existente entre la resolución y la distorsión armónica, para el caso de señales de audio se utiliza la ventana de Hanning. Ésta suaviza la señal en los extremos, eliminando discontinuidades generadas por aplicar la DFT a una señal no periódica, y por tanto el 'ripple'. La desventaja que se experimenta es la distorsión que sufre la señal original.

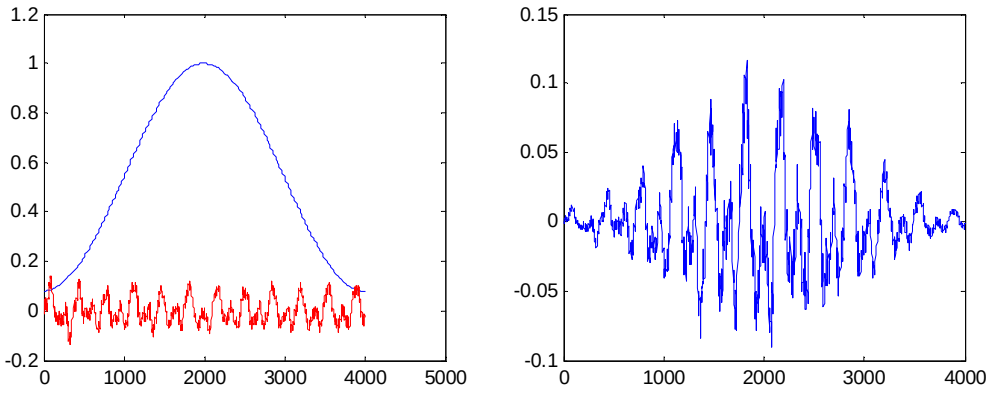


Figura 2-6: A la izquierda, ventana de Hanning en azul y señal de audio en rojo; a la derecha la multiplicación de ambas.

2.3.3 Transformada discreta del coseno (DCT)

La DCT es una transformada muy similar a la DFT en donde las secuencias base son cosenos y la representación de una señal real mediante esta transformada, es también real [1].

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos\left[\frac{k\pi}{N}\left(n + \frac{1}{2}\right)\right]$$

Se utiliza en muchas aplicaciones de compresión de datos con preferencia sobre la DFT debido a una propiedad que se denomina generalmente “compactación de la energía”. La DCT tiende a concentrar la mayor parte de la información de la señal en los coeficientes de baja frecuencia. Gracias a esto, se necesita un menor número de coeficientes para representarla.

Esto se ejemplifica en la figura siguiente:

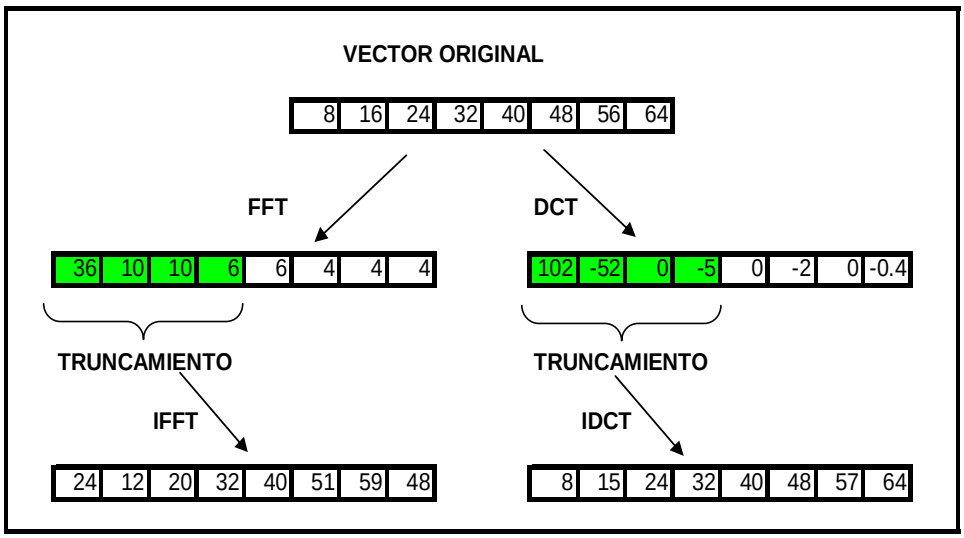


Figura 2-7: Compactación de la DCT

2.4 Filtrado de señales

Los filtros son una clase de sistemas lineales e invariantes con el tiempo particularmente importante. Estrictamente hablando, el filtro selectivo en frecuencia sugiere un sistema que deja pasar ciertas componentes de frecuencia y rechaza completamente otras. Pero en un sentido más amplio, cualquier sistema que modifique ciertas frecuencias con respecto a otras, también se denomina filtro [1].

En las siguientes figuras se pueden ver tres tipos de filtros según los rangos de frecuencias que conservan del espectro: Filtro pasaaltos, pasabanda y pasabajos.

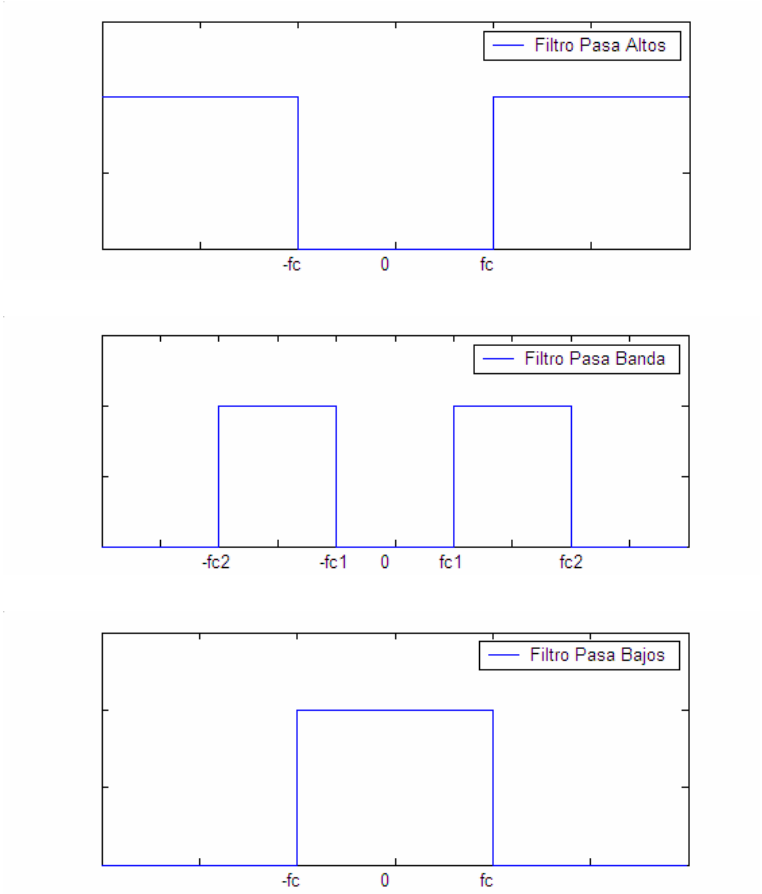


Figura 2-8: Tipos de filtros

3 Fundamentos de la música

3.1 Resumen

En este capítulo se presentarán conceptos básicos de la música, necesarios para un mejor entendimiento de los capítulos posteriores.

3.2 Características de los sonidos musicales

Para describir un sonido musical se utilizan tres términos: altura, timbre e intensidad. Todo sonido tiene una duración y, a lo largo de ésta, cualquiera de estos tres parámetros puede variar (los sonidos naturales jamás son perfectamente estables o constantes).

3.2.1 La altura y la frecuencia

La altura está directamente relacionada con la frecuencia de oscilación de una onda sonora, pero ambos términos no son sinónimos. De hecho, muchos sonidos (como los percusivos) no tienen una altura definida. El motivo de esta aparente paradoja es que, mientras la frecuencia es una propiedad física indisociable de todo aquello que, como el sonido, vibra u oscila, la altura es una cualidad subjetiva que percibimos sólo en algunos sonidos [27].

Al golpear, por ejemplo, un bombo o un platillo, se puede sin duda afirmar que el platillo suena más agudo que el bombo, pero no se puede decir si estos sonidos correspondían a un *Do* o a un *La*.

Lo que hace que un sonido posea o no una altura clara es básicamente, su periodicidad. Es necesario que un sonido sea aproximadamente periódico, es decir que su frecuencia de oscilación no varíe (o varíe poco) dentro de un determinado lapso de tiempo, para que se llegue a percibir una altura. Si se analiza un lapso mayor de tiempo, la frecuencia sí puede variar, y en este caso, lo que se percibe son alturas variables en el tiempo.

3.2.2 El timbre y los armónicos

El timbre podría definirse como el "color" de un sonido, y es lo que ayuda a caracterizar y distinguir diferentes tipos de instrumentos, o a reconocer a las personas por su voz.

Cada instrumento musical tiene propiedades acústicas determinadas por su forma y material. Los instrumentos musicales requieren de la intervención del intérprete que provea la energía que iniciará el sonido. Para que el sonido se considere musical con una altura específica, debe provenir de vibraciones “periódicas”.

El carácter del sonido que produce cada instrumento es en parte debido a las vibraciones asociadas al proceso de iniciación, la forma de tocar las cuerdas en una guitarra por ejemplo. En parte es debido también a las vibraciones características asociadas al elemento resonador, o sea la caja de madera en una guitarra. El sonido provocado por estas vibraciones es sostenido inicialmente, pero generalmente decae cuando se deja de brindar energía al instrumento.

El proceso de iniciación también determina el tipo de sonido, en el piano por ejemplo, no es lo mismo presionar una tecla de forma abrupta que mantenerla presionada. Se ha observado que si se elimina el efecto de la iniciación o ataque en una grabación de música, es más difícil distinguir un instrumento de otro. Los factores externos también pueden influenciar en el timbre, por ejemplo un instrumento desplazándose en una sala [28]

Un *La* de 440 Hz en una clarinete suena diferente que el mismo *La* en un saxo; aunque ambos tienen la misma altura, sus timbres no son iguales [27].

En la Figura 3-1, se muestran las variaciones en el tiempo de estos dos sonidos, comparados con una onda sinusoidal pura de la misma altura. Dicha figura puede ayudar a comprender la naturaleza física del timbre. En los tres fragmentos, el período es el mismo (pues tienen la misma altura), pero sus formas son diferentes.

El motivo de esta diferencia de forma, es que las ondas de los sonidos naturales son más complejas porque vibran con varias frecuencias simultáneas. En la naturaleza no se encuentran sonidos puros con una sola frecuencia, como el de la Figura 3-1 c); éstos sólo son obtenibles por medios electrónicos.

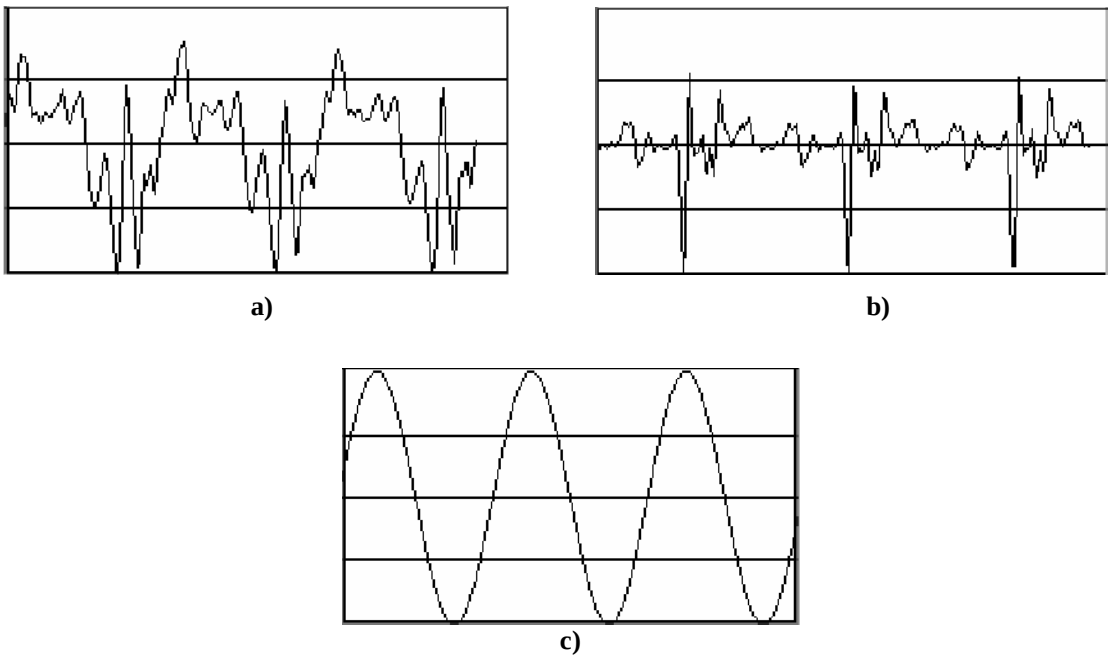


Figura 3-1: Fragmentos de tres sonidos de la misma altura (La de 440 Hz) en un clarinete (a), un saxofón (b) y una onda sinusoidal pura (c)

En los sonidos naturales, la frecuencia de vibración más baja es la que determina normalmente el período y la altura, y se denomina *frecuencia fundamental*. Las restantes frecuencias, que suelen ser múltiplos de la frecuencia fundamental junto con esta última se denominan *armónicos*. Cada tipo de instrumento tiene, según la forma en que se construyen, una serie diferente de armónicos de amplitudes diferentes, que son los que definen su timbre y otorgan las "señas de identidad" al instrumento.

Resumiendo, el timbre indica la forma en que la energía se distribuye entre los distintos armónicos y la forma en que esta distribución cambia en el tiempo.

En las canciones, se produce una superposición de sonidos de varios instrumentos. Para este caso se puede definir una textura de la música, que coincide con la idea de un timbre polifónico. Esta textura queda determinada por los aportes del timbre de cada instrumento presente en el fragmento de canción que se esté analizando.

3.2.3 La intensidad y los decibelios

La altura queda determinada por el número de oscilaciones por unidad de tiempo, mientras que la intensidad depende del cuadrado de la amplitud de estas oscilaciones. La percepción de la intensidad sonora es, en realidad, un fenómeno auditivo muy complejo, mucho más que el de la altura, y lo que sigue es una simplificación [27].

Las intensidades de diferentes sonidos pueden variar, en varios millones de órdenes de magnitud (es decir, el sonido más intenso que podamos oír, lo será varios millones de veces más, que el más tenue). Por ello, la intensidad se mide en una escala logarítmica, los decibeles (dB), de acuerdo con la siguiente fórmula:

$$\text{Nivel de intensidad en decibeles (dB)} = 10 \times \log_{10} (\text{amplitud}^2 / \text{amplitud referencia}^2).$$

Esta expresión determina un nivel o diferencia de intensidad entre dos amplitudes. El origen (0dB) corresponde al umbral de audición (mínimo sonido audible). Por debajo de este valor se obtendría el auténtico silencio.

Por encima de los 130 dB se produce una sensación dolorosa. Valores superiores prolongados llegan a destrozar el tímpano. En la Tabla 3-1 se muestran algunos valores típicos.

Descripción	Nivel (dB)	Relación de intensidad
Despegue de cohete espacial	190	10^{19}
Despegue de un reactor	150	10^{15}
Umbral de dolor	130	10^{13}
Concierto de heavy metal	120	10^{12}
Martillazos sobre una plancha metálica (a 50 cm)	110	10^{11}
Tráfico en calle concurrida	70	10.000.000
Conversación normal (a 1 m)	60	1.000.000
Restaurante concurrido	50	100.000
Casa en la ciudad	40	10.000
Iglesia vacía	30	1.000
Estudio de grabación	20	100
Umbral de audición	0	1

Tabla 3-1: Ejemplos de niveles sonoros en dB respecto al umbral de audición.

3.3 Percepción de altura

La percepción de la altura está íntimamente vinculada a la frecuencia, aunque la afectan en menor medida, la intensidad, la complejidad espectral y la duración.

Se utilizan diferentes escalas para representar las alturas.

3.3.1 Escala cromática

Un fenómeno muy importante relacionado con la percepción de las alturas, es el de la octava. Si escuchamos dos sonidos cuyas frecuencias guardan una relación de 2:1 (por ejemplo 400 Hz y 200 Hz), serán percibidos como muy similares.

El motivo es que los dos distan exactamente una octava. Dado que cada vez que se dobla la frecuencia se sube una octava, un sonido de 880 Hz estará dos octavas por encima de uno de 220 Hz.

Esta idea de octava se repite en casi todas las culturas, a lo largo de la historia. Lo que sí varía enormemente de una cultura a otra es el número de subdivisiones que se aplican a la octava.

En la música occidental, la octava se divide en doce alturas o semitonos, de las cuales siete tienen “nombre propio” y corresponden a las teclas blancas de un piano. Las cinco restantes (que corresponden a las teclas negras) pueden tomar el nombre de la inmediatamente anterior, en cuyo caso se les añade el símbolo # (sostenido), o bien de la posterior, en cuyo caso se les añade el símbolo b (bemol).

De esta forma las doce notas de una octava pueden nombrarse de dos maneras diferentes (*"do, do#, re, re#, mi, fa, fa#, sol, sol#, la, la#* y *si"*, o bien "*do, reb, re, mib, mi, fa, solb, sol, lab, la, sib* y *si"*). En la nomenclatura sajona las notas se designan mediante letras mayúsculas, de acuerdo con la siguiente equivalencia:

Do	Re	Mi	Fa	Sol	La	Si
C	D	E	F	G	A	B

La relación de frecuencias entre cualquier nota y la siguiente es siempre igual a $2^{1/12}$ (1.05946). De esta forma, al avanzar doce semitonos (una octava), obtenemos un factor de $(2^{1/12})^{12}$ que es efectivamente igual a 2.

Los intervalos musicales corresponden a la relación entre dos notas de una escala musical. Por ejemplo, una *quinta* es el intervalo entre un *do* y un *sol* pues en la secuencia *do-re-mi-fa-sol* hay 5 notas. También pueden describirse en términos de la relación de frecuencias, por ejemplo, la quinta corresponde a una relación de 3/2.

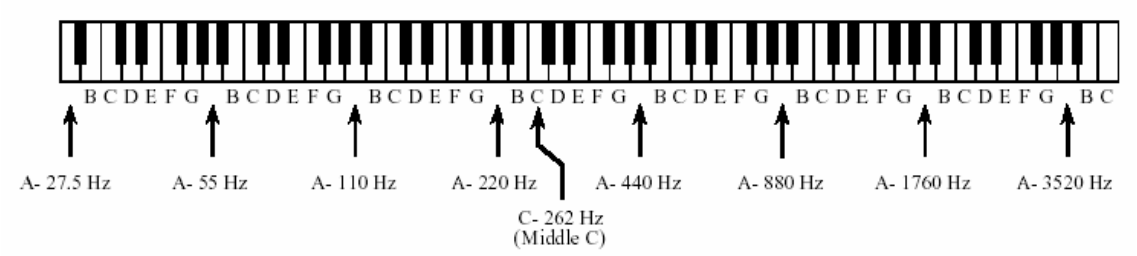


Figura 3-2: Siete octavas representadas en un piano

Escala Cent

La escala Cent es una representación lineal de las frecuencias de las notas. En esta escala 100 *cents* corresponden a un semitono y 1200 *cents* a una octava:

$$fcent = 1200 \cdot \log(f_{Hz} / (440 \cdot 2^{\frac{3}{12}-5}))$$

En la Figura 3-3 se puede ver la relación que existe entre los *cents* y las frecuencias en Hz.

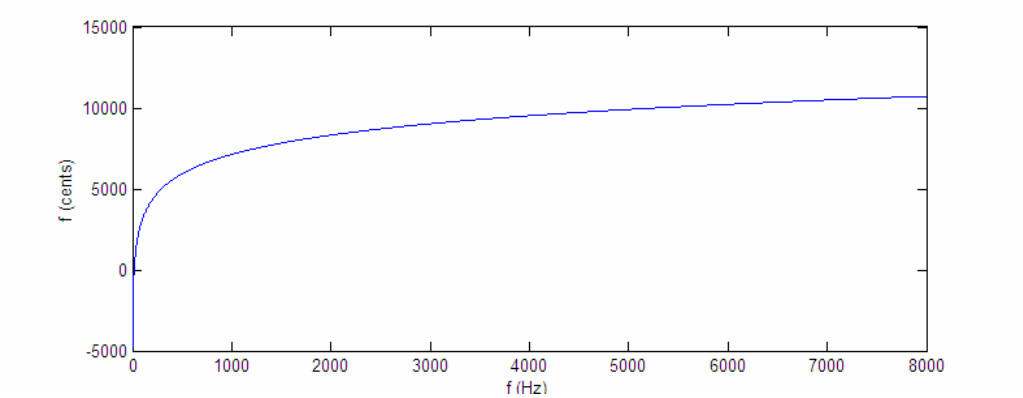


Figura 3-3: Cents en función de Hz.

3.3.2 Escala de Mel

Hoy en día, se conoce con bastante precisión el funcionamiento del oído humano como detector de frecuencias, a modo de analizador espectral. Concretamente, el órgano de Corti realiza un análisis espectral por bandas discretas superpuestas denominadas bandas críticas.

El ancho de las bandas críticas tiene un crecimiento de forma aproximadamente lineal hasta la frecuencia de 1 KHz y aproximadamente logarítmico por encima de ella. El paso de la primera a la segunda zona no es brusco sino progresivo[12].

Stevens y Volkman (1940) desarrollaron una escala basada en la percepción auditiva humana llamada escala de Mel. Esta escala mapea las frecuencias medidas en Hz, en lo que se percibe como altura [4].

Para crearla se utiliza el siguiente procedimiento:

1. Se toma como frecuencia de referencia 1000 Hz y se le asignan 1000 Mels.

2. Se les presenta una señal a los oyentes y se les pide que cambien la frecuencia hasta que la altura percibida sea el doble de la señal de referencia, luego 10 veces la referencia, etc. Y luego la mitad de la referencia, 1/10 de la referencia, etc.

3. De estos datos se construye la escala de Mel:

$$Mel(f) = 1127.01048 \log\left(1 + \frac{f}{700}\right)$$

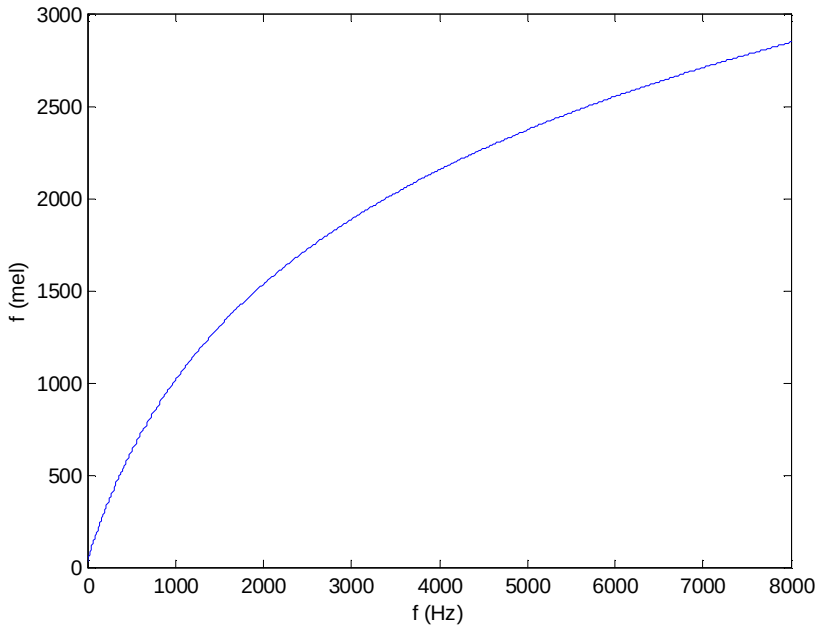


Figura 3-4: Frecuencia en mels en función de la frecuencia en Hz.

4 Métodos de extracción de características

4.1 Resumen

A continuación se presentan los métodos estudiados para la extracción de características. Primeramente se explicará el método de Coeficientes Cepstrales de frecuencia Mel (*MFCC*), que permite medir la similitud de segmentos mediante el análisis de las diferencias del timbre.

Seguidamente se estudian los métodos denominados Vectores de Croma (VC) y Transformada de constante Q (CQT), que se basan en las diferencias en la melodía. Finalmente se realiza un análisis y una comparación cualitativa de los resultados obtenidos.

4.2 Introducción

La caracterización del audio mediante el análisis del espectro de frecuencias, permite capturar ciertos aspectos como ser cambios de timbre y altura. Un ejemplo de estas transiciones se observa en la Figura 4-1.

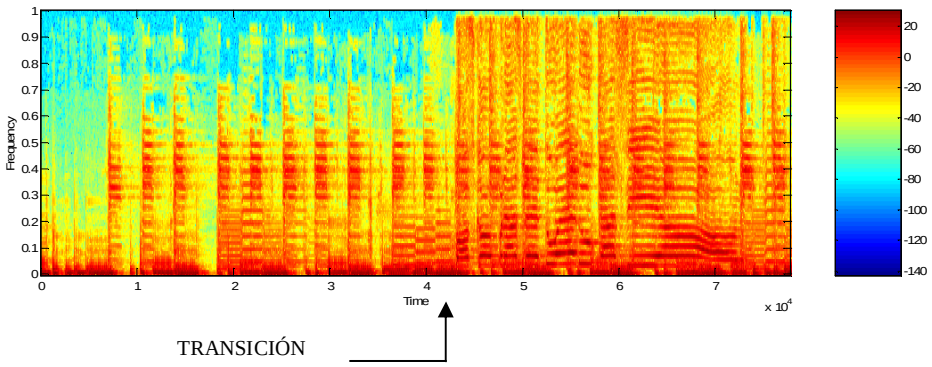


Figura 4-1: Espectrograma: se evidencia la estructura del audio.

En los métodos presentados a continuación se estudia la señal realizando un enventanado temporal y comparando los parámetros extraídos al espectro de cada ventana[6].

[6] Ver 2.3.2 Transformada de Fourier de tiempo corto (STFT)

4.3 Coeficientes Cepstrales de Frecuencia Mel (MFCC)

4.3.1 Introducción

Este algoritmo fue originalmente propuesto por Davis y Mermelstein (1980) [12] en un sistema de reconocimiento automático de voz. Actualmente es uno de los más utilizados en esta área. Ha demostrado también, buenos resultados en extracción de información en música polifónica, aplicándose a la detección de instrumentos, de cantantes y de generación de listas de reproducción entre otros [10][19].

Los *MFCC* son una representación compacta del espectro de una señal de audio, ya que se utilizan unos pocos coeficientes para representarla. Además, y no menos importante, considera la percepción humana de la altura, mediante la escala de Mel [16].

4.3.2 Procedimiento extracción de los MFCC

El procedimiento es el siguiente: se realiza el enventanado de la señal en bloques de T segundos solapados, mediante ventanas de Hanning.

Los pasos a seguir para cada bloque $v(n)$ son los siguientes:

- DFT: cálculo de la transformada discreta de Fourier (DFT)
- DEP: estimación de la Densidad Espectral de Potencia, $DEP(k) = \left|DFT(v(n))\right|^2$.
- Vector de coeficientes Mel (VCM): obtenido del filtrado de la DEP mediante un banco de N_f filtros triangulares solapados (ver Figura 4-4), integrando la energía presente en cada banda (se obtienen N_f componentes).
- Coeficientes cepstrales *MFCC*: representación del vector de coeficientes Mel en decibeles y aplicación de la transformada coseno (DCT). Esta transformada realiza una compactación de los coeficientes obteniéndose N_{mfcc} elementos.

Figura 4-2: Diagrama del procedimiento de extracción de características mediante MFCC para cada bloque de muestras $v(n)$

Se pasará a detallar los puntos más relevantes del procedimiento.

Banco de filtros Mel

Para caracterizar la señal de audio, se utiliza un banco de N_f filtros triangulares solapados, que se adecuan al comportamiento del oído humano. De esta forma se intenta analizar el contenido energético de la señal, presente en cada una de las bandas críticas del sistema auditivo.

Se eligen frecuencias linealmente espaciadas en la escala Mel[7], y sus correspondientes frecuencias en Hz son los centros de los filtros del banco (ver Figura 4-3). Por esta razón se le denomina "Banco de filtros Mel".

La base de cada triángulo está comprendida entre las frecuencias centrales de sus filtros adyacentes como se ve en la Figura 4-4. Las alturas de los filtros se determinan manteniendo el área de cada triángulo unitaria.

La DEP es filtrada utilizando este banco, y se integra la energía de cada una de las N_f bandas, obteniéndose así el vector de coeficientes Mel (VCM). Sus componentes representan la información de la DEP presente en cada banda, ponderada por la forma del filtro.

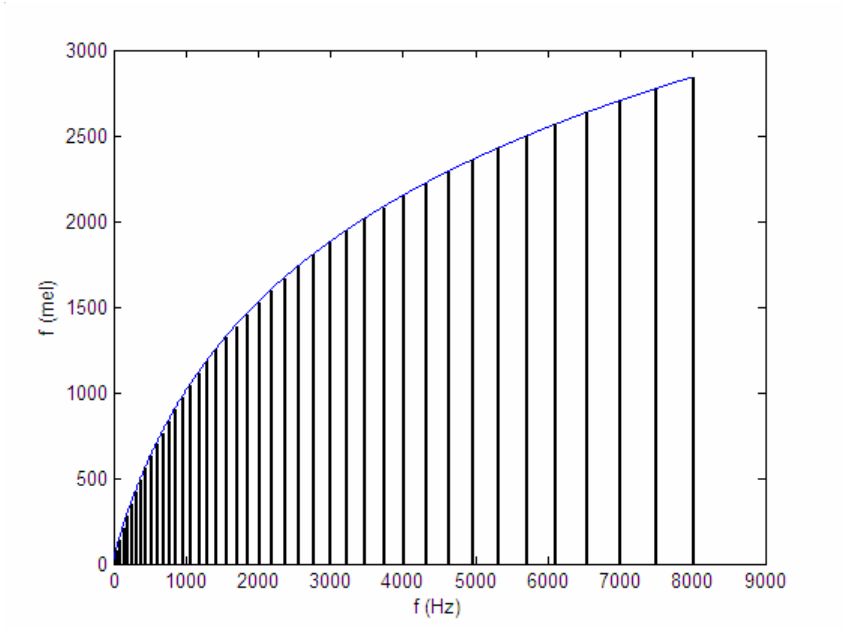


Figura 4-3: Las barras indican las frecuencias de los centros de los filtros triangulares

[7] Ver capítulo 4 "Fundamentos de la música"

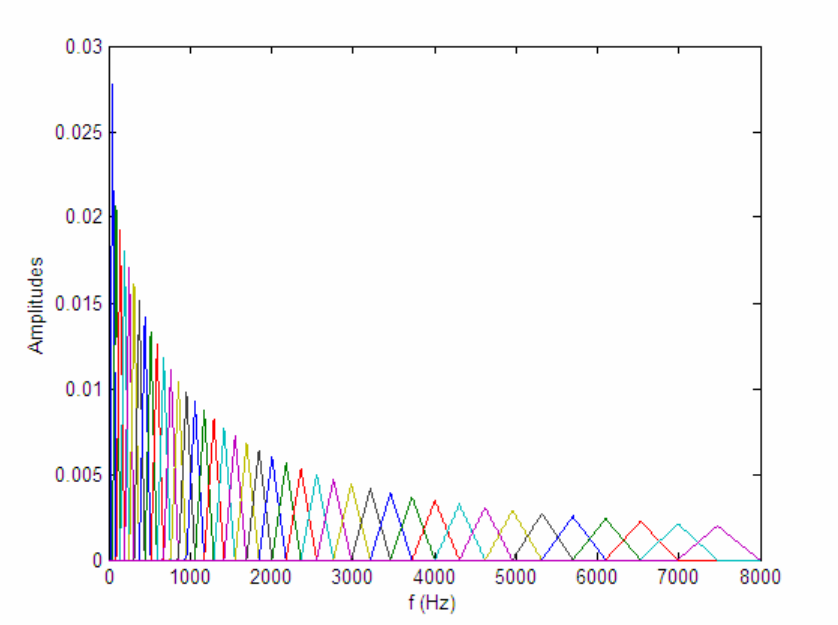


Figura 4-4: Banco de filtros triangulares de MFCC

Coeficientes Cepstrales

Previo a la caracterización cepstral se presentará el mecanismo de producción de la voz para explicar el uso del cepstrum en audio.

Mecanismo de producción de la voz

La voz es producida por el pasaje de aire proveniente de los pulmones excitando las cuerdas vocales ubicadas en la glotis y continuando su recorrido a través del aparato fonador. Ver Figura 4-5.

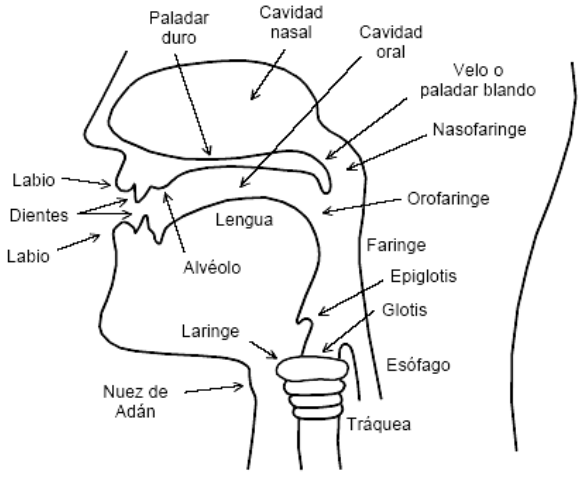


Figura 4-5: Corte esquemático del aparato fonador [13]

Inicialmente el aire hace vibrar las cuerdas vocales, generando ondas periódicas (tonos) o ruido, dependiendo de la posición de la glotis. Las ondas siguen su trayectoria a través de la faringe hacia la cavidad oral y/o nasal, hasta salir por los labios. En ese pasaje el espectro de las ondas cambia de forma, debido a las variantes en el diámetro de la laringe, la posición de la lengua, la mandíbula, los dientes y los labios. Este proceso corresponde a la articulación de los sonidos [11].

Una vez diferenciados los dos aspectos de excitación y articulación, se observa que la articulación está vinculada a la sucesión de sonidos (fonemas) emitidos. Por el contrario, la excitación, y concretamente la fluctuación de la frecuencia fundamental y de la energía, aportan información sobre la expresión del lenguaje, la entonación, el sexo del locutor, el estado emocional, etc.

Ambos procesos se pueden considerar independientes el uno del otro. Por ejemplo, con una misma cadena de palabras y con diferentes formas de expresarlas se puede realizar una afirmación, una interrogación, etc [12].

Modelado del mecanismo de producción de la voz

La producción de la voz se modela entonces como el filtrado de la excitación (señal periódica de las cuerdas vocales o señal de ruido producida por una constricción al paso del aire) por la transferencia del aparato fonador. Esta transferencia se considera un filtro lineal a efectos de simplificar el modelo; se observa en el dominio del tiempo como la convolución entre la señal de excitación y la respuesta impulsiva del aparato fonador.

Cepstrum

Los coeficientes cepstrales se obtienen a partir del Análisis del *Cepstrum* real, que ha probado ser una herramienta muy útil en el estudio del reconocimiento automático de voz. El *cepstrum* se define como la transformada inversa de Fourier (IFT) del logaritmo de la DEP:

$$C(\tau) = IFT(\log(DEP)).$$

Donde τ es una variable en un nuevo dominio del tiempo, llamado dominio cepstral. Éste se mide en unidades de cuefrencias, un anagrama de la palabra frecuencias; de la misma forma se relaciona la palabra cepstrum con espectro (spectrum en inglés).

Dado que la DEP es una función real y par, se puede demostrar que aplicar la IFT es equivalente a aplicarle su transformada directa FT, por lo tanto :

$$C(\tau) = IFT(\log_{10}(DEP)).$$

Los coeficientes cepstrales de frecuencia Mel son una variante de la ecuación anterior. En primer lugar, el logaritmo no se aplica directamente a la DEP, sino a los coeficientes Mel (VCM). La segunda modificación consiste en la utilización de la DCT en lugar de la FT. Esto busca lograr una reducción en el número de coeficientes[8] del cepstrum, resaltando los de más bajas cuefrencias.

$$C(\tau) = DCT(10 * \log_{10}(VCM))$$

El cepstrum es un análisis en frecuencia de la DEP (expresada en decibeles) de la señal de voz. Sirve para analizar dos tipos de periodicidad que aparecen en la DEP:

- Periodicidades rápidas, debidas a la estructura armónica del espectro, que se repiten en los múltiplos de la frecuencia fundamental Fo. Corresponden a la información de la fuente excitadora y se sitúan en la parte alta del cepstrum (altas cuefrencias).

- Fluctuaciones mucho más lentas, no periódicas, que dan la envolvente del espectro. Se manifiestan en la parte baja del cepstrum (bajas cuefrencias) y caracterizan la transferencia del aparato fonador.

Sea *s(n)* la señal de voz muestreada, *u(n)* la señal de excitación y *h(n)* la respuesta al impulso del aparato fonador. El modelo de producción de la voz está dado por la expresión siguiente, siendo * la convolución:

$$s(n) = u(n) * h(n)$$

Aplicando la transformada de Fourier a la expresión anterior, y recordando que la convolución de dos secuencias en el tiempo, implica su producto en frecuencia [1] :

$$S(k) = U(k).H(k)\,,$$

Tomando el módulo de S(k) al cuadrado, se obtiene la DEP:

$$DEP(k) = |S(k)|^2 = |U(k)|^2.|H(k)|^2$$

Ahora calculando el cepstrum,

$$Cep(\tau) = IFT\left(\log\left(|U(k)|^2.|H(k)|^2\right)\right) = IFT\left(2.\log\left(|U(k)|\right)\right) + IFT\left(2.\log\left(|H(k)|\right)\right)$$

Este resultado indica que el cepstrum de una señal de voz está dado por la suma del cepstrum de la excitación y el de la transferencia del aparato fonador. Como se explicó anteriormente, estas señales ocupan partes disjuntas del dominio cepstral, por lo tanto, se pueden obtener sus coeficientes cepstrales separadamente (ver Figura 4-6). Las diferentes informaciones se utilizan respectivamente para estimar la frecuencia fundamental (la altura) y para el reconocimiento automático de voz.

[8] Ver sección 2.3.3

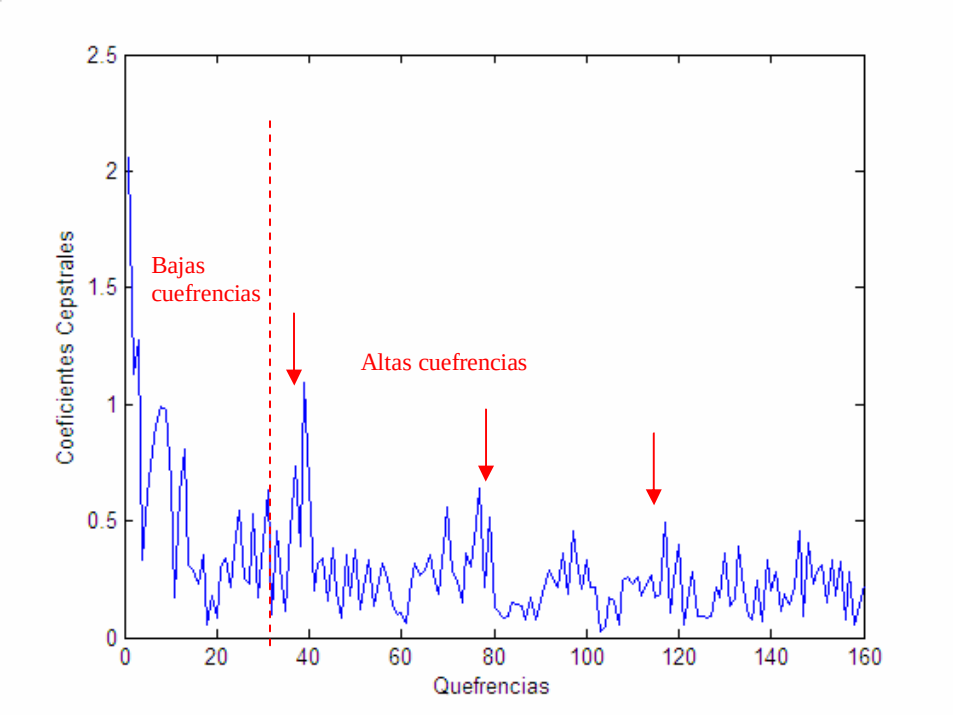


Figura 4-6: Coeficientes Cepstrales de una señal de 160 muestras.

Aplicación del cepstrum a la música polifónica

En la bibliografía referente a este tema, [15][16], se constató que la información del cepstrum más comúnmente utilizada para analizar la música polifónica, es la de la envolvente espectral, o sea, los coeficientes bajos del cepstrum.

Esta envolvente brinda información sobre la distribución y las amplitudes de los armónicos, que como se vio en la sección 3.2.2, son una representación aproximada de la textura tímbrica de la canción.

Como se dijo anteriormente, la aplicación de la DCT enfatiza los coeficientes cepstrales bajos, por lo tanto los *MFCC*s contienen principalmente información referida a la envolvente de los coeficientes Mel, obtenidos luego del filtrado de la DEP.

En la Figura 4-7, se reconstruyeron los coeficientes Mel para un tramo de un segundo de canción a partir de los *MFCC*. Se puede ver como la señal reconstruida aproxima a la envolvente de la señal original.

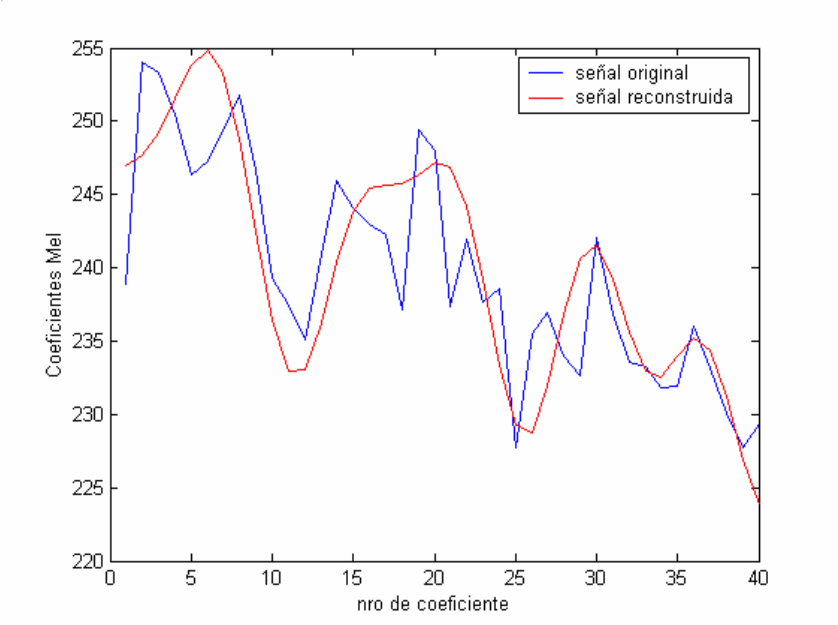


Figura 4-7: Reconstrucción de los coeficientes Mel a partir de los MFCC.

4.3.3 Cálculo de la similitud: medida de distancia

Una vez obtenidos los vectores *MFCC*, es necesario encontrar una medida que refleje la diferencia entre ellos. Una opción es considerar la distancia coseno normalizada:

$$d(v_i, v_j) = \frac{< v_i, v_j >}{\|v_i\| \cdot \|v_j\|}.$$

Otra de las distancias entre vectores que se aplica normalmente es la distancia euclideana, que como se recuerda se define:

$$d(v_i, v_j) = \sqrt{\sum_{kl=1}^{L} \left(v_i(k) - v_j(k)\right)^2}.$$

Si bien pueden utilizarse ambas indistintamente, la distancia coseno es la más utilizadas en los trabajos anteriores sobre *MFCC* [18] [19], por eso fue la elegida.

$$S(i, j) = \frac{< v_i, v_j >}{\|v_i\| \cdot \|v_j\|}.$$

4.3.4 Análisis de parámetros

Para la selección de los parámetros del algoritmo *MFCC* se presenta a continuación un estudio cualitativo de dichos parámetros. En el capítulo 9 e realizará un análisis estadístico de los resultados para elegir la configuración más adecuada.

Los parámetros son los siguientes:

- Tamaño de ventana
- Número de coeficientes *MFCC* : N_{mfcc}
- Uso del primer coeficiente *MFCC*
- Forma de los filtros
- Número de filtros del banco: N_f
- Rango en frecuencias del banco: $f_{f\max}$

Tamaño de ventana

El tamaño de ventana utilizado al analizar señales de audio, depende de la aplicación a realizar. El reconocimiento de palabras, por ejemplo, necesita una resolución temporal que permita reconocer los distintos fonemas correspondientes a las letras que componen la palabra. Por lo tanto es necesario utilizar una ventana de 10ms aproximadamente.

En la presente aplicación, basada en búsqueda de similitudes en música, la resolución temporal debe ser menor que la de la aplicación anterior, ya que con ese nivel de resolución no se podría encontrar similitudes de la duración de un estribillo. Se decidió estudiar entonces, matrices generadas a partir de ventanas de 200, 500, 1000 y 2000ms , solapadas en un 50%. Finalmente como se verá en la evaluación final, se utilizaron ventanas de 1000ms.

Número de coeficientes *MFCC*

El número de coeficientes N_{mfcc} con que se caracteriza el espectro de una ventana, determina con qué precisión se aproxima la DEP. La mayoría de los trabajos consultados utiliza 12 coeficientes.

En la Figura 4-8 se puede ver la reconstrucción de la DEP utilizando tres valores espaciados de N_{mfcc} : 6, 12 y 30. De estas figuras se puede concluir que para el valor 6 la aproximación es demasiado vaga, y para el valor 30, demasiado exacta. En cambio para 12 coeficientes se aprecia la envolvente de forma clara.

Uso del primer coeficiente

El primer coeficiente *MFCC* contiene información referente a la energía de la DEP, por lo que se analizó su influencia en los resultados.

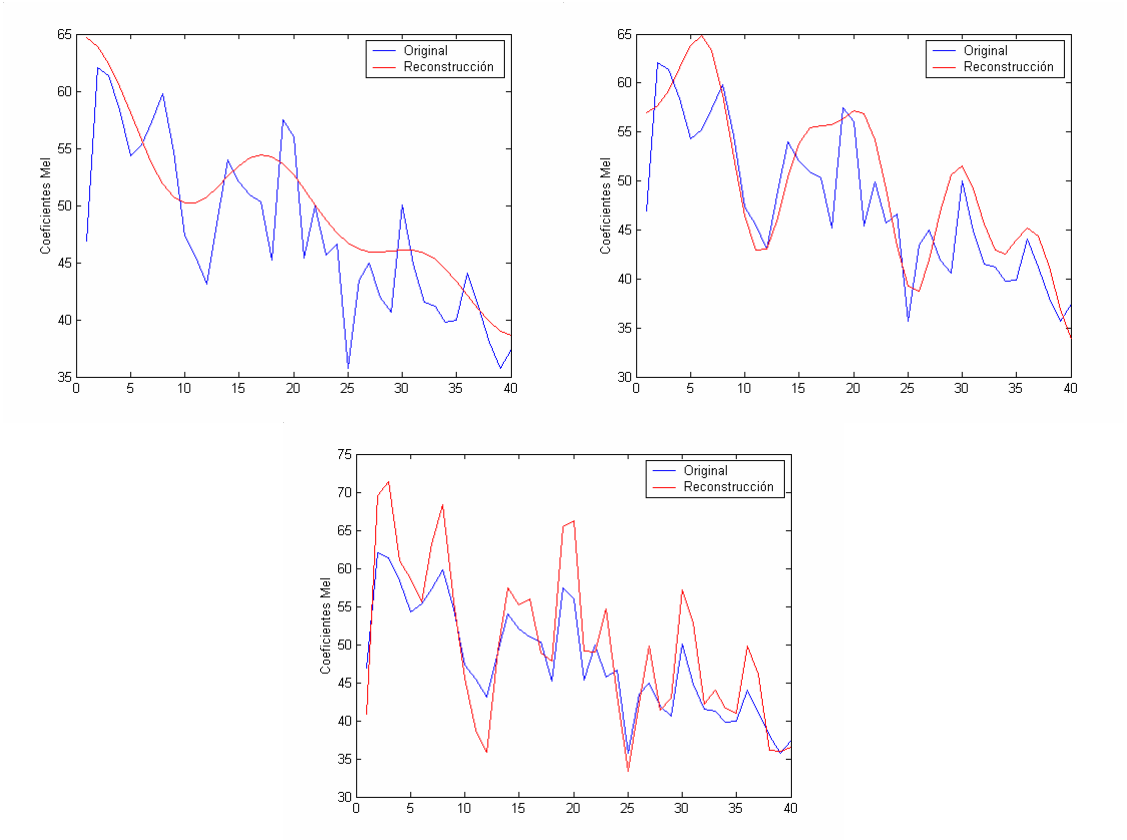


Figura 4-8: Reconstrucción de los coeficientes Mel para 6, 12 y 30 coeficientes.

Forma de los filtros

Este parámetro no fue analizado, se optó por la utilización de un banco de filtros triangular, con solapamiento de acuerdo a la mayoría de los trabajos consultados; específicamente esta definido en [3].

Número de filtros del banco y rango de frecuencias

La resolución espectral depende del número de filtros del banco y del rango de frecuencias que cubre. Por lo tanto se eligió fijar el número de filtros en 40, ya que es el valor más comúnmente utilizado en algoritmos de *MFCC*; variando únicamente el rango de frecuencias.

El rango analizado es un parámetro fundamental, ya que de él depende el tipo de información que estamos conservando. Los rangos se analizan a partir de 20 Hz, por lo tanto se elimina la información referente a la energía de la señal. Se analizaron dos rangos: hasta 3 kHz y hasta 8 kHz. En el primero[9] se encuentra la información más relevante de la voz, cuyo espectro está solapado con las frecuencias fundamentales de algunos instrumentos como ser la guitarra. El segundo se corresponde con el rango completo de frecuencias disponible; al contar con una base de datos muestreada a 16 kHz, se tiene un rango hasta los 8 kHz.

[9] En telefonía se emplea un rango de frecuencia hasta 4kHz. [1]

4.4 Vectores de Croma (VC)

4.4.1 Introducción

En esta sección se expone el método de vectores de croma [8]. Éste fue estudiado e implementado. El desempeño evaluado fue inferior al de *MFCC* y por este motivo se decidió no incluirlo en la solución final.

A diferencia del anterior, se basa en observar la señal como una secuencia de notas, y comparar líneas melódicas.

4.4.2 Procedimiento

Primero se realiza el enventanado con ventanas de Hanning a bloques de N muestras.
A cada bloque se le extrae un vector de características $v_{CV}(c)$, de 12 componentes llamado Vector de Croma. Como se vio en el capítulo 3 "Fundamentos de la música", existen 12 semitonos para cada octava. El croma *c*, corresponde al nombre de la nota, mientras que la altura *h*, es el número de octava (ver Figura 4-10).

El procedimiento para la extracción de características de cada bloque es el siguiente:

- DFT: cálculo de la transformada discreta de Fourier (DFT).
- DEP: estimación de la Densidad Espectral de Potencia, $DEP(k) = \left|DFT(v(n))\right|^2$.
- Vector de Croma: los coeficientes corresponden a los valores resultantes del filtrado de la DEP mediante un banco de filtros cromáticos[10], y la posterior integración del resultado para cada croma. No se tiene en cuenta de esta forma la octava en la que se encuentran los cromas.

$$v_{CV}(c) = \sum_{h=Oct_L}^{Oct_H} \sum_{k=1}^{N} BPF_{c,h}(k) DEP(k)$$

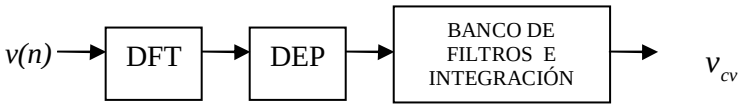


Figura 4-9: Diagrama del procedimiento de extracción de características mediante VC para cada bloque de muestras $v(n)$

[10] Se denomina banco de filtros cromático, porque cada filtro se corresponde con un semitono de las octavas analizadas.

NOTA	Frecuencia [Hz]					
	Octava 2	Octava 3	Octava 4 (central)	Octava 5	Octava 6	Octava 7
DO	65,41	130,81	261,63	523,25	1046,50	2093,00
DO#	69,30	138,59	277,18	554,36	1108,73	2217,46
RE	73,42	146,83	293,66	587,33	1174,66	2349,32
RE#	77,78	155,56	311,13	622,25	1244,51	2489,02
MI	82,41	164,81	329,63	659,26	1318,51	2637,02
FA	87,31	174,61	349,23	698,45	1396,91	2793,83
FA#	92,50	185,00	369,99	739,99	1479,98	2959,96
SOL	98,00	196,00	392,00	783,99	1567,98	3135,96
SOL#	103,83	207,65	415,30	830,61	1661,22	3322,44
LA	110,00	220,00	**440,00**	880,00	1760,00	3520,00
LA#	116,54	233,08	466,16	932,33	1864,66	3729,31
SI	123,47	246,94	493,88	987,77	1975,53	3951,07

Figura 4-10: Frecuencias de los semitonos de las octavas 2 a 7.

A continuación se explican en detalle los pasos más relevantes del procedimiento de extracción.

Banco de filtros cromático

El banco de filtros está compuesto por las bandas de frecuencia de todos los semitonos existentes entre las octavas Oct_L y Oct_H .

Los semitonos no son lineales con la frecuencia, por lo que para calcular los centros de los filtros, se utiliza la escala logarítmica Cent[11].

Se localizan los valores en cents de cada semitono, o sea de cada combinación *(c, h)* :

$$F_{c,h} = 1200h + 100(c-1)$$

A partir de estos valores se buscan las f_{cent} tales que $F_{c,h} - 100 < f_{cent} < F_{c,h} + 100$, y para las correspondientes frecuencias en *Hz* se calcula la forma del filtro:

$$BPF_{c,h}(f_{Hz}) = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi(fcent - (F_{c,h} - 100))}{200}\right)\right)$$

La forma elegida para realizar este filtro es la de la ventana de Hanning; como se ve en la Figura 4-12.

[11] Ver capítulo 4: "Fundamentos de la música"

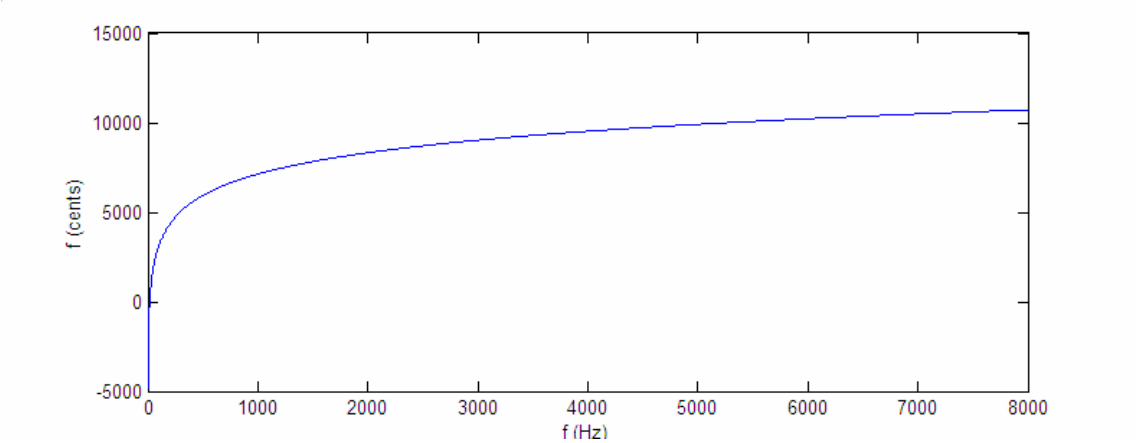


Figura 4-11: Cents en función de Hz.

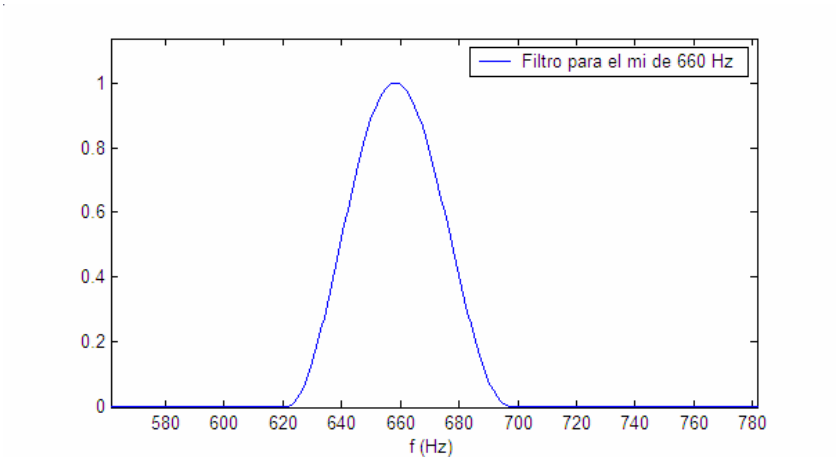


Figura 4-12: Filtro para el mi de 660 *Hz*.

En la Figura 4-13 se ve un ejemplo del filtrado de una ventana para el croma *mi*.

4.4.3 Cálculo de la similitud: medida de distancia

Obtenidos los VC, debe precisarse una medida de similitud entre los mismos; en [8] se define de la siguiente forma:

$$S(i,j) = 1 - \frac{\left| \frac{v_i}{\max_c(v_i)} - \frac{v_j}{\max_c(v_j)} \right|}{\sqrt{12}}.$$

La definición del segundo término se justifica por la norma común de vectores (con coeficientes normalizados). La aplicación del denominador $\sqrt{12}$ para normalizar la diferencia, es la longitud de la diagonal de un hipercubo de 12 dimensiones (12 cromas).

Al contrario de la distancia coseno esta es cero cuando los vectores son similares, por esto se toma la diferencia con respecto a uno, de modo de trabajar en la matriz de la misma forma.

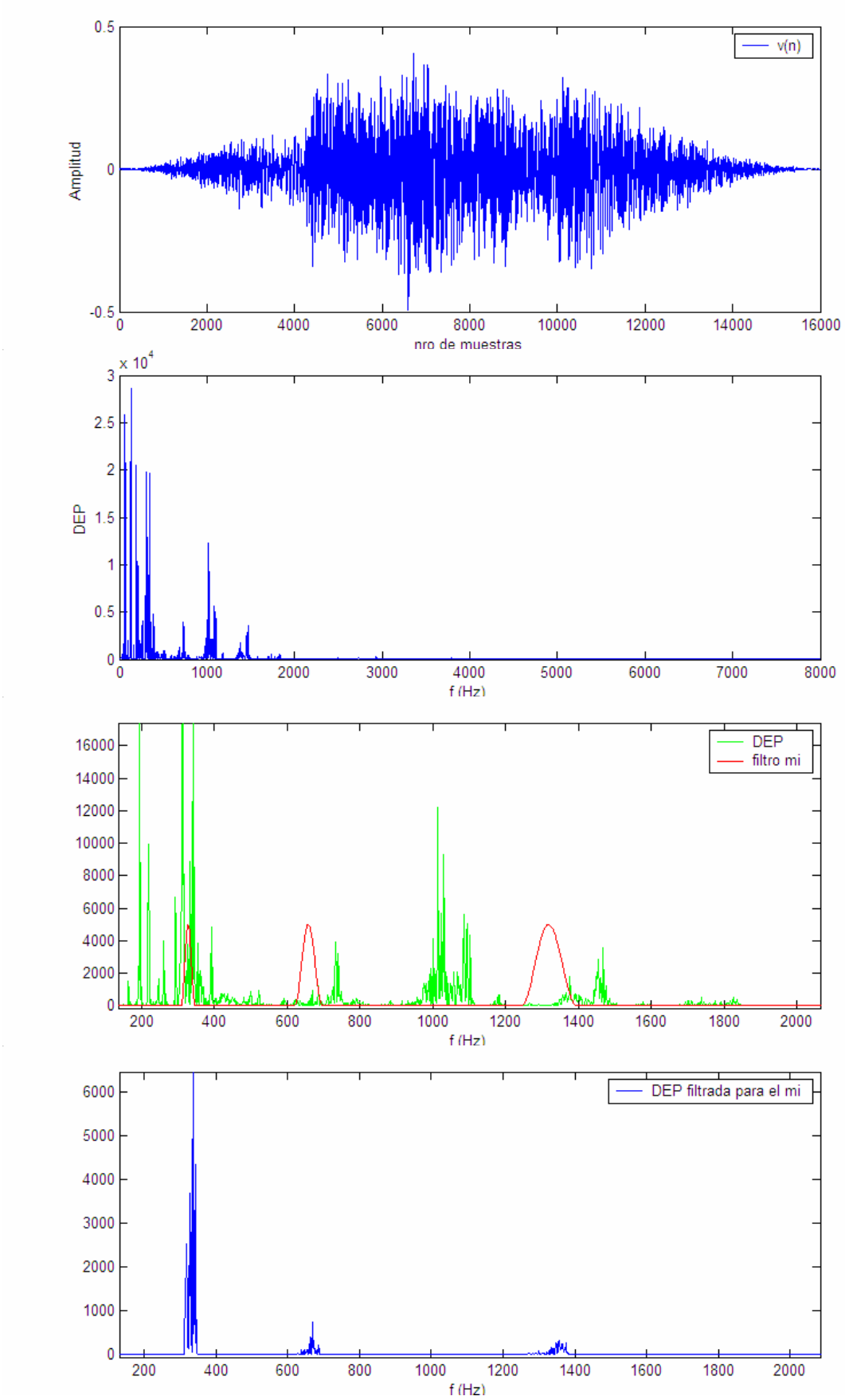


Figura 4-13: Procedimiento realizado a una ventana con los filtros del semitono *mi.*

4.4.4 Análisis de parámetros

Los parámetros para este método son los siguientes:

- Tamaño de ventana
- Forma de los filtros
- Rango en frecuencias a analizar (cantidad de octavas)

El análisis efectuado no difiere con lo dicho antes para *MFCC*, es decir, se estudiaron de la misma forma las variaciones del tamaño de ventana y rango de frecuencia.

4.5 Transformada CQT

4.5.1 Introducción

En esta sección se explica el tercer método de extracción de características que se estudió, pero no fue implementado. A diferencia de los métodos anteriores, en este no se aplica directamente la DFT.

La distancia que se utiliza para comparar los vectores de características de las ventanas es novedosa porque suprime la información referente al timbre, y compara únicamente según la melodía. Por ello, en teoría, este método es capaz de detectar repeticiones de un estribillo tocadas con distintos instrumentos.

4.5.2 Extracción de vectores de características

Primero se realiza el enventanado de la señal como en los métodos anteriores. Luego para cada ventana, se utiliza un banco de filtros para extraer los coeficientes de 36 semitonos distintos, cubriendo 3 octavas.

Como se observa en la siguiente ecuación, este método no realiza la DFT. A continuación se explicarán los parámetros que aparecen:

$$X(k) = \frac{1}{N_k} \sum_{n=0}^{N_k - 1} x(n) e^{\frac{-j\,2\pi Q n}{N_k}}$$

$X(k)$ representa la energía espectral de la k-ésima nota, centrada en la frecuencia f_k para una ventana dada [6].

$$f_k = f_0 \cdot 2^{k/12}, \quad k = 0,1,2,...,35$$

Se elige como frecuencia mínima $f_0 = 130.8Hz$, que representa al *DO3*. Esta elección se basa en la hipótesis de que en la música analizada la mayoría de las notas tienen frecuencias mayores a ésta.

Q es una constante que relaciona la resolución temporal con la frecuencia:

$$Q = \frac{f_k}{f_{k+1} - f_k} = \frac{1}{2^{1/12} - 1}$$

Para cada filtro k, se toma un ancho de ventana $N_k = \frac{f_s Q}{f_k}$, donde f_s denota la frecuencia de muestreo. El utilizar distintos anchos de ventana para representar los semitonos implica una resolución fina en frecuencia ya que se adapta el tiempo de análisis a las frecuencias en las que se está trabajando.

4.5.3 Medida de distancia

Esta medida de distancia pretende calcular la similitud de la melodía sin considerar diferencias en el timbre. Las características del timbre de una nota están representadas generalmente por la energía espectral de los armónicos, que están comprendidos en el vector de características.

Melodía vs. Timbre

Si dos instrumentos tocan la misma nota, los sonidos tendrán la misma frecuencia fundamental, pero su timbre será diferente. La distancia coseno determinará que las notas son lejanas, ya que considera el valor absoluto de la diferencia.

Para discriminar la diferencia por melodía de la diferencia por timbre, se examina el vector diferencia entre dos notas, definido como:

$$\Delta V = V_1 - V_2 = \left[\left| v_{11} - v_{21} \right|, ..., \left| v_{1N} - v_{2N} \right| \right]$$

Donde V_1 y V_2 son los vectores de características de dos notas y N es la dimensión del vector (en este caso, 36).

Los armónicos para una frecuencia fundamental f, se encuentran en 2f, 3f, 4f, etc. Como se vio en la sección 3.3.1, 2f se encuentra a 12 semitonos de f, 3f a 7 semitonos de 2f, y 4f a 4 semitonos de 3f.

$$2 \cdot f = f \cdot 2^{12/12}$$
$$3 \cdot f \cong 2 \cdot f \cdot 2^{7/12}$$
$$4 \cdot f \cong 3 \cdot f \cdot 2^{4/12}$$

En la Figura 4-14se puede ver un ejemplo de esto. En el primer caso se compara la misma nota tocada con dos instrumentos diferentes y las componentes que prevalecen son las que se distancian 12, 7 o 4 semitonos, o sea los valores de los armónicos. A estos apartamientos entre componentes se les llamará apartamientos armónicos.

En el segundo caso se comparan distintas notas con el mismo instrumento y las distancias entre componentes no son de esos valores. El valor absoluto del vector diferencia en los dos casos será similar a pesar de que en el primer caso la nota es la misma y en el segundo caso no.

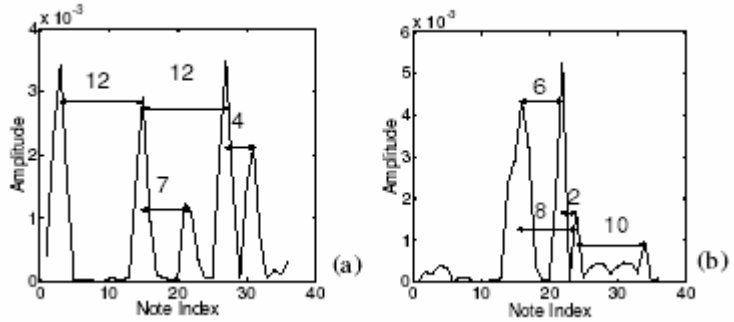


Figura 4-14: Vector diferencia entre: a) D3 tocado por cello y alto-trombón; b) D3 y D#3 tocado por cello

Estructura del vector diferencia

Del ejemplo anterior se concluye que para que se pueda diferenciar los tonos independientemente del timbre, hay que considerar la estructura del vector diferencia y no solamente su norma.

Lo que se propone para solucionar este problema es adjudicar distintos pesos al vector diferencia. Esto quiere decir que si las componentes significativas del vector están apartadas intervalos armónicos, la distancia debería ser chica, y si no es así, debería ser grande.

Para describir la estructura del vector ΔV, se considera su auto-correlación:

$$r(m) = \sum_{n=0}^{N-m-1} \Delta v_{n+m} \Delta v_n, \quad 0 \le m \le N-1$$

Donde Δv_i es la i-ésima componente de ΔV, y m es el apartamiento entre componentes. $r(m)$ es el coeficiente de auto-correlación y representa el grado en que las notas están distanciadas m. Por ejemplo, los coeficientes con apartamientos armónicos como $r(12)$ o $r(7)$ representan la posibilidad de que los dos sonidos sean la misma nota, por lo tanto deben estar suprimidos en la medida de distancia.

Se forma un vector $R = [r(0), r(1), ..., r(N-1)]^T$ con los coeficientes de auto-correlación. Para reflejar la contribución de las distintas componentes se le adjudican distintos pesos a cada una de ellas. Para esto se crea un vector de pesos $W = [w(0), w(1), ..., w(N-1)]^T$, y se halla la distancia multiplicando los dos vectores:

$$d_{ij} = W^T R_{ij}$$

El vector de pesos *W* debe cumplir que las posiciones de apartamientos armónicos ($r(12)$ y $r(7)$ por ejemplo) estén multiplicadas por valores pequeños en comparación con el resto.

4.6 Comparación cualitativa de los métodos de extracción de características

Para analizar cualitativamente los métodos anteriores se enumerarán las características más importantes de cada uno:

- **MFCC**: Compara parejas de fragmentos de un segundo de duración evaluando la similitud en el timbre. Las bandas de frecuencia que analiza están espaciadas según la escala Mel que es una representación de la percepción humana de la altura. Dichas bandas cubren las frecuencias de 20Hz a 3kHz. El vector de características se compone de 12 coeficientes.

- **VC**: Compara parejas de fragmentos de un segundo de duración, pero a diferencia de *MFCC*, analiza el contenido melódico (semitonos) en lugar del timbre. Las bandas de frecuencia analizadas son las correspondientes a seis octavas de la escala cromática (130 Hz a 4 kHz). Se consideran los componentes de cada croma (12 coeficientes) sin distinguir entre las distintas octavas.

- **CQT**: Analiza las bandas de la escala cromática distinguiendo los semitonos entre tres octavas; de esta forma, el vector de características está compuesto por 36 elementos cubriéndose las frecuencias de 130 Hz a 1kHz. Se representan los semitonos de una forma más exacta que en VC ya que se utilizan distintas resoluciones temporales para analizar distintas frecuencias, por los que tiene una resolución más fina en frecuencia

Como se explicó anteriormente, la caracterización del espectro en *MFCC* se realiza "vagamente" por la envolvente de los armónicos, y no por los componentes armónicos mismos. Por lo tanto al comparar los *MFCC* de distintos fragmentos, se buscará más la similitud entre timbres que entre altura.

En CQT y VC, el contenido en frecuencia analizado es más específico que en *MFCC*, ya que se analizan las componentes correspondientes a cada semitono. Esto significa que analizando el mismo fragmento de canción, *MFCC* podría encontrar similitud cuando los otros dos no.

En la Figura 4-15 se encuentran graficados, coeficientes *MFCC* para fragmentos de un segundo, de una canción con cuatro repeticiones del estribillo. Los distintos colores se corresponden a cada una de las repeticiones, y los cuadros muestran su evolución sincronizada en el tiempo cada dos segundos

Al sincronizar las repeticiones se puede ver que los *MFCC* calculados son muy similares, adicionalmente difieren entre cuadros, verificándose que en cada cuadro, las texturas tímbricas se mantienen constantes.

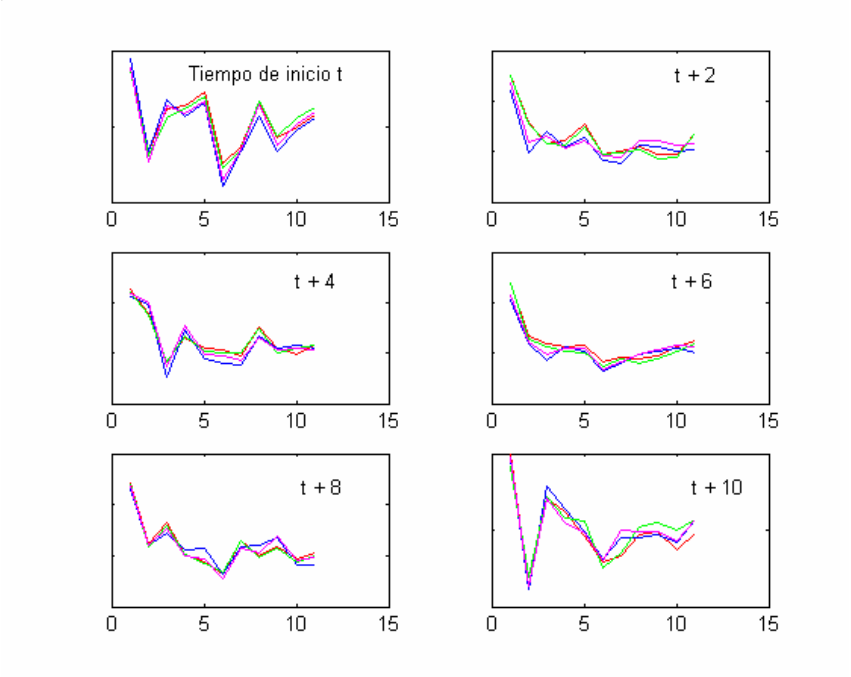


Figura 4-15: Coeficientes MFCC para 6 instantes de tiempo de 4 repeticiones sincronizadas.

En la Figura 4-16 se graficaron los coeficientes del VC para la misma canción y con el mismo sincronismo que en el ejemplo anterior. La similitud en base a líneas melódicas también parecería ser un buen método, pero a grandes rasgos, los cuadros parecen ser menos similares que para *MFCC*, confirmando lo que se explicaba anteriormente.

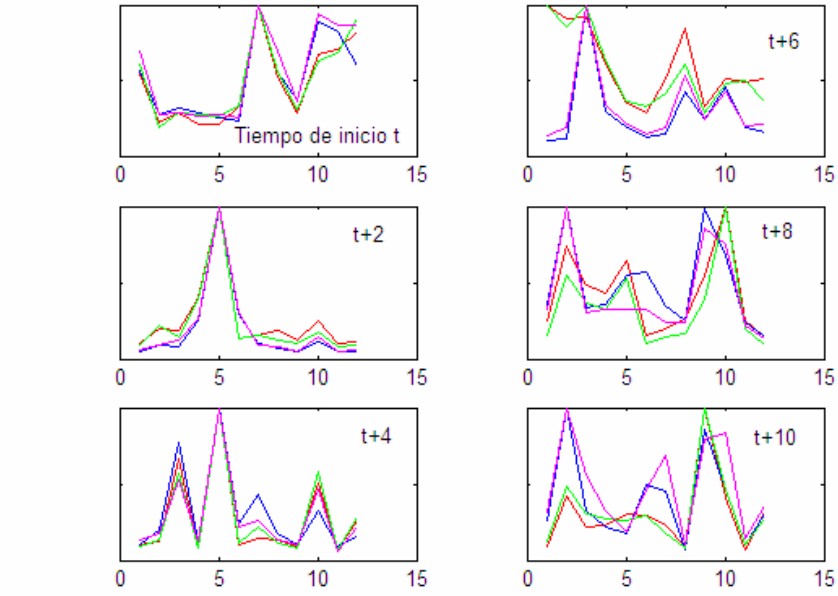


Figura 4-16: Coeficientes del VC para 6 instantes de tiempo de 4 repeticiones sincronizados.

5 Matriz de Similitud

5.1 Resumen

En este capítulo se explicará qué es y cómo se analiza la matriz de similitud, el elemento fundamental en la identificación del resumen. En ella se almacena la información de la canción necesaria para realizar la detección.

5.2 Introducción

El propósito de la matriz de similitud es visualizar la comparación de dos secuencias de datos. Cada secuencia a comparar se representa en el eje vertical y horizontal respectivamente.

Cada elemento de la matriz representa la comparación entre dos fragmentos de una canción. Los valores corresponden a la distancia entre ciertas características extraídas de ambos fragmentos según alguna métrica elegida.

En este caso se utiliza la matriz de similitud propia, que en adelante llamaremos *S*. Ésta compara una secuencia de datos consigo misma. La secuencia utilizada es la conformada por las ventanas en que se divide la canción, y se comparan los vectores de características de cada ventana antes extraídos.

5.3 ¿Cómo "leer" la matriz?

S es una matriz cuadrada y simétrica respecto a su diagonal, donde el valor de *S(i,j)* es proporcional a la similitud entre los elementos *i* y *j*.

Los valores obtenidos son exhibidos como representación en dos dimensiones de la canción, en una imagen en escala de grises, con valores entre 0 (negro) y 1 (blanco)[12]. Los píxeles blancos corresponden a intervalos idénticos.

La esquina izquierda superior de la imagen corresponde al principio de la canción, mientras que la derecha inferior al final.

El análisis de la matriz de similitud propia, permite visualizar la estructura de la canción en el tiempo, ya que los intervalos de canción similares aparecen como líneas diagonales claras. La diagonal principal siempre es blanca, ya que representa la comparación de cada instante consigo mismo.

Para extraer las características se utilizaron ventanas de un segundo, solapadas en un 50%. Queda definido entonces, un espaciamiento temporal entre ventanas de medio segundo, por lo que esta será la resolución de la imagen.

[12] Se elige arbitrariamente esta configuración, pero se podría utilizar el 0 para el blanco y el 1 para el negro.

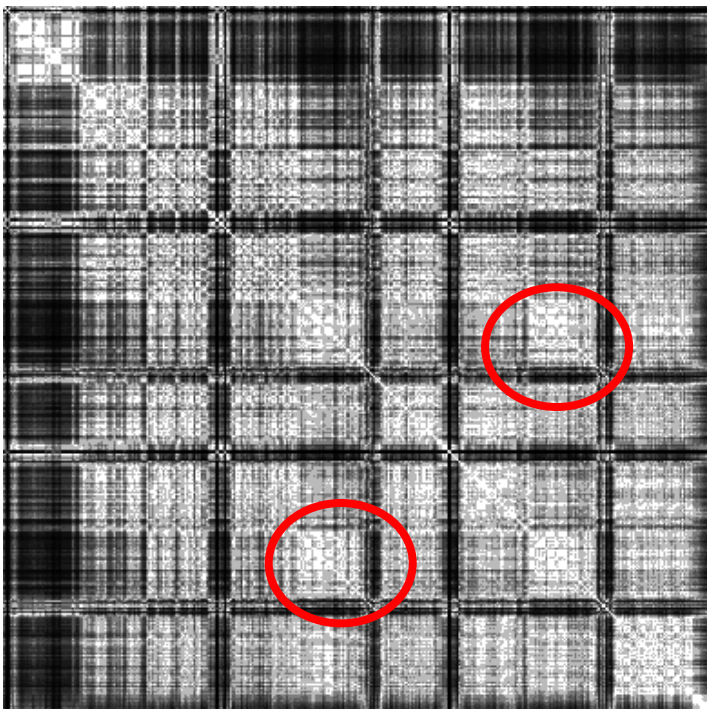


Figura 5-1: Matriz de Similitud Propia S de la canción Tearjerker de Red Hot Chilli Peppers

Además de las diagonales claras se pueden observar cuadrados claros, esto se debe a que en la canción existen intervalos más homogéneos musicalmente que otros.

5.4 Representación "time-lag": Matriz T

Esta representación se utiliza en el método de identificación de estribillo (IE). Consiste en mapear la matriz S en otra matriz T calculada a partir de la rotación del triángulo inferior izquierdo de la matriz S. De esa forma se buscan líneas horizontales en lugar de diagonales y el costo de procesamiento se reduce [6].

En el capítulo siguiente se explicarán en detalle los procesos que se le realizan a la matriz S para obtener la matriz T, que luego se continuará procesando para identificar el estribillo y sus repeticiones.

El eje horizontal de T representa el número de ventana (tiempo) y el vertical es el lag, o sea el tiempo que transcurre desde el tiempo i hasta i+lag. El lag de las líneas horizontales detectadas representa el tiempo tras el cual ocurrirá la repetición.

$$T(i, lag) = S(i, i + lag)$$

En la matriz *T* se visualizan únicamente las líneas que aparecen en el triángulo inferior de *S*. Esto no implica falta de información, porque la matriz es simétrica y estos tiempos se encuentran al sumarle los valores de lag a las líneas de *T*.

Las imágenes se trabajaron con la configuración de nivel de grises en que la máxima similitud corresponde al blanco. De todas formas algunas imágenes se presentarán con la configuración contraria, para obtener una mejor visualización.

En la Figura 5-2 se ilustra con un ejemplo la matriz T. En la matriz original se observan dos diagonales a los lados de la diagonal principal, que se corresponden con las repeticiones del estribillo. Como se indica ambas incluyen la misma información, es decir la primera aparición del estribillo entre t_1 y t_2 y su repetición entre $t_1 + lag$ y $t_2 + lag$. Para la construcción de la matriz T se mapea únicamente la diagonal señalada en rojo (ubicada en el triángulo inferior).

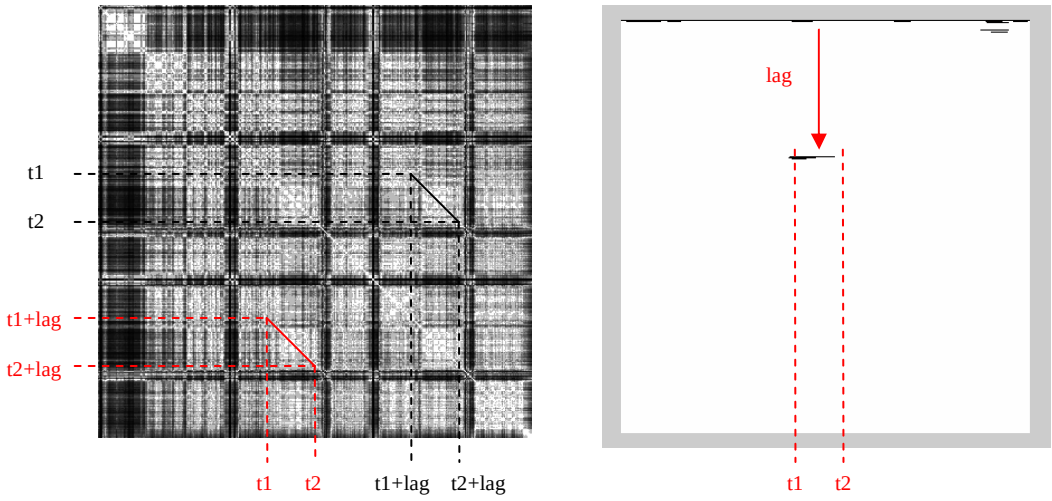


Figura 5-2: Representación time-lag

5.5 Significado de la posición y la cantidad de líneas en la matriz T

En la matriz T aparecen columnas de líneas que representan el número de repeticiones de cada estribillo.

En la Figura 5-3 se puede ver una primera columna con tres líneas. Esas líneas son las comparaciones de la primera repetición del estribillo con las tres restantes. En la segunda columna se ve la comparación de la segunda repetición con la tercera y cuarta. La última columna representa la comparación de la tercer y cuarta repetición.

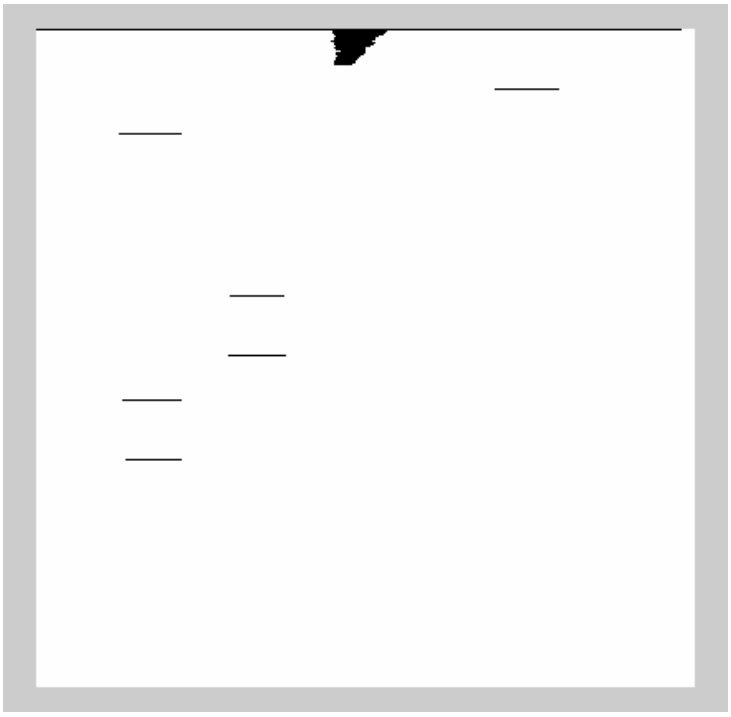


Figura 5-3: Matriz T con cuatro repeticiones.

En las siguientes figuras se observan matrices T para canciones con dos y tres repeticiones del estribillo. El análisis es análogo al de la Figura 5-3.

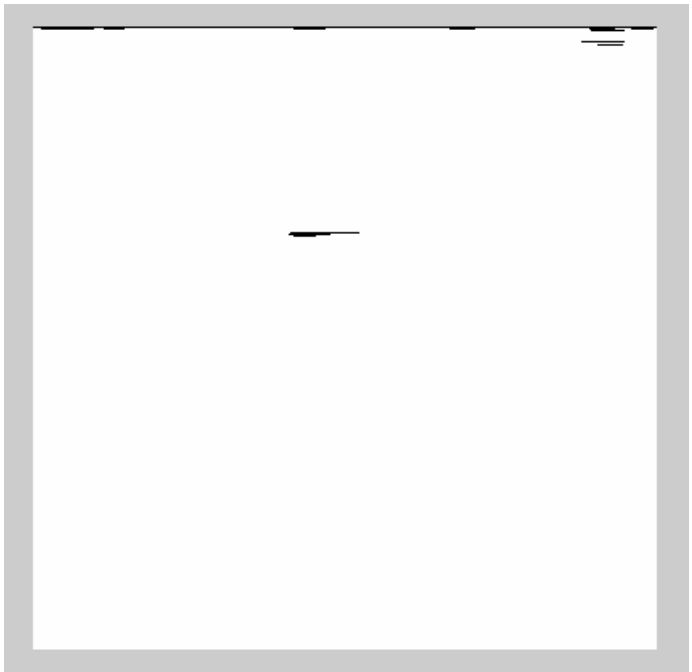


Figura 5-4: Matriz T con dos repeticiones.

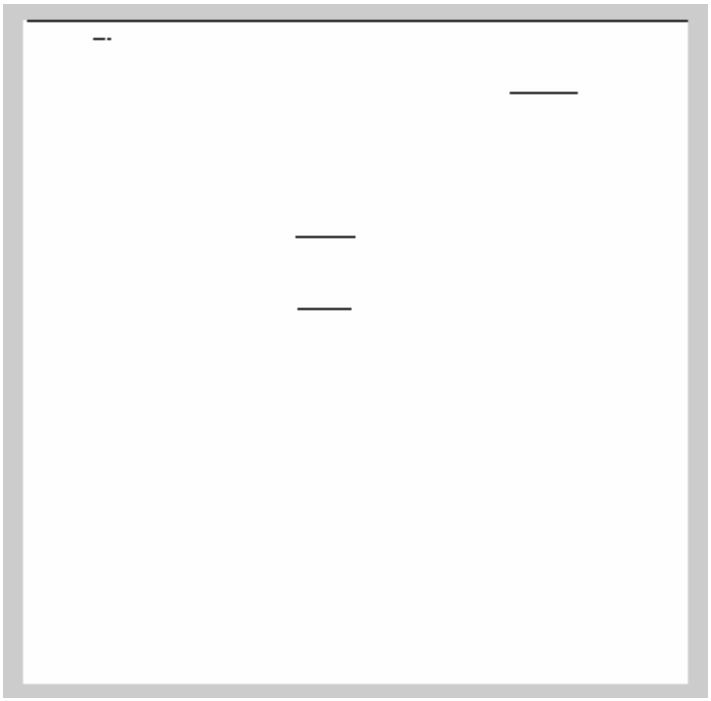


Figura 5-5: Matriz T con tres repeticiones.

6 Identificación de estribillo (IE)

6.1 Resumen

En este capítulo se presenta el método de análisis principal, que busca en la Matriz de Similitud las secciones de canción que se repiten.

6.2 Introducción

El procedimiento definido consta de dos etapas. En la primera se procesa la imagen para conservar únicamente las líneas que corresponden a las repeticiones y eliminar la información que no es necesaria. En la segunda etapa, se analiza la imagen resultante para identificar el comienzo y fin de los fragmentos candidatos a estribillo.

Este procedimiento se realiza de forma iterativa, variando ciertos parámetros y almacenando todos los resultados. Finalmente se selecciona la configuración óptima.

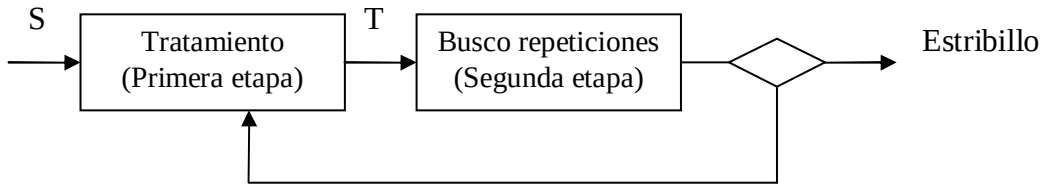


Figura 6-1: Esquema básico de la identificación del estribillo

6.3 Primera etapa: Tratamiento de la imagen

Se parte de la matriz S, donde la resolución corresponde a medio segundo de canción, y su tamaño varía según el largo de la canción. Para el procesamiento se trabaja con la representación time-lag, por lo que se intentará identificar las líneas horizontales presentes en la matriz T.

El proceso que define el resultado de esta etapa es la convolución[13] de la imagen con un elemento estructurante horizontal. El largo óptimo del elemento estructurante varía de canción en canción. Se determinó empíricamente un rango de valores en los que efectivamente se detectaban repeticiones. Este rango comprende los valores de horizontal desde 8 hasta 47.

La convolución realizada implica una limpieza de la imagen (erosión), donde se resaltan las áreas de la misma que se asemejen al elemento estructurante horizontal.

[13] Ver Apéndice B

En la Figura 6-2 se presenta el procedimiento general de identificación. La primera y la segunda etapa se repiten variando el valor de la horizontal y evaluando si la detección fue exitosa o no. En las posteriores secciones de este capítulo se explicará en detalle los procesos realizados.

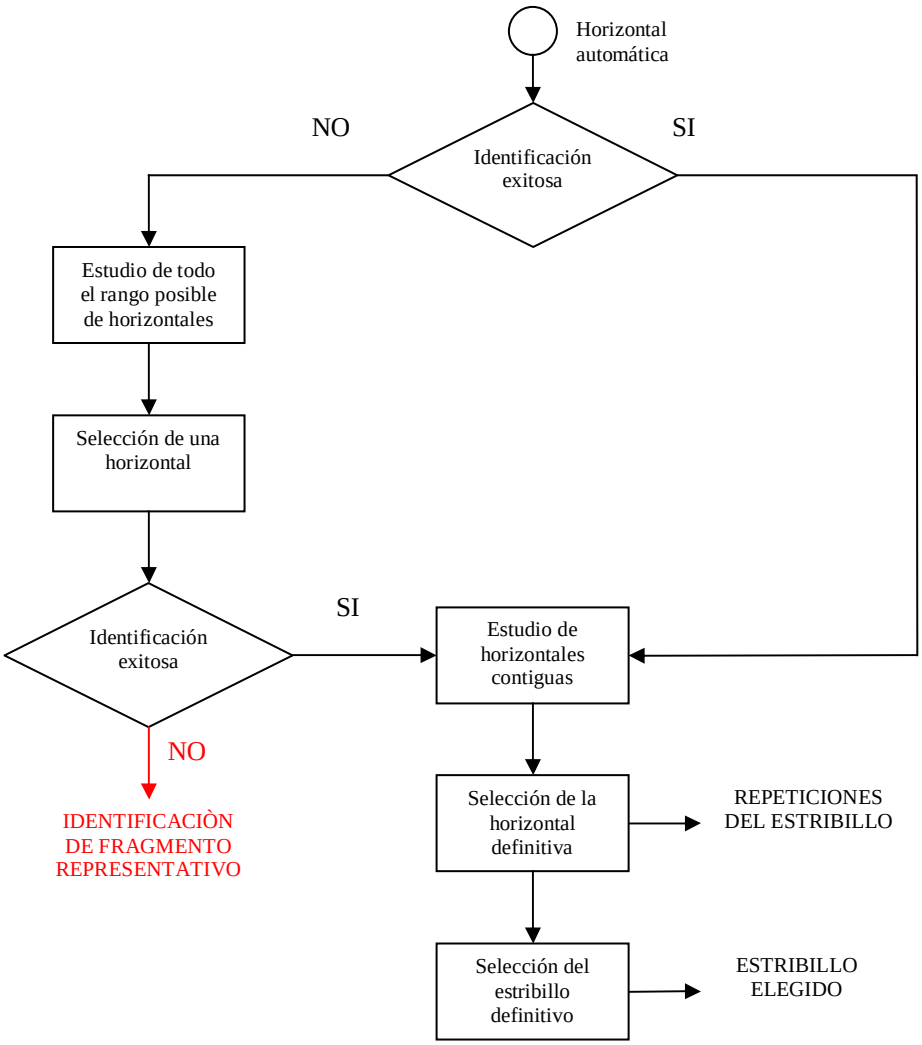


Figura 6-2: Procedimiento general de identificación del estribillo y sus repeticiones

En la Figura 6-3 se puede ver el principio y el final de esta etapa.

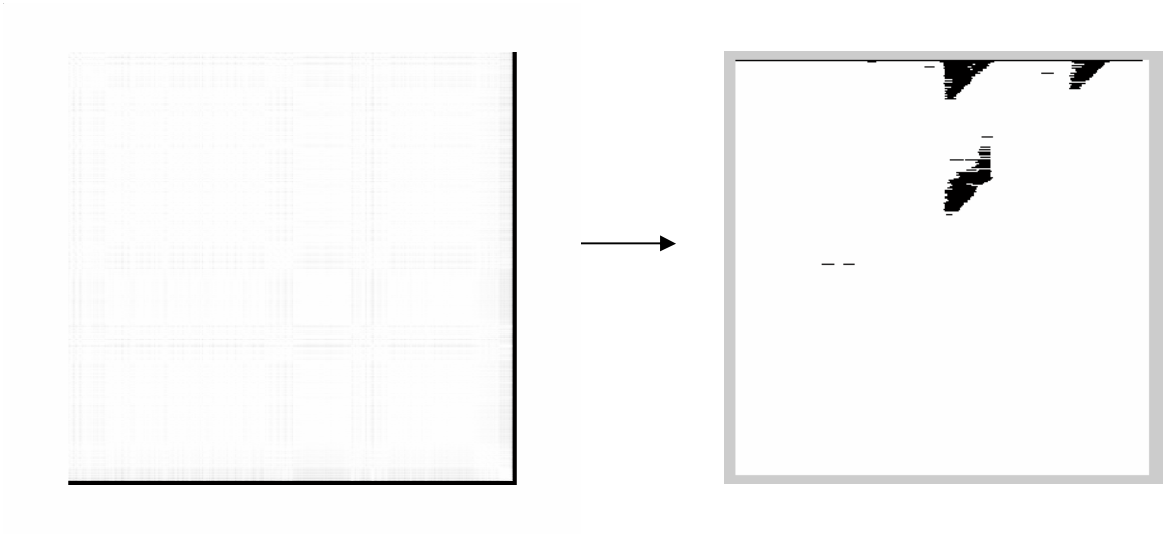


Figura 6-3: Matriz de similitud original S y su representaciòn time-lag T procesada.

El procesamiento de la imagen es sencillo y se ilustra en la Figura 6-4. En primer lugar, se rota la imagen S ecualizada, Seq, para obtener su representación time-lag T. La imagen obtenida es convolucionada con el elemento estructurante, para un determinado valor de horizontal. El resultado de la convolución se binariza.

Se suman la imagen binarizada y la imagen T. Esta operación busca enfatizar las horizontales. Finalmente, se vuelve a realizar la convolución con el elemento estructurante y se binariza la imagen. Las binarizaciones se realizan con un umbral igual a cero.

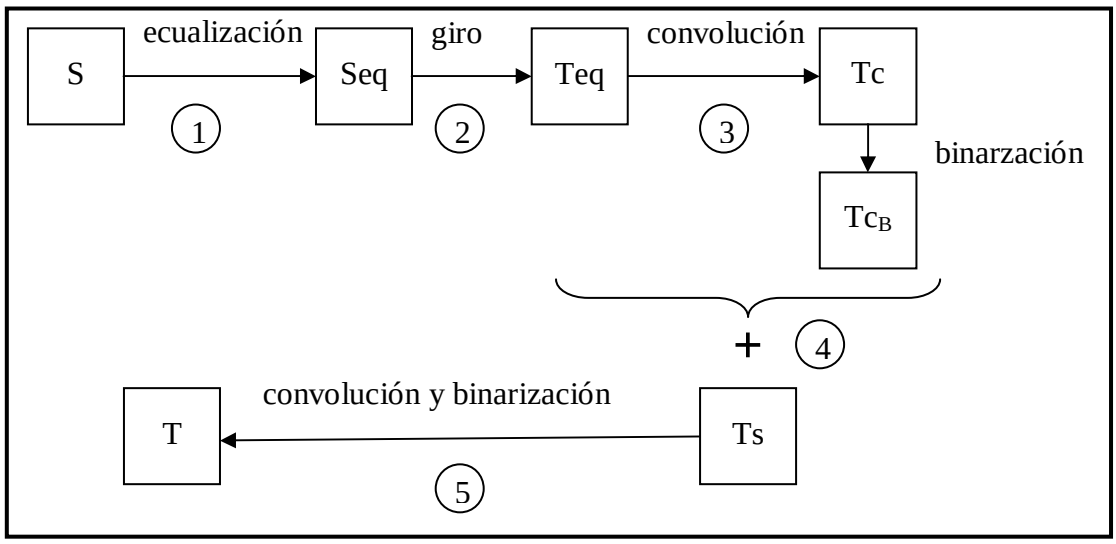


Figura 6-4: Primera etapa

A continuación se muestran las imágenes resultantes de cada proceso del procedimiento.

1. Ecualización de la matriz S: Seq (Figura 6-5).

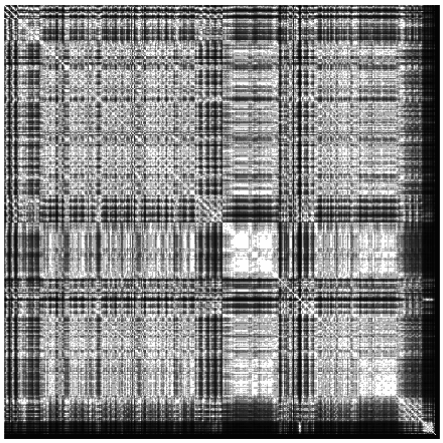


Figura 6-5: Seq

2. Representación time-lag de Seq[14]: Teq (Figura 6-6).

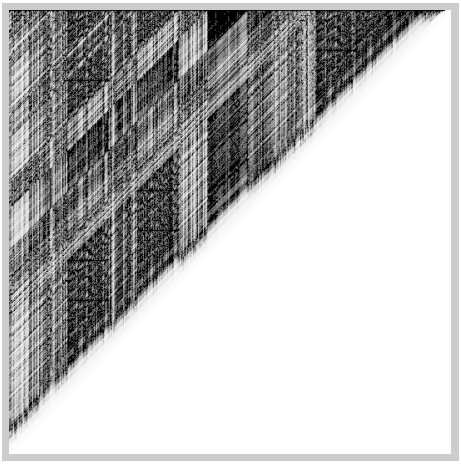


Figura 6-6: Teq

3. Convolución de Teq con el elemento estructurante de largo 12, h12: Tc (Figura 6-7).

[14] A partir de este punto se comenzará a visualizar las imágenes de tal forma que el negro representa similitud.

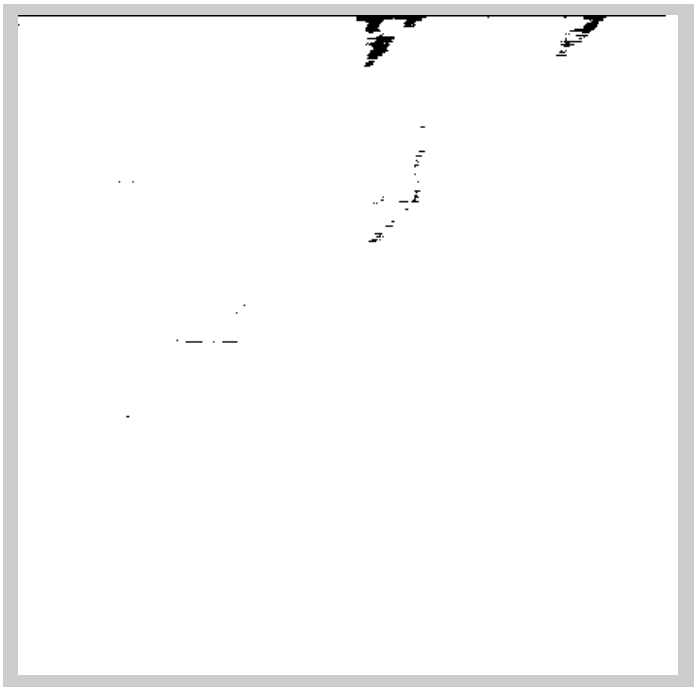


Figura 6-7: Tc

4. Suma de Teq con Tc binarizada: Ts (Figura 6-8).

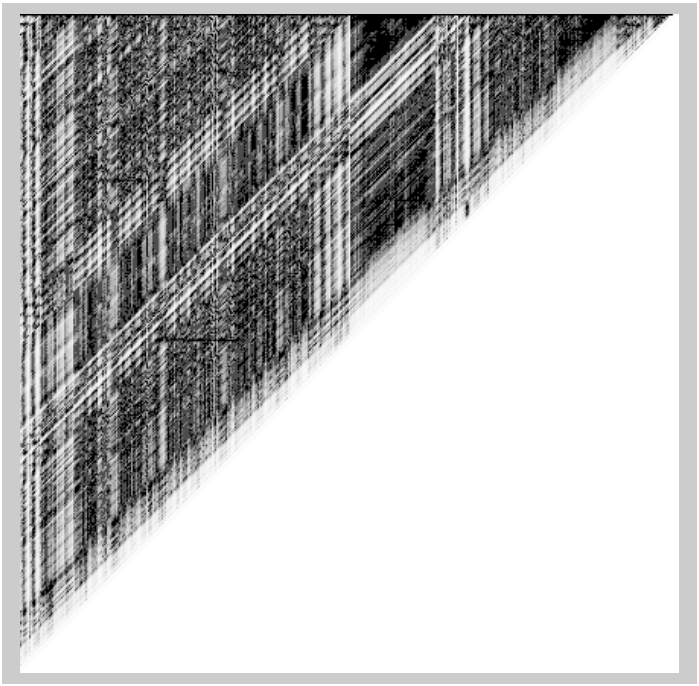


Figura 6-8: Ts

5. Nueva convolución de Ts con h12 y binarización: T (Figura 6-9).

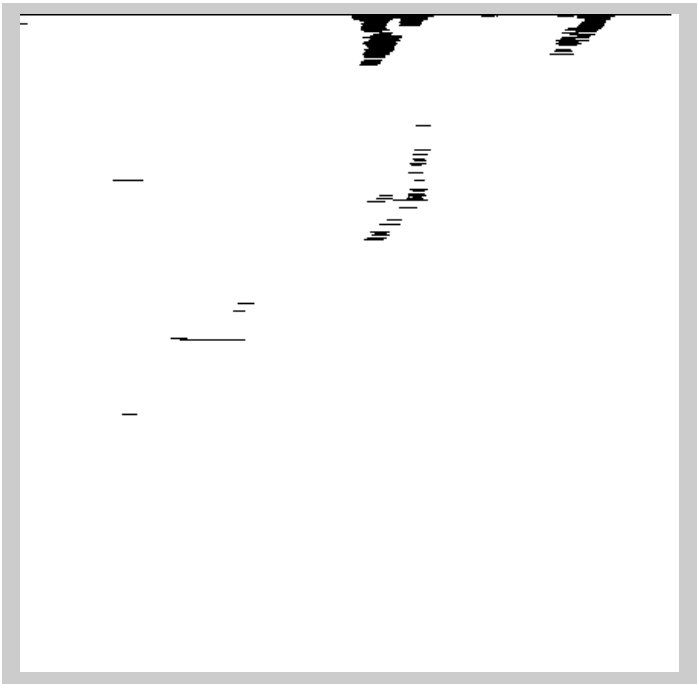


Figura 6-9: Sc

Como se observa en las imágenes, el efecto de la convolución con la horizontal permite visualizar las líneas correspondientes a las repeticiones en forma clara. El grado de "limpieza" de la matriz depende del valor de horizontal elegido. Para valores pequeños de dicha horizontal la imagen tendrá más líneas que para valores mayores.

Antes de pasar a la segunda etapa se analizarán los problemas que surgen en las imágenes y dificultan la detección correcta de las repeticiones.

6.4 Tipos de problemas en las imágenes

No todas las matrices de similitud se comportan de la misma forma ya que las canciones están compuestas de forma muy diferente. Estas diferencias se reflejan en determinados problemas en las imágenes para encontrar las líneas.

A continuación se explicará en que consisten dichos problemas, y en la próxima sección se presentará la forma de solucionarlos.

6.4.1 Problema 1: Hueco en la línea correspondiente al estribillo

Este tipo de problema aparece cuando las repeticiones del estribillo se diferencian en un lapso corto de tiempo. Generalmente ocurre por la entonación distinta de una sílaba o palabra, por la adición de instrumentos, o por arreglos musicales distintos[15].

[15] Arreglos musicales distintos son ejecuciones alteradas de la misma melodía.

Se considera la presencia de un hueco cuando dos líneas están distanciadas 4 segundos como máximo (ver Figura 6-10). Este valor fue elegido a partir del análisis de la base de datos.

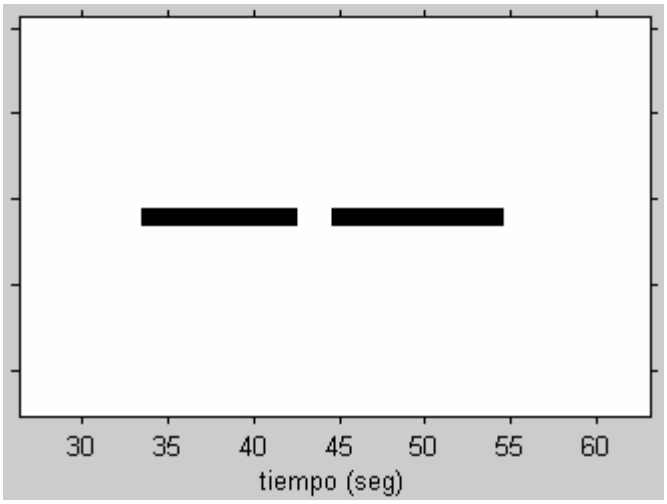


Figura 6-10: Ejemplo de hueco

6.4.2 Problema 2: Homogeneidad en la canción (Triángulos)

Los triángulos aparecen en la matriz T como un conjunto de líneas con lags cercanos (ver Figura 6-11). Esto se debe a que segmentos de la canción están compuestos por pequeños fragmentos de música similares. Generalmente los triángulos no se corresponden con el estribillo, pero hay algunas excepciones.

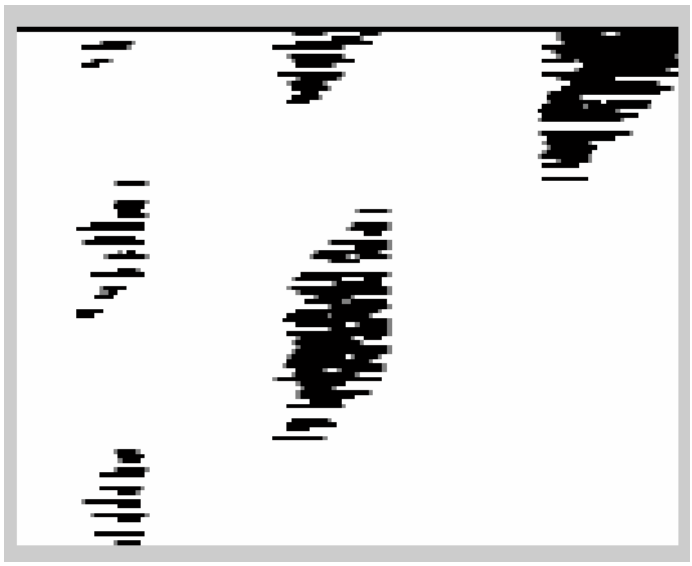


Figura 6-11: Ejemplo de triángulos

Los triángulos de la matriz T provienen de los cuadrados claros que se veían en la matriz S, y al utilizar la representación time-lag se transforman en triángulos (ver Figura 6-12).

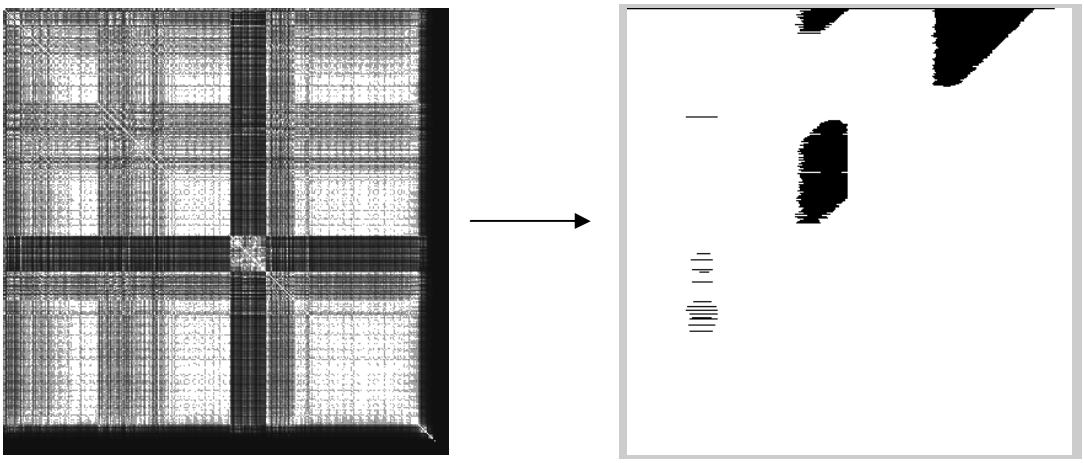


Figura 6-12: Matriz S donde la similitud se ve como cuadrados

La presencia de triángulos depende en algunos casos de cuan "limpia" esté la imagen. Suele pasar que para un valor de horizontal pequeño se observan triángulos, mientras que para valores mayores éstos no aparecen (ver Figura 6-13).

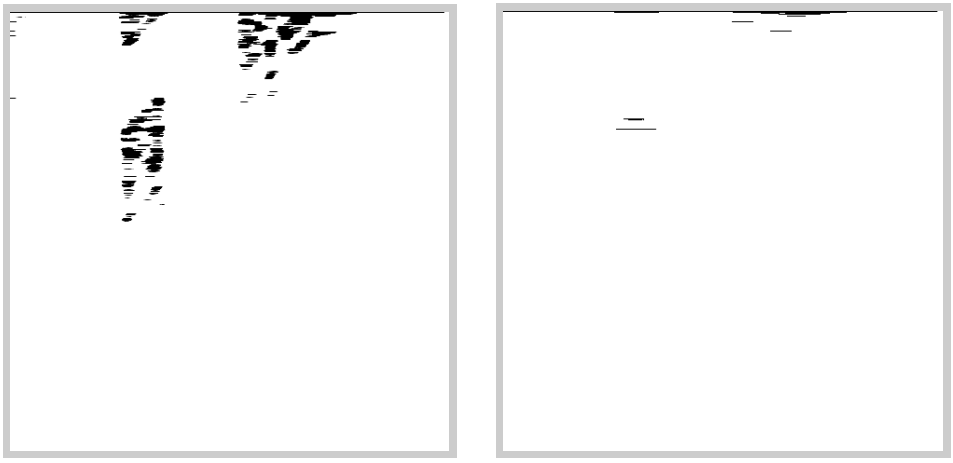


Figura 6-13: A la izquierda la imagen se convolucionó con un valor de horizontal de 11, y a la derecha con 20.

En la sección 6.5 se analizará los pasos a seguir cuando aparece un triángulo.

6.4.3 Problema 3: Repetición de fragmentos distintos

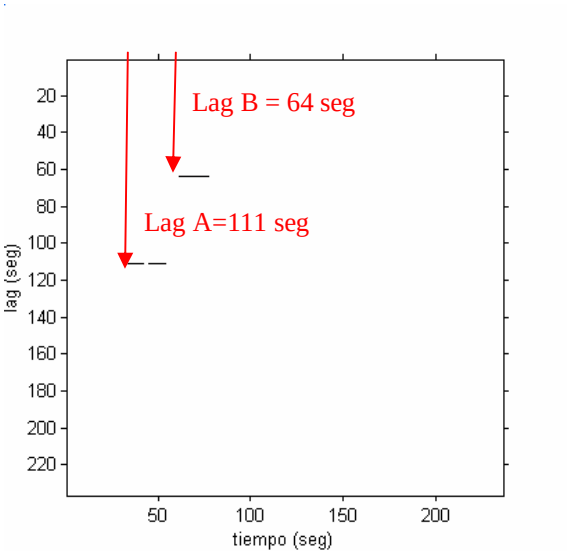
Determinar el estribillo de una canción es una tarea difícil de realizar. En numerosas canciones la estructura es tal, que existe más de un fragmento con letra y música que se repiten. Podría decirse entonces que estas canciones tienen dos o en algunos casos hasta tres estribillos diferentes.

En estos casos, lo primero que hay que hacer es identificar el problema, y luego elegir uno de ellos.

Por ejemplo, en la Figura 6-14, si las dos líneas que aparecen correspondieran al mismo estribillo, una debería ser repetición de la otra. En otras palabras, los tiempos de inicio y fin de una línea más su lag, deberían coincidir con los tiempos de la otra.

Sean t_{1A}, t_{2A} y t_{1B}, t_{2B} los respectivos tiempos de inicio y fin de los fragmento A y B; y lag A y lag B sus lags. Al sumarle los lags a dichos tiempos se verifica que no coinciden en el tiempo, por lo que un fragmento no puede ser repetición del otro.

Efectivamente, son fragmentos de música distintos y como se verá a continuación, la herramienta realiza la elección, considerando ciertas características de cada uno de ellos.



t_{1A} = 62 seg
t_{2A} = 77 seg

t_{1B} = 34 seg
t_{2B} = 54 seg

t_{1A} + lag A = 126 seg
t_{2A} + lag A = 141 seg

t_{1B} + lag B = 145 seg
t_{2B} + lag B = 165 seg

No coinciden con los fragmentos anteriores o entre sí

Figura 6-14: Ejemplo de una canción con dos estribillos diferentes

6.5 Segunda etapa: Búsqueda de repeticiones

En esta etapa, se describirá el análisis efectuado a la matriz T para encontrar los tiempos del estribillo y sus repeticiones.

La implementación de esta búsqueda se basó en el análisis de la matriz T, sus características y patologías anteriormente explicadas. Se intentó plasmar el proceso lógico que se desarrolla al observar la imagen.

Para disminuir el costo de procesamiento se submuestrea la matriz T, obteniéndose una resolución de 1 segundo, que es la resolución esperada al identificar un estribillo. A la nueva matriz se le llamará R.

6.5.1 Búsqueda de líneas en R

Se considera por hipótesis que el largo de un estribillo está comprendido entre 8 y 45 segundos. Por lo tanto, se buscarán las líneas que tengan al menos el largo mínimo y no superen el máximo. También se tendrán en cuenta los pares de líneas que formen un posible estribillo con hueco. Se asume como ruido el resto de las líneas.

Lo que se obtiene es una tabla, con los tiempos de inicio, finales y los valores de lag de cada línea hallada.

Como ejemplo se puede ver la Tabla 6-1 donde se encuentran 11 líneas. Aparecen resaltados con distintos colores los conjuntos de líneas presentes en la Figura 6-15. En el conjunto anaranjado y en el violeta no se incluyeron algunas de las líneas porque su duración es menor al mínimo establecido.

Inicio (seg)	Fin (seg)	Lag (seg)
114	125	9
177	186	9
113	123	10
176	184	10
114	123	11
114	123	12
113	121	13
175	181	13
123	133	61
54	63	107
65	73	107

Tabla 6-1: Líneas halladas

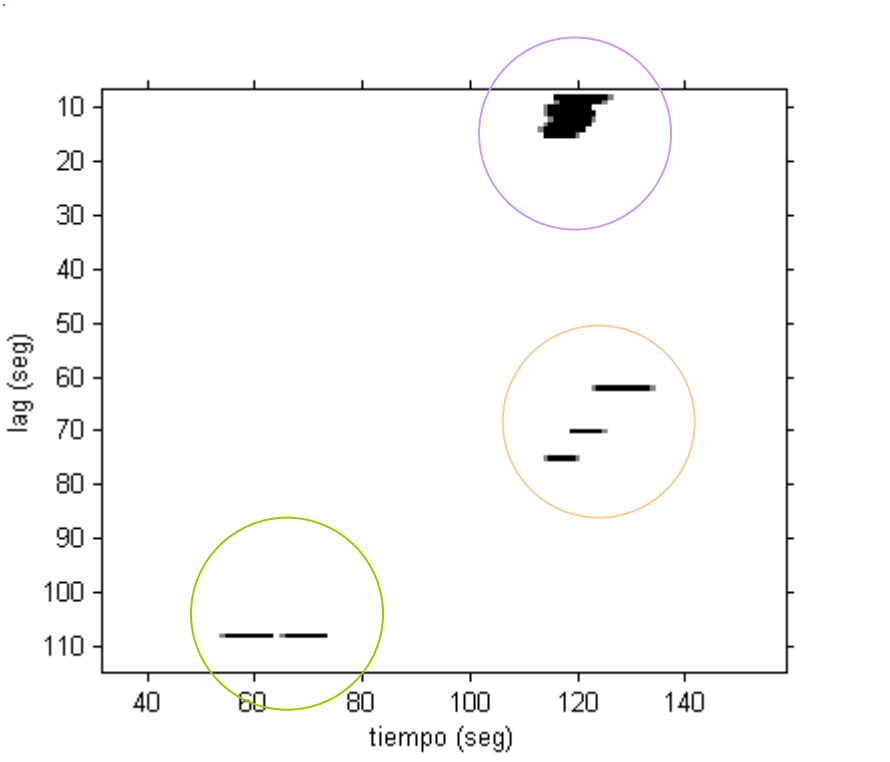


Figura 6-15: En los círculos se identifican 3 conjuntos de líneas encontrados.

Cuando la imagen presenta triángulos (problema 2) y el número de líneas halladas asciende a 15, se considera que la identificación no es exitosa para esa horizontal ya que se presentan demasiadas similitudes. Lo mismo ocurre cuando no se halla ninguna línea. En estos casos se interrumpe el procesamiento para el valor de horizontal utilizado, y se continúa el análisis para otro valor.

A lo largo del procesamiento se realiza en varias ocasiones una limpieza de las tablas calculando un largo estimado de estribillo como el largo de la línea más larga hallada hasta el momento, y eliminando las líneas cuyo lag sea menor a ese largo estimado. Con esto es posible deshacerse de algunas de las líneas que resultan de similitudes propias dentro del segmento.

6.5.2 Extensión de líneas

En este punto se analiza para cada línea hallada, la imagen original Seq (matriz de similitud ecualizada); con esto se procura recuperar los puntos de las líneas perdidos en la convolución (erosión). Extensión de líneas

Se considera la línea extendida, hasta un máximo de 5 segundos a cada lado de la misma. Luego se calcula la desviación estándar y la media de los valores de los puntos de la línea original, pero en Seq. La desviación estándar se compara con la diferencia entre los nuevos puntos y la media de los originales antes hallados. En los casos en que la diferencia calculada sea menor a la desviación estándar, se extenderá la línea incluyendo estos puntos en la tabla.

En la Tabla 6-2 se puede ver un ejemplo del resultado de la extensión de líneas.

Inicio (seg)	Fin (seg)	Lag (seg)
123	133	61
54	63	107
65	73	107

→

Inicio (seg)	Fin (seg)	Lag (seg)
123	134	61
54	64	107
65	74	107

Tabla 6-2: Líneas extendidas

6.5.3 Rellenado de huecos

Rellenado de huecos

Como se explicó anteriormente, se considera como hueco a un apartamiento entre dos líneas de 4 segundos como máximo (problema 1). Por esto si hay líneas con igual valor de lag y la distancia entre ellas es menor o igual a 4 segundos, se convierten en una única línea.

En la Tabla 6-3 se encuentran coloreados en verde las líneas originales y la resultante luego de rellenar el hueco que se presentaba.

Inicio (seg)	Fin (seg)	Lag (seg)
123	134	61
54	64	107
65	74	107

→

Inicio (seg)	Fin (seg)	Lag (seg)
123	134	61
54	74	107

Tabla 6-3: Relleno de huecos

6.5.4 Agregado de lags

Hasta ahora se trabajó con las líneas que se visualizan en la matriz R sin utilizar la información que se obtiene de sumar los lags a las mismas. A partir de este punto se agregan a la tabla los tiempos de inicio y fin que corresponden a esta suma.

Al incluir esa información luego del punto anterior, se asegura que los huecos estén completos tanto para los fragmentos originales como para los nuevos.

Inicio (seg)	Fin (seg)	Lag (seg)
123	134	61
54	74	107

→

Inicio (seg)	Fin (seg)
123	134
184	195
54	74
161	181

Tabla 6-4: Se agregan lags

6.5.5 Fundición de extremos de fragmentos coincidentes en el tiempo

Una vez que se tienen los fragmentos encontrados, se comparan uno a uno para analizar si coinciden en el tiempo (se solapan). Los inicios y finales de los fragmentos que se solapen se reemplazan por el mínimo inicio y el máximo final. Este procedimiento se denomina “fundición de extremos”.

Cuando se funden los extremos de dos fragmentos y el largo del resultado es mayor al 150% del largo estimado definido anteriormente, se considera que la fundición fue incorrecta, ya que esos dos fragmentos eran independientes. Si esto se cumple se “desolapa”, asignando el punto medio del solapamiento como final del primer fragmento e inicio del segundo.

Se puede ver un ejemplo de la fundición de extremos en la Tabla 6-5.

Inicio (seg)	Fin (seg)
51	61
102	112
48	70
100	122
106	120
183	197
54	68
182	196

→

Inicio (seg)	Fin (seg)
48	70
100	122
182	197

Tabla 6-5: Fundición de extremos

6.5.6 Asignación de grupos

En este punto se decide para el caso de que haya más de un estribillo en la canción (problema 3), cuál de ellos es el más apropiado para devolver. Para esto se asignan grupos a los fragmentos.

Se denomina que dos líneas pertenecen a un mismo grupo si una es generada por la otra más su lag, o si coinciden en el tiempo.

En la Tabla 6-6 y la Figura 6-16 se observa un ejemplo de la asignación de grupos. En la imagen aparecen únicamente las líneas halladas y no las generadas por la suma del lag.

Inicio (seg)	Fin (seg)	Grupo
53	88	1
114	150	1
9	35	2
88	113	2
162	179	2

Tabla 6-6: Asignación de grupos

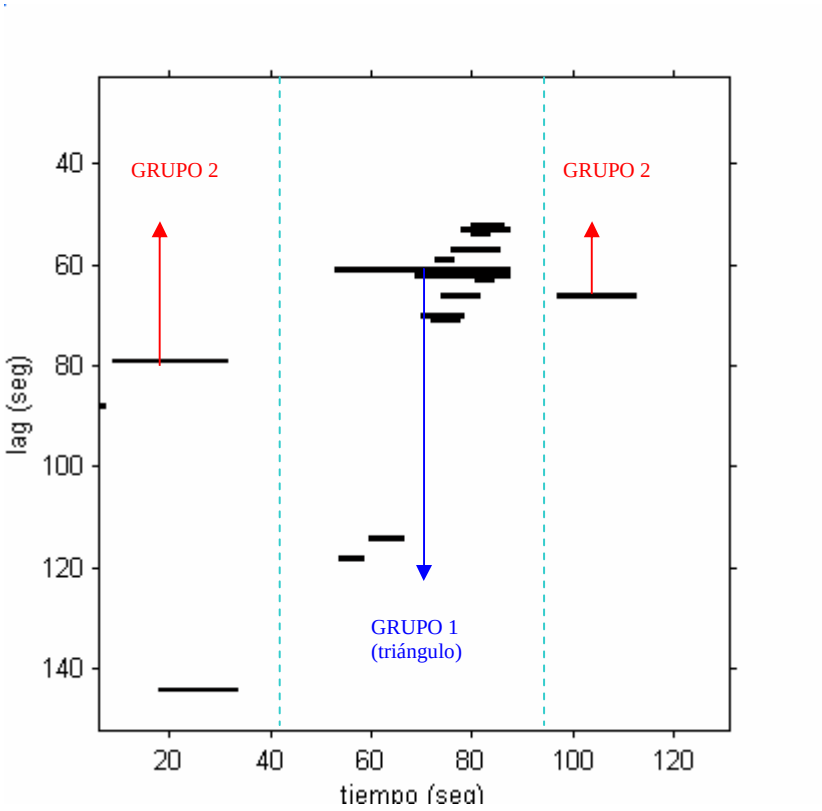


Figura 6-16: Matriz T con distintos grupos.

6.5.7 Elección de grupo

Para la elección del grupo se consideran tres características, cada una con distinto peso sobre la decisión:

- Grupo con algún componente proveniente de un triángulo. Se considera que una línea pertenece a un triángulo cuando hay más de 4 líneas que coinciden con ella en el tiempo, y tienen valores de lag cercanos.
- Número de líneas que componen el grupo.
- Máxima duración de fragmento que haya en el mismo.

En función de estos parámetros se define qué grupo es el que se devuelve para la horizontal seleccionada de acuerdo al siguiente criterio:

1. Primero se analiza si hay algún grupo con triángulos (problema 2), de ser así se eliminará ese grupo siempre y cuando exista otro grupo que no los tenga.

2. Luego, se toma como prioridad la cantidad de elementos en el grupo, que corresponde al grupo con más repeticiones de estribillo.

3. En el caso que la condición anterior encuentre más de un grupo, se elegirá el que tenga máxima duración.

4. Si las condiciones anteriores no definen un único grupo, se elegirá arbitrariamente el primer grupo.

En el ejemplo de la Tabla 6-7, de acuerdo al criterio explicado anteriormente, se elige el grupo dos ya que ninguno de sus elementos proviene de un triángulo.

Inicio (seg)	Fin (seg)	Grupo	Triángulo	Duración (seg)
53	88	1	si	36
114	150	1	si	37
9	35	2	no	27
88	113	2	no	26
162	179	2	no	18

Tabla 6-7: Asignación de grupos

En este caso las repeticiones reales se ven en la Tabla 6-8. Se verifica que el grupo elegido fue el correcto.

Inicio (seg)	Fin (seg)	Duración (seg)
18	33	16
97	112	16
163	178	16

Tabla 6-8: Repeticiones reales

6.6 Selección de las repeticiones del estribillo

El proceso descrito se realiza para distintos valores de largo de horizontal, almacenándose todos los resultados obtenidos y seleccionando uno de ellos (ver Figura 6-2). El proceso de selección consiste en evaluar en orden descendente de prioridades los siguientes parámetros:

- Número de repeticiones
- Largo de las repeticiones

Para reducir el tiempo de análisis se toma como punto de partida para el valor de largo de horizontal, el equivalente al 10% de la duración total de la canción. Este es el porcentaje promedio de duración de estribillo respecto a la duración total, calculado para las canciones analizadas. En adelante se llamará a este valor, horizontal automática. Si existe detección al utilizarla, se analizan únicamente valores de horizontales contiguos a la automática.

En caso contrario se analiza todo el rango de horizontales posible, considerando los valores 10, 15, 20, 25, 30, 35, 40 y 45. En esta primera selección se elige una de ellas y como en el caso anterior se estudian valores contiguos y se realiza la segunda selección.

Si para todos los valores de horizontal analizados se encuentran más de 15 líneas o ninguna, no se puede seleccionar una horizontal. En este caso, la identificación del estribillo se define no exitosa para esa canción y se recurre al método de identificación de fragmento representativo (IFR).

El caso en que se encuentran más de 15 líneas para todas las horizontales analizadas, implica que la matriz T contiene triángulos, y en consecuencia, en la matriz S la similitud se ve como cuadrados en lugar de diagonales. Con el método IFR se obtienen mejores resultados para este tipo de matrices ya que como se verá a continuación se busca la similitud buscando áreas lo más blancas posible.

En la Tabla 6-9 se puede ver un ejemplo de las dos selecciones antes explicadas. En este caso para la horizontal automática no existe detección. Se procede a analizar, entonces, el rango de 10 a 45. La horizontal seleccionada es la 15.

	Inicio	**Fin**	**Dur**	**Inicio**	**Fin**	**Dur**	**Inicio**	**Fin**	**Dur**
10	0	0	**0**	0	0	**0**	0	0	**0**
15	44	66	**23**	92	116	**25**	165	177	**13**
20	45	57	**13**	96	108	**13**	0	0	**0**
25	0	0	**0**	0	0	**0**	0	0	**0**
30	0	0	**0**	0	0	**0**	0	0	**0**
35	0	0	**0**	0	0	**0**	0	0	**0**
40	0	0	**0**	0	0	**0**	0	0	**0**
45	0	0	**0**	0	0	**0**	0	0	**0**

Tabla 6-9: Resultados para todo el rango de horizontales

Para la segunda selección del ejemplo (Tabla 6-10) se estudian las horizontales contiguas a la 15. Se obtiene como resultado la horizontal 14, ya que tiene la misma cantidad de repeticiones que la 15 pero una de las repeticiones es más larga.

Horiz	**Inicio**	**Fin**	**Dur**	**Inicio**	**Fin**	**Dur**	**Inicio**	**Fin**	**Dur**
11	0	0	**0**	0	0	**0**	0	0	**0**
12	0	0	**0**	0	0	**0**	0	0	**0**
13	0	0	**0**	0	0	**0**	0	0	**0**
14	44	66	23	92	119	28	165	177	**13**
15	44	66	23	92	116	25	165	177	**13**
16	44	65	23	92	116	25	0	0	**0**
17	44	57	23	92	108	17	0	0	**0**
18	44	57	23	96	108	13	0	0	**0**

Tabla 6-10: Resultados para horizontales contiguas alrededor de la elegida en la primera selección

6.7 Determinación del estribillo

Una vez definidas todas las repeticiones del estribillo de la canción, queda especificar cuál será la que se elija como resumen. Para esto se analizan nuevamente los resultados obtenidos para el resto de las horizontales, buscando la repetición más "estable".

La estabilidad de un fragmento implica que para todos los valores de horizontal analizados, ése fragmento aparezca varias veces. En caso de encontrarse dos o más fragmentos igualmente estables, se elige el más largo.

En el ejemplo, para la horizontal seleccionada las primeras dos repeticiones aparecen la misma cantidad de veces para las horizontales analizadas, pero la segunda repetición es más larga. Se designa entonces a esta repetición como estribillo y por lo tanto resumen de la canción.

Horiz	**Inicio**	**Fin**	**Dur**	**Inicio**	**Fin**	**Dur**	**Inicio**	**Fin**	**Dur**
14	44	66	**23**	92	119	**28**	165	177	**13**

Tabla 6-11: Resultados para la horizontal elegida y en verde la repetición seleccionada.

7 Identificación de fragmento representativo (IFR)

7.1 Resumen

Este método es un análisis de la matriz de similitud, alternativo al visto en el capítulo anterior. No se basa en el reconocimiento de patrones de repetición, sino que simplemente evalúa las zonas de la matriz donde se encuentra más similitud.

Esto lo hace un método muy sencillo de implementar y con un corto tiempo de procesamiento, aunque los resultados obtenidos son muy inferiores a los del método IE como se verá en las evaluaciones finales.

7.2 Procedimiento

Partiendo de la Matriz de Similitud ecualizada Seq vista en el Capítulo 5, se define una medida de similitud llamada semejanza media, $\bar{S}$, que analiza la similitud de un segmento con respecto al resto de la canción.

Dado un segmento que comienza en un punto q y termina en r, $\bar{S}$ se calcula como el promedio de los valores de los puntos de S, en el rectángulo formado por la comparación del segmento con la canción en su totalidad. El rectángulo descripto se puede visualizar en la Figura 7-1.

La semejanza media está definida entonces, de la siguiente forma:

$$\bar{S}(r,q) = \frac{1}{N(r-q-1)} \cdot \sum_{i=q}^{r-1} \sum_{j=1}^{N} S_{eq}(i,j),$$

donde N es el número total de ventanas (ancho de la matriz Seq).

De conocerse o deducirse información relevante sobre la estructura de la matriz, puede ser incluida en el análisis mediante un vector de pesos:

$$\bar{S}(r,q) = \frac{1}{N(r-q-1)} \cdot \sum_{i=q}^{r-1} \sum_{j=1}^{N} w(j) \cdot S_{eq}(i,j),$$

Generalizando para cualquier segmento de largo L, la semejanza media se calcula según el vector $Q_L(k)$ cuyo k-ésimo componente (k correspondiente al comienzo de cada segmento) se define como:

$$Q_L(k) = \bar{S}(k, k+L-1) = \frac{1}{N(L-1)} \cdot \sum_{i=k}^{k+L-1} \sum_{j=1}^{N} S_{eq}(i,j), \qquad k = 1,2,...N-L+1.$$

Q_L tiene entonces, N – L+1 componentes. Para hallar el comienzo del fragmento que se va a devolver, hay que calcular el k que corresponde al máximo valor de Q_L.

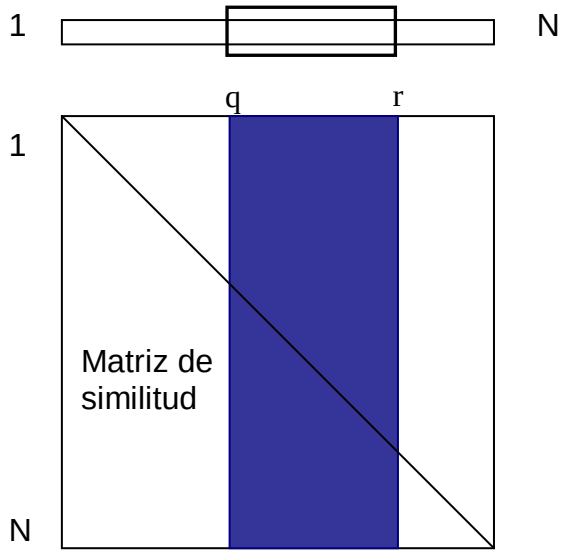


Figura 7-1: El promedio de los puntos del rectánguloes la semejanza media del segmento con el resto de la canción.

En capítulos posteriores se presenta un análisis estadístico de los resultados de este método en comparación con el IE.

8 Evaluación y elección de métodos para el sistema

8.1 Resumen

Este capítulo está dedicado a detallar la evaluación y elección de los métodos presentados, así como también de sus parámetros. Por tanto, contiene la estructura y métodos empleados en el sistema de identificación de resumen implementado.

En las secciones 8.2 y 8.4 se explican los criterios empleados en la construcción de la base de datos de canciones utilizada y la definición de resumen válido, empleados en la evaluación respectivamente.

Inicialmente se analizan los métodos de *MFCC* y VC para elegir las configuraciones óptimas de sus parámetros. Luego se realiza una comparación entre éstos para elegir el que presente el mejor desempeño y se analiza IFR para seleccionar también la mejor configuración. Finalmente se evalúan los métodos de identificación del resumen: *Identificación de Estribillo* e *Identificación de Fragmento Representativo*.

8.2 Base de datos

El universo de trabajo se limitará mediante la selección de un conjunto de canciones, que se procuró fuera lo más amplio posible, teniendo como único requisito el cumplimiento de las hipótesis exigidas: canciones que contaran con estribillos, principalmente de rock y pop.

Para verificar que una canción contiene estribillos y realizar posteriormente la validación de los resultados, se procedió a encontrar los estribillos manualmente. Se escucharon las canciones atentamente hasta poder identificar los inicios y finales de cada una de sus repeticiones.

En la elección de los "estribillos reales", fue inevitable tener cierta duda en canciones con distintos estribillos posibles. Esto es, dos fragmentos que se repiten a lo largo de la canción, pero distintos entre sí (ver sección 6.4.3).

Dada la subjetividad de los datos se solicitó a 20 personas que etiquetaran los estribillos. Se presentaron las canciones a cada persona junto con su estribillo correspondiente, pidiendo que lo evaluaran según fuera correcto y preciso[16].

Inicialmente los algoritmos fueron calibrados utilizando una base de 100 canciones, conocido como *conjunto de entrenamiento* (*Apéndice A.a*). A partir de estos resultados, se establecieron los parámetros del sistema.

[16] La encuesta se realizó en base a la lista del *Apéndice A.a* y los resultados se presentan en el apéndice B.

Cumplida esta primera etapa de calibración, se configuró una segunda lista de 50 canciones, *conjunto de validación* (*Apéndice A.b*), disjunta con la anterior. Estas procuran complementar el conjunto de entrenamiento con temas y artistas ampliamente difundidos, no incluidos antes. También pretende ampliar el espectro de estilos musicales, incluyendo temas actuales y de las últimas décadas.

El motivo para evaluar el desempeño de los métodos utilizando una nueva base, es verificar que los parámetros obtenidos en la calibración sean independientes de la base de datos antes utilizada.

8.3 Procedimiento de evaluación

En la siguiente figura se muestra el esquema del procedimiento utilizado para realizar la evaluación.

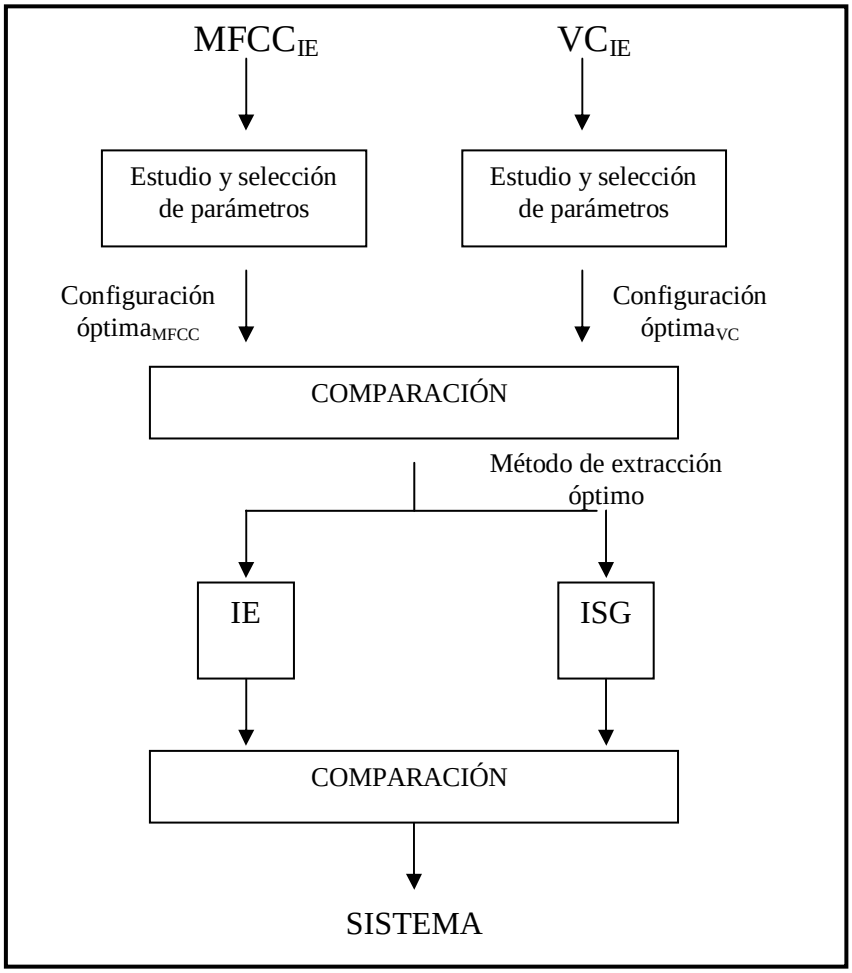


Figura 8-1: Procedimiento de evaluación del sistema

La evaluación se realizó en base a la implementación en MATLAB de los métodos de extracción de características *MFCC* y VC, así como también de los métodos de identificación IE e IFR.

Para la identificación del resumen se empleó el método de *Identificación de Estribillo (IE)*[17]. Primeramente se estudian distintas configuraciones de parámetros para seleccionar la configuración óptima para cada método, en base a los criterios explicados a continuación.

A partir de la configuración óptima se comparan *MFCC* y VC mediante *IE*, eligiendo el que presente mejores resultados. Éste será el método de extracción de características implementado en la herramienta.

Los parámetros de *IFR* son estudiados para el método de extracción que resulte de la selección anterior. Dada la complejidad de *IE* no se analizaron mediante estadísticas los parámetros, sino en forma empírica. Finalmente se analizan los métodos de identificación *IE* e *IFR* con el método de extracción seleccionado.

8.4 Criterios de Validación

El proceso de validación consiste en comparar los resultados obtenidos mediante la herramienta, con los obtenidos manualmente: *estribillos reales* (E_R). Este proceso se puede observar en la Figura 8-2.

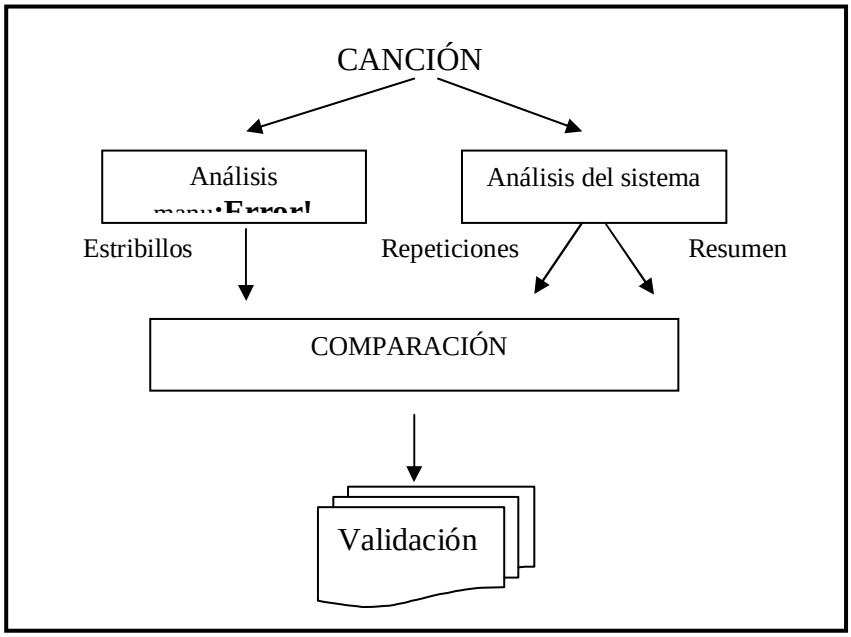


Figura 8-2: Esquema del proceso de validación

Los resultados para la devolución del resumen (R), y la de sus repeticiones (RR) se analizaron de forma independiente

[17] Se realiza la evaluación sólo para el método de *Identificación de Estribillo*. Esto se debe a la experimentación inicial donde se observa que los resúmenes devueltos por IE se aproximan mejor al estribillo que los de *ISG*. De todas formas se realiza posteriormente esta verificación.

8.4.1 Criterios de validación de R

Para estipular el éxito de la devolución de R, se consideraron las siguientes variables:

- Cobertura: representa el porcentaje de *R* que coincide con algún E_R, es una medida de cuanta información del estribillo real se encuentra en el resumen.

$$C_R = \frac{\text{longitud}(\text{R I E}_\text{R})}{\text{longitud}(\text{E}_\text{R})}$$

- Precisión: representa el porcentaje de E_R que coincide con *R*, es una medida de cuan precisa es la detección del resumen.

$$P_R = \frac{\text{longitud}(\text{R I E}_\text{R})}{\text{longitud}(\text{R})}$$

Se realizan tres análisis diferentes utilizando la cobertura y la precisión para evaluar distintos aspectos de los resultados. También se tendrá en cuenta el caso en que el sistema no devuelva resultados para *IE*, que llamaremos *vacío*.

Análisis 1: Cobertura

En un primer análisis, se considera únicamente la cobertura, y un resumen es calificado como válido si el porcentaje de cobertura es mayor a 70%. De esta forma, se admite que la duración de algunos de los resúmenes sea mayor a la detectada manualmente, pero el estribillo estará presente. Es justamente por esto que daremos prioridad a este análisis, porque está en concordancia con nuestro objetivo de detección completa.

Por medio de este análisis se diferenciaron los casos que presentan una cobertura del 0%, lo que implica que la identificación resulta en un estribillo *falso*, de aquellos en que existió un mínimo de cobertura.

Se define la *cobertura promedio* como el porcentaje de cobertura obtenido en las canciones donde se detecta algo del estribillo, o sea las que no son *falso* ni *vacío*.

Análisis 2: Combinación de cobertura y precisión

Representa un análisis más estricto que el anterior, donde se da prioridad a la cobertura pero considerando también la precisión de los resultados. Califica al resumen identificado como válido por el criterio anterior, exigiéndole adicionalmente una precisión de al menos un 50 %.

Para el estudio de este análisis, se observa en la Figura 8-3 un diagrama de la estructura básica de una canción y los posibles fragmentos hallados, discriminando su validez según la posición.

Los fragmentos coloreados en verde son aquellos considerados válidos. En los fragmentos número uno, dos y tres los extremos no se corresponden con exactitud a los reales. El caso cuatro es el idealmente hallado.

Los amarillos representan devoluciones consideradas inválidas para este análisis. En el número 5 es debido a que la precisión es menor a 50% y en el número 6 por no alcanzar la cobertura necesaria.

Por último, los fragmentos anaranjados muestran aquellos que no cumplen ninguno de los requerimientos y los rojos los que no representan en lo absoluto al estribillo (estribillo *falso*).

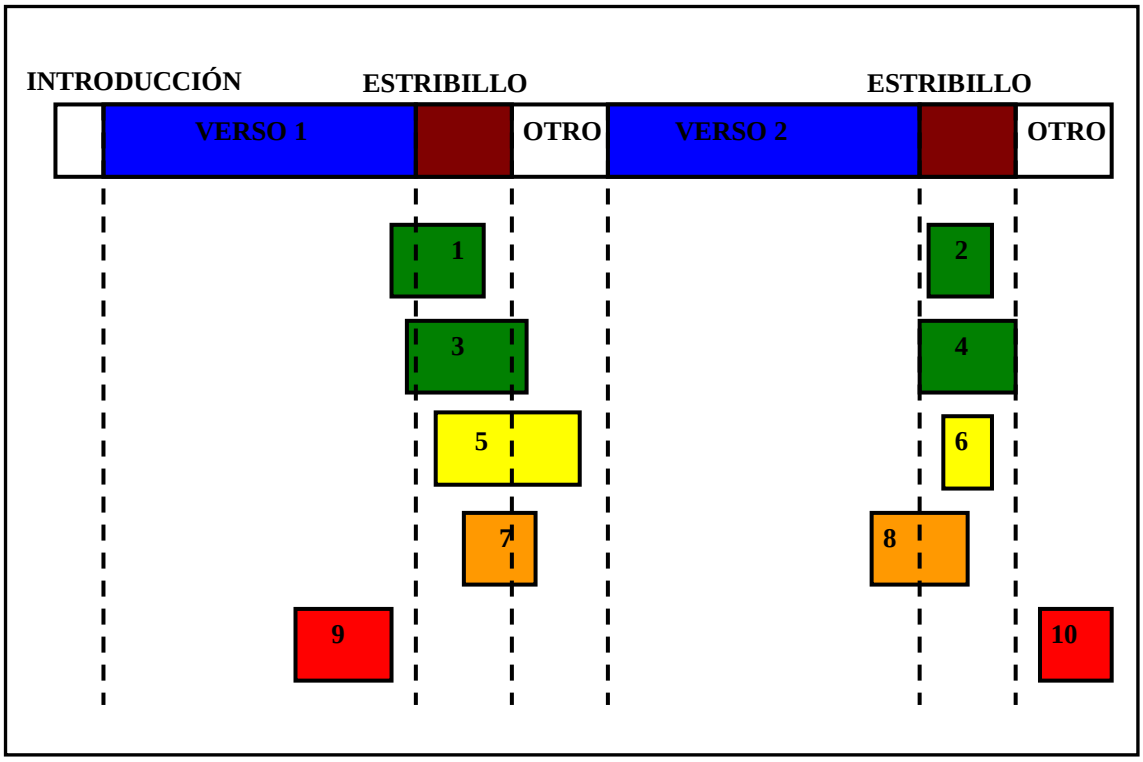


Figura 8-3: Diagrama de las posibles posiciones de los resúmenes con respecto a los estribillos parael estudio del segundo análisis.

Análisis 3: Medida F_1

La medida F_1[18] consiste en realizar un promedio armónico de la cobertura y la precisión:

$$F_1 = \frac{2CP}{C+P}.$$

La F_1 presenta igual peso tanto en la cobertura como en la precisión. Es por esto que puede resultar una medida ambigua como se puede ver en el ejemplo de laFigura 8-4:

[18] Ver apéndice E

Ejemplo de resultados de cobertura, precisión y F_1.: da un mismo resultado para coberturas y precisiones muy distintas.

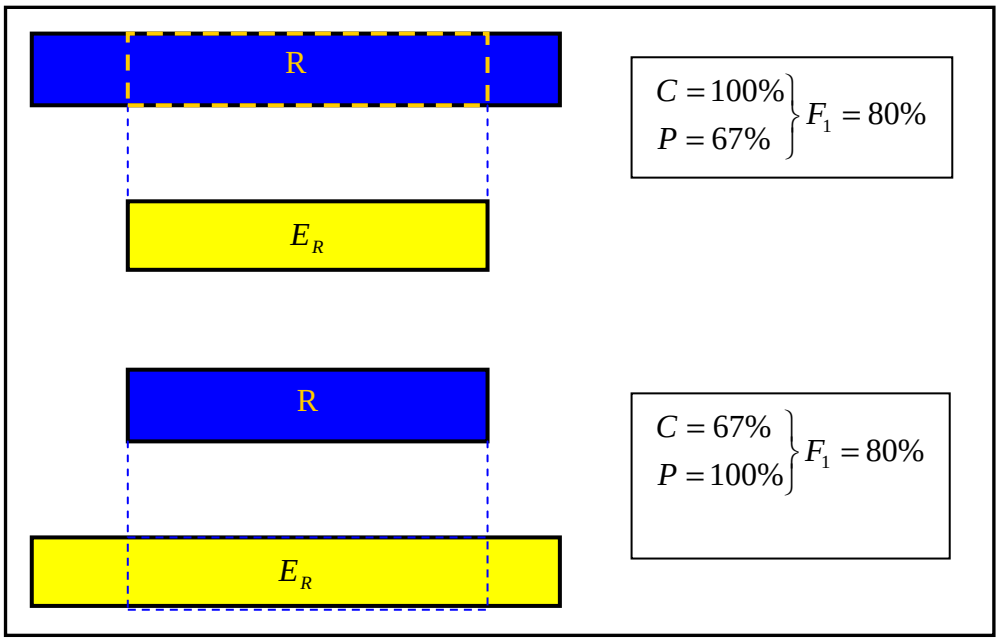


Figura 8-4: Ejemplo de resultados de cobertura, precisión y F_1.

Debido a esta ambigüedad y a que se pretende encontrar un buen porcentaje del estribillo, este análisis no se usa en la selección de parámetros y métodos. De todas formas se analizará la F_1 para las canciones del *conjunto de validación*, ya que también fue utilizada como evaluación en trabajos previos [6], [8]. Los resúmenes identificados se definen como válidos si la F_1 es mayor al 70%.

8.4.2 Criterios de validación de RR

Para la validación de las repeticiones se consideran las mismas nociones de cobertura y precisión anteriormente explicadas pero adaptadas de la forma siguiente:

- Cobertura: representa el porcentaje total de tiempo en los que los R_i coinciden con los E_{Rj}.

$$C_{RR} = \frac{\sum_{i,j} \text{longitud}\left(R_i \; I \; E_{Rj}\right)}{\sum_{j} \text{longitud}\left(E_{Rj}\right)}$$

- Precisión: representa el porcentaje total de tiempo en que los E_{Ri} coinciden con los R_j.

$$P_{RR} = \frac{\sum_{i,j} \text{longitud}(R_i \text{ I } E_{Rj})}{\sum_{i} \text{longitud}(R_i)}$$

En este caso también se realizan los tres análisis antes explicados.

8.5 Evaluación de las configuraciones de MFCC y VC

Luego de haber explicado los criterios de validación, se pasa a evaluar los métodos.

8.5.1 Elección de parámetros para MFCC (con IE)

Las estadísticas para *MFCC* se hallaron con los distintos parámetros descriptos en la sección 4.3.4. Primeramente se buscó la "mejor" configuración de parámetros de forma experimental, resultando ser la opción 1 de la Tabla 8-1. En base a dicha opción, se modificó un parámetro por vez con el fin de elegir la configuración óptima y analizar los cambios que provocan los parámetros.

	Rango de frecuencias	N_{mfcc}	Uso del 1er coeficiente	Datos ventana
Opción 1	20 a 3000	12	Si	1s, solap 50%
Opción 2	20 a 3000	12	No	1s, solap 50%
Opción 3	20 a 3000	12	Si	2s, solap 50%
Opción 4	20 a 3000	12	Si	0.5s, solap 50%
Opción 5	20 a 3000	12	Si	0.2s, solap 50%
Opción 6	20 a 3000	6	Si	1s, solap 50%
Opción 7	20 a 3000	30	Si	1s, solap 50%
Opción 8	20 a 4000	12	Si	1s, solap 50%
Opción 9	20 a 8000	12	Si	1s, solap 50%
Opción 10	20 a 8000	12	No	1s, solap 50%

Tabla 8-1: Opciones de análisis para MFCC

En el Apéndice C.a se detallan las estadísticas. En las figuras siguientes se pueden apreciar los resultados para los criterios más relevantes.

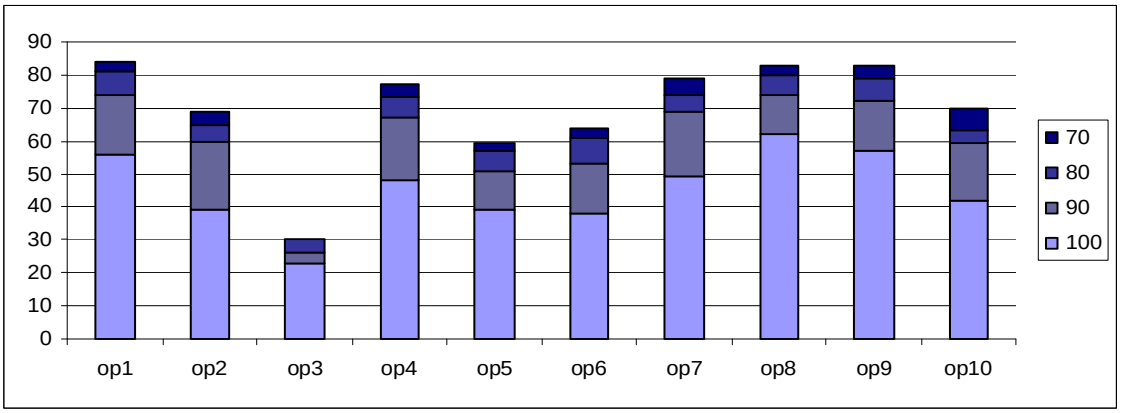


Figura 8-5: Análisis de cobertura de resumen para distintas opciones de MFCC con IE

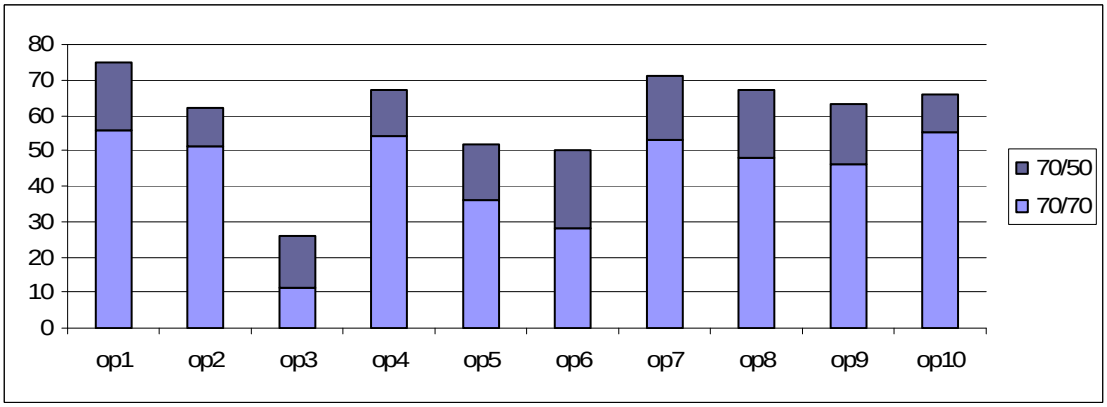


Figura 8-6: Análisis de combinación de cobertura y precisión de resumen para distintas opciones de MFCC con IE

En la Figura 8-5 se aprecian los porcentajes de canciones que verifican una cobertura del 70% del resumen (análisis 1). Las mejores opciones resultaron ser la 1, la 8 y la 9. Del mismo modo, en la Figura 8-6 se presenta el análisis 2; se observa que las mejores opciones son la 1 y 7.

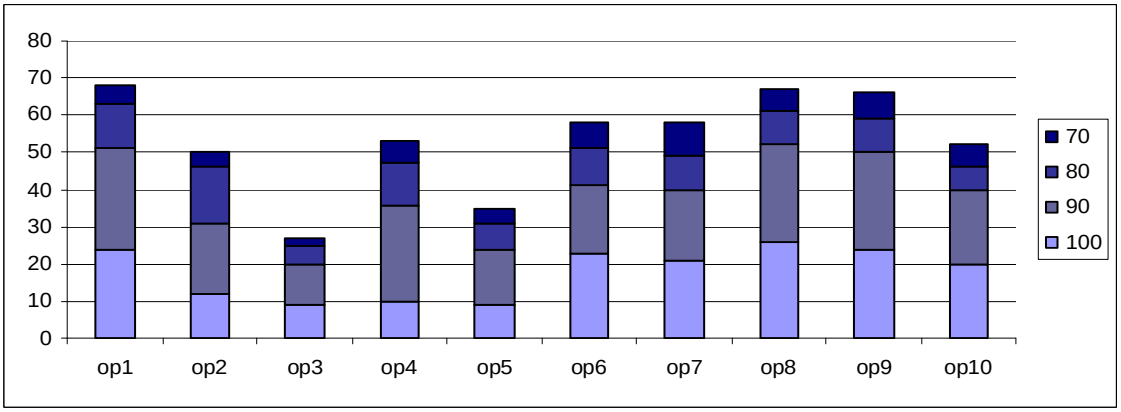


Figura 8-7: Análisis de cobertura de repeticiones para distintas opciones de MFCC con IE

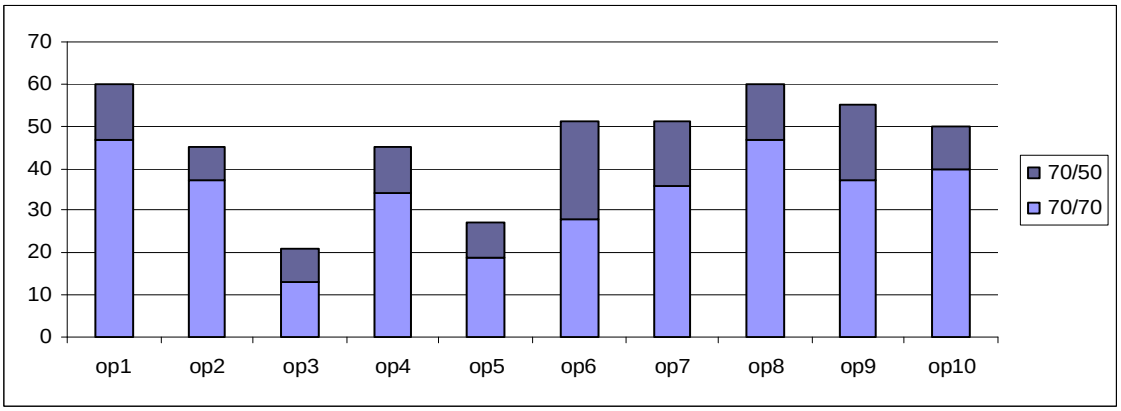


Figura 8-8: Análisis de cobertura y precisión de repeticiones para distintas opciones de MFCC con IE

El análisis 1 para las repeticiones del resumen (Figura 8-7) devuelve como mejores opciones la 1, 8 y 9, mientras que el análisis 2 (Figura 8-8) encuentra como mejores opciones la 1 y 8.

Las sutiles diferencias en ciertas estadísticas pueden dar lugar a incertidumbre en la elección de la configuración óptima. Si se contara con una base de datos de mayor tamaño, las estadísticas podrían ser más determinantes. Es por esto que en el momento de elegir la configuración óptima, se tuvieron en cuenta las estadísticas y el estudio de los parámetros realizado en la sección 4.4.4.

En conclusión, encontramos que la mejor opción es la 1.

	Rango de frecuencias	N_{mfcc}	**Uso del 1er coeficiente**	**Datos ventana**
Opción 1	20 a 3000	12	si	1s, solap 50%

Al analizar ciertos parámetros, se observó que las estadísticas no devuelven resultados intuitivos, debido a las características de *IE*. Este es el caso de la comparación de la opción 1 con la 6, que se diferencian en el número de coeficientes *MFCC*, 12 y 6 respectivamente. Ya que el uso de un número bajo de coeficientes implica una aproximación menos exacta del timbre, se debería encontrar una mayor similitud en la comparación y por tanto un mayor porcentaje de canciones con resumen válido.

El método IE puede no resultar el más adecuado para este tipo de análisis ya que en los casos que se encuentra demasiada similitud no devuelve resultados[19], no diferenciando de aquellos en que no se encuentra similitud alguna. Hay un 22% de canciones donde no se encuentra ningún estribillo en la opción 6 de acuerdo a la Tabla C-1. Luego del análisis de las matrices de estos grupos se concluye que es debido al exceso de similitud que no se obtienen resultados en la mayoría de los casos.

[19] Ver sección 6.4.2

Las opciones 7 y 1 analizan los efectos de utilizar un mayor número de coeficientes *MFCC*, con 30 y 12 coeficientes respectivamente. La última opción da mejores resultados como se advierte de la Figura 8-5 a la Figura 8-8, detectando una menor similitud con 30 coeficientes y presentando la opción 7 una mayor exactitud en el estribillo hallado.

Se analizaron distintos tamaños de ventana en las opciones 1, 3, 4 y 5. La opción 3 de ventanas de 2 segundos presentó una muy baja precisión, con un buen porcentaje de cobertura. Esto es debido a que se encuentra demasiada similitud por ser una ventana grande. Las opciones 4 y 5 por el contrario tienen un 14 y un 19% de canciones con cobertura menor al 70%, mientras que la opción 1 tiene un 6% de acuerdo a la Tabla C-1. De aquí se concluye que utilizar ventanas de 1 segundo (opción 1) es un buen compromiso entre cobertura y precisión para encontrar similitud mediante *MFCC*.

Las opciones 1 y 2 comparan el uso del primer coeficiente para los rangos de filtros de hasta 3 kHz, y la 9 y 10 hasta 8kHz. Para ambos rangos se observa que sin utilizar el primer coeficiente los estribillos se hallan cortos, o sea que se encuentra poca similitud,. Esto se extrae de la Tabla C-1, sin utilizar el primer coeficiente hay un 69% de canciones con cobertura mayor al 70% y un 16% con menor cobertura que la anterior. Por el contrario, utilizando el primer coeficiente se tiene un 84% y un 6% respectivamente.

Por último los rangos de filtros se analizan en las opciones 1 y 9. Para el análisis 1 ambos dan resultados muy similares; con el criterio 2 se obtiene un 5% más éxitos que con la opción 1. Analizando en detalle las tablas del Apéndice de estadísticas C.a, y a partir del análisis cualitativo sobre los parámetros realizado en la sección 4.3.4, se concluye que el uso de un banco hasta 3 kHz presenta mejores resultados respecto a la precisión.

8.5.2 Configuración de parámetros para VC (con IE)

Se busca la configuración óptima para la extracción de características mediante VC. En la Tabla 8-2 se presentan las distintas opciones analizadas empleando IE: rangos de frecuencias a analizar y tamaños de ventana.

	Rango de octavas[20]	**Datos ventana**
Opción 1	1 a 8	1s, solap 50%
Opción 2	1 a 7	1s, solap 50%
Opción 3	1 a 8	0.2s, solap 50%
Opción 4	1 a 8	0.5s, solap 50%

Tabla 8-2: Opciones de análisis para VC

Las figuras presentadas a continuación contienen el análisis 1 y 2 de para la identificación del estribillo y sus repeticiones.

[20] Los rangos de frecuencia correspondientes a las octavas se explican en la sección 4.4.2. Un rango de 3 a 8 octavas es de 130 a 8000 Hz mientras que el rango de 3 a 7 es de 130 a 4000 Hz.

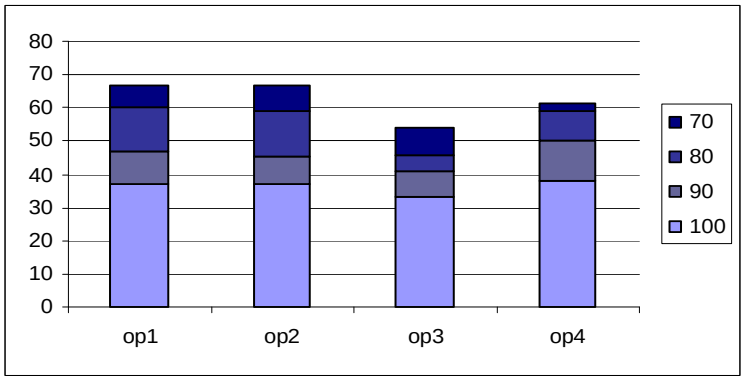


Figura 8-9: Análisis de cobertura de resumen para distintas opciones de VC con IE

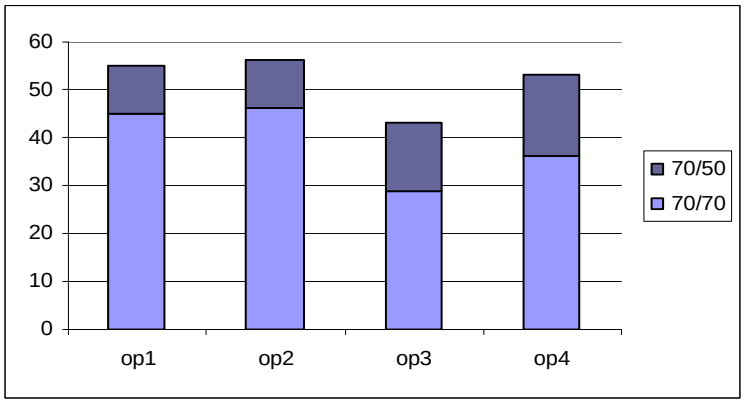


Figura 8-10: Análisis de combinación de cobertura y precisión de resumen para distintas opciones de VC con IE

En la identificación del estribillo los mejores desempeños se obtienen para las opciones 1 y 2, resultando mejor la 2 por una mínima diferencia.

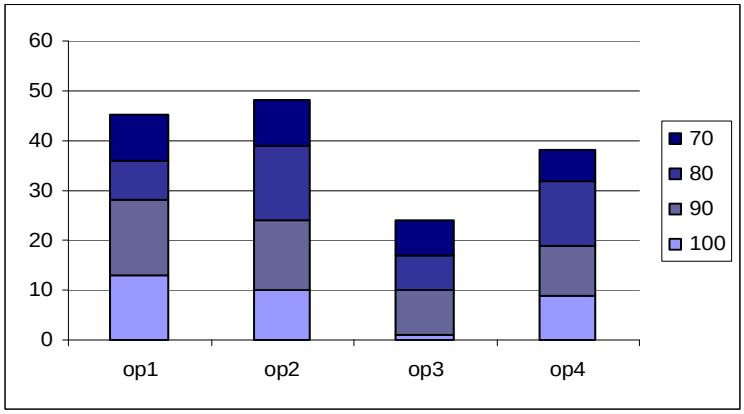


Figura 8-11: Análisis de cobertura de repeticiones para distintas opciones de VC con IE

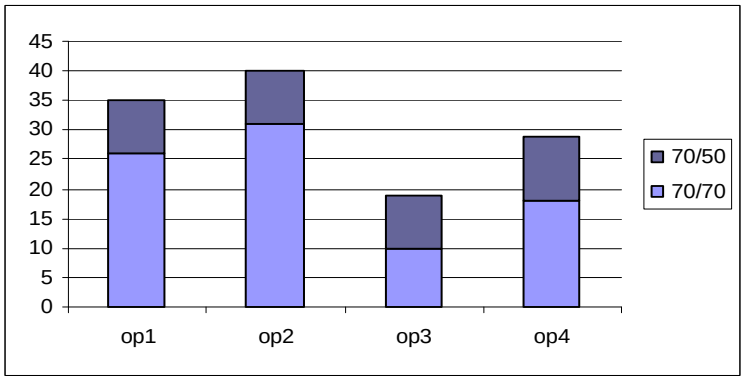


Figura 8-12: Análisis de cobertura y precisión de repeticiones para distintas opciones de VC con IE

Las estadísticas de las repeticiones muestran que el rango de octavas no sería un parámetro determinante, ya que no se aprecia una variación significativa. Las opciones 1 y 2 presentan poca diferencia en sus estadísticas. El tamaño de ventana si resulta ser decisivo, siendo 1 segundo la mejor opción.

Comparando las opciones 1, 3 y 4, aplicando el mismo análisis que se realizó en *MFCC* a las tablas del apéndice de estadísticas, se observa que una ventana menor a 1 segundo, encuentra los estribillos demasiado cortos.

Esto se puede deber al tamaño de la base de datos, de todas formas de las figuras anteriores se observa que las estadísticas más altas se dan en la opción 2. Esto sumado a que el rango de frecuencias, representadas por las octavas, se corresponde con la selección de configuración óptima para *MFCC*, define la opción 2 como la configuración óptima.

Configuración Óptima	Rango de octavas	Datos ventana
Opción 2	3 a 8	1s, solap 50%

8.6 Comparación de los métodos de extracción de características: MFCC y VC (óptimos con IE)

De la comparación cualitativa entre las diferencias entre *MFCC* y VC presentado en la sección 4.6, se procede a evaluar los resultados obtenidos con estos métodos. Para ello se hallan las estadísticas, detalladas en el apéndice C, para los distintos análisis en identificación de resumen y sus repeticiones.

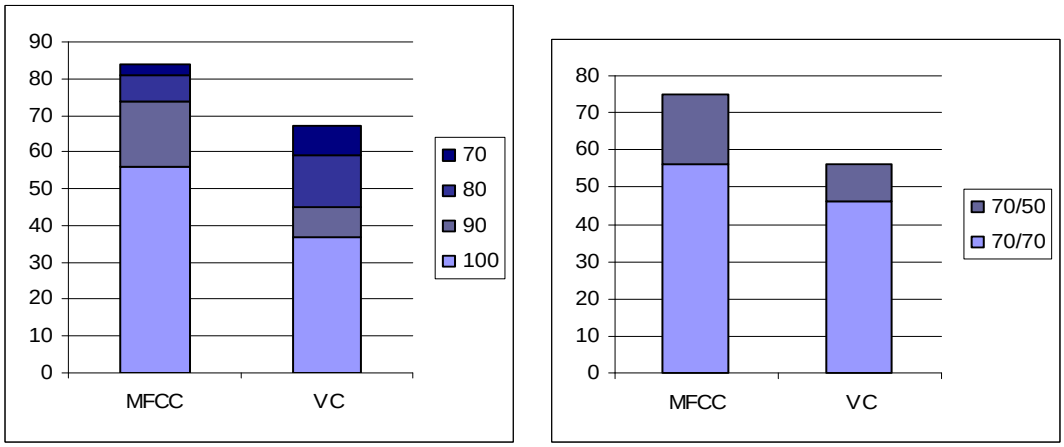


Figura 8-13: Porcentaje de canciones según análisis de cobertura y de cobertura-precisión para estribillo

En la figura anterior se puede apreciar un desempeño superior de *MFCC* frente a VC en la identificación del estribillo. Cabe destacar que el porcentaje de cobertura para *MFCC* dentro de los estribillos identificados es del 94% mientras que para VC es un 86%. Esto indica que los estribillos se hallan más cortos empleando VC. Este último presenta a su vez un 16% de canciones con estribillo falso, mientras que *MFCC* presenta un 7%.

Estas diferencias de desempeño se mantienen para la identificación de las repeticiones del estribillo como se puede apreciar en el apéndice C. Por esto se decide emplear *MFCC* como método de extracción de características en el sistema de identificación de resumen.

8.7 Análisis de IFR (con MFCC óptima)

Cabe destacar nuevamente que IFR no detecta repeticiones, sino que busca un fragmento de largo fijo que presente similitud global con el resto de la canción. Este no se adapta a la canción, por lo tanto es de esperar que el estribillo no se encuentre con exactitud.

Habiéndose detectado experimentalmente que en numerosas ocasiones retornaba la música del principio o final, se estudió la aplicación de un vector de pesos sencillo. El mismo cumple la función de suprimir los inicios y finales, de modo de no considerarlos en el análisis.

Por consiguiente los parámetros a determinar son el largo óptimo del fragmento a devolver y el efecto del vector de pesos (según el porcentaje de canción truncado). En la tabla Tabla 8-1 se detallan las configuraciones analizadas.

	Largo	Vector de peso
Opción 1	20	15%
Opción 2	25	15%
Opción 3	30	15%
Opción 4	25	10%
Opción 5	30	20%

Tabla 8-3: Opciones de análisis para IFR

En la Figura 8-14 se demuestran los resultados para el análisis de cobertura, detallando el resto del estudio estadístico en el apéndice C.d. Como es de esperar cuanto más largo es el fragmento, mayores son los porcentajes de acierto (existe más probabilidad de encontrar algún trozo de estribillo). Las opciones 4 y 5 muestran los mejores resultados, sin embargo se selecciona el menor largo por considerarse 30 segundos demasiado largo para muchas canciones.

PORCENTAJE COBERTURA	op1	Op2	op3	op4	op5
100	7	15	23	11	19
MAYOR A 90	12	21	30	19	23
MAYOR A 80	22	33	38	30	33
MAYOR A 70	27	43	50	64	64
PROMEDIO	60	68	71	66	67
FALSO	30	24	22	26	23

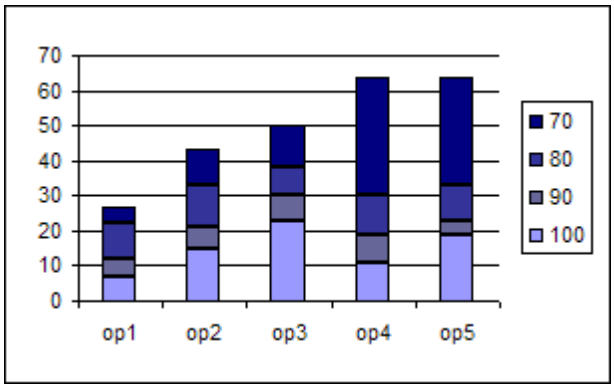


Figura 8-14: Análisis de cobertura para resumen

8.8 Comparación de los métodos de identificación de resumen: IE e IFR (con MFCC óptima)

A continuación se presenta la comparación entre los métodos *IE* e *IFG*, analizando su desempeño en la identificación del estribillo, presentando el primero un desempeño superior en todos los análisis.

PORCENTAJE COBERTURA	IE	IFR
100	56	11
MAYOR A 90	74	19
MAYOR A 80	81	30
MAYOR A 70	84	64
PROMEDIO	94	66
FALSO	7	24
NINGUNO	3	0

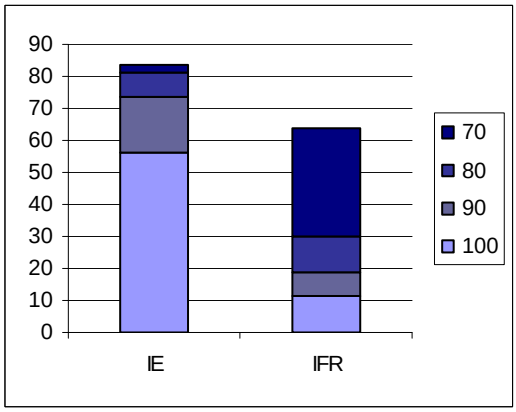


Figura 8-15: Porcentaje de canciones según cobertura mediante IE e IFS.

Se observa que IFR tiene un porcentaje considerable de estribillos válidos (según el análisis de cobertura), pero se confirma la poca exactitud ya que tiene un porcentaje de cobertura promedio del 66%. Aunque el resumen pueda no ser exacto y coincidir en un corto tiempo con el estribillo, el método tiene la cualidad de encontrar siempre un resumen.

Es justamente por esto que se decide incluir este método alternativo en la implementación del sistema, en los casos que no haya identificación por encontrar demasiada similitud. La matriz procesada con IFR detectará así el estribillo, sin obtener las repeticiones. Cabe destacar que el uso de IFR brinda al sistema la posibilidad de identificar un fragmento representativo en canciones que no posean estribillo.

Para la base de datos analizada se obtienen 3 canciones en las cuales no se identifica ningún estribillo. A continuación se analizan estos casos con IFR y se comparan con estribillo y sus repeticiones. Estas canciones presentan gran similitud en ciertas zonas, impidiendo la detección de líneas en la matriz, este es el problema 2 visto en la sección 6.4.2. Por lo tanto se puede considerar a este método como una solución al mismo.

Arranca Corazones - Ataque 77			
	Inicio (s)	**Fin (s)**	**Duración (s)**
estribillo 1	46	71	26
estribillo 2	114	140	27
estribillo 3	157	180	24
estribillo 4	181	205	25
resumen representativo	**171**	**195**	**25**

Can´t stand loosing you - The Police			
	Inicio (s)	**Fin (s)**	**Duración (s)**
estribillo 1	29	39	11
estribillo 2	73	93	21
estribillo 3	126	168	43
resumen representativo	**77**	**101**	**25**

When I Grow Up – Garbage			
	Inicio (s)	**Fin (s)**	**Duración (s)**
estribillo 1	57	69	13
estribillo 2	84	98	15
estribillo 3	125	139	15
resumen representativo	**99**	**123**	**25**

8.9 Sistema de identificación de resumen

En la siguiente figura se puede apreciar el diseño del sistema definitivo, con los métodos elegidos de extracción de características e identificación.

Los parámetros de *MFCC* de la configuración óptima son los siguientes:

	Rango de frecuencias	N_{mfcc}	**Uso del 1er coeficiente**	**Datos ventana**
Opción 1	20 a 3000	12	si	1s, solap 50%

La detección de IE (si devuelve o no resultados el método, de acuerdo a los criterios expuestos en el capítulo 6), determina el tipo de resultados. Si hay detección exitosa, el sistema devolverá el estribillo y sus repeticiones. En caso de no tenerla, se halla un resumen alternativo mediante IFR, presentando una única devolución.

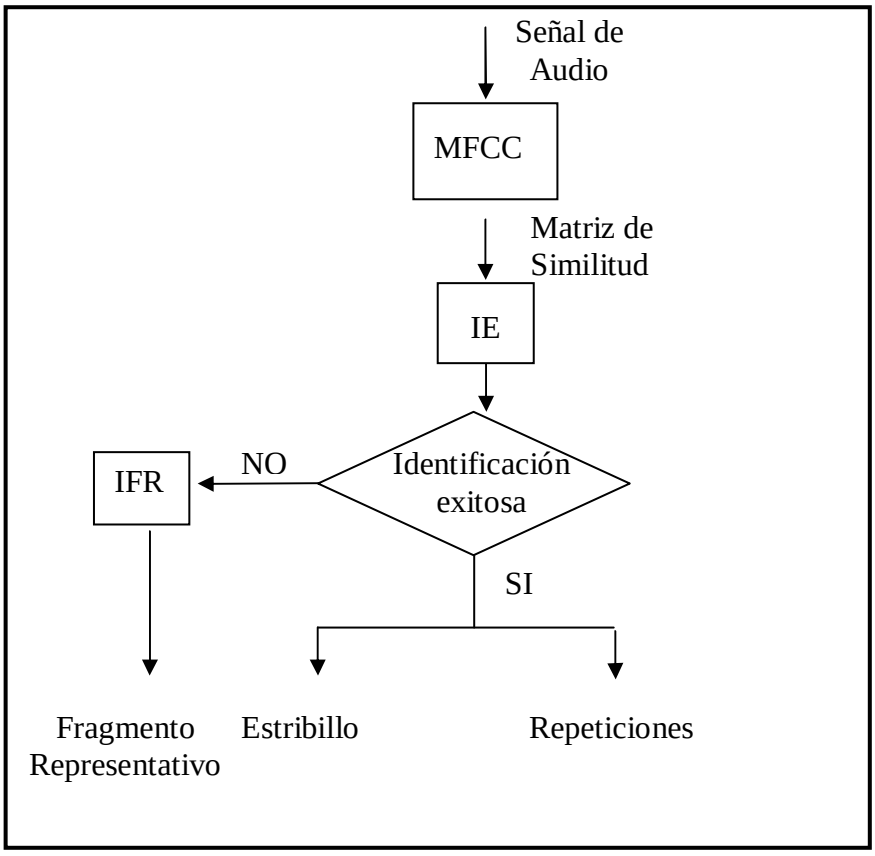


Figura 8-16: Diseño del sistema

9 Implementación

9.1 Resumen

El presente capítulo describe la implementación de la herramienta de identificación automática de resumen en canciones (IARC), desarrollada según los análisis descriptos en el capítulo 8.

9.2 Introducción

Los resultados obtenidos en el capítulo anterior definieron los parámetros y métodos a utilizar en la herramienta de identificación automática de resumen en canciones (IARC).

La herramienta fue desarrollada completamente en MATLAB e incluye una interfaz gráfica. Si bien sabemos que un programa en MATLAB restringe su utilización por parte del público, esto se debió a que se optó por dar prioridad al perfeccionamiento del algoritmo que a la aplicabilidad. La implementación de una aplicación no dependiente de MATLAB, e independiente de la plataforma, queda como tarea a futuro. Igualmente se comenzó la programación en C++, incluyendo los procedimientos de extracción de características hasta la conformación de la *Matriz de Similitud.*

9.3 Implementación en MATLAB

La implementación se adaptó para archivos de audio en formato MP3 por ser uno de los más difundidos actualmente. Se realiza una descompresión previa del archivo mediante el programa *mpg123.exe*, llamado desde MATLAB, lo que permite trabajar con la señal PCM.

Para reducir el costo computacional se decide trabajar con canciones monofónicas muestreadas a 16 kHz y 56 kbps; de todas formas el sistema admite archivos muestreados a una frecuencia superior y/o estéreo (dos canales de audio). En este último caso, las señales son submuestrads y convertidas a señales monofónicas (promediando ambos canales).

Para la extracción de características con *MFCC* se emplea la función *ma_mfcc.m* del Ma Toolbox [24]. El resto del procesamiento se realiza con funciones estándar de MATLAB.

Interfaz gráfica

La interfaz gráfica desarrollada en MATLAB procura acercar a quienes no están familiarizados con el programa. Se puede ver su presentación en la Figura 9-1.

Por medio de la interfaz es posible seleccionar la canción a procesar (en formato MP3) y extraer automáticamente su resumen.

Finalizado el procesamiento e identificado el resumen, este se reproduce de forma automática. Igualmente existe la posibilidad de volver a reproducirlo, así como sus repeticiones, y el resumen alternativo.

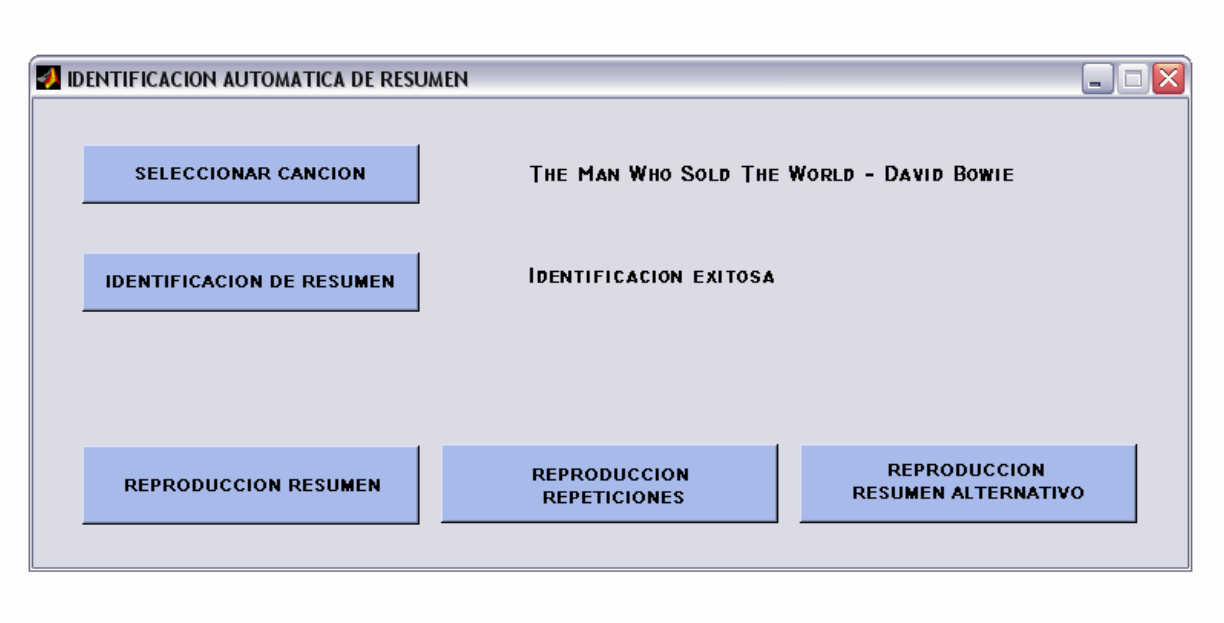


Figura 9-1: Interfaz gráfica de la herramienta en MATLAB.

9.4 Implementación en C++

La implementación en el lenguaje C++ elaborada hasta el momento, permite la creación de la matriz de similitud mediante *MFCC* de un archivo de audio. El desarrollo de la etapa de análisis de la matriz de similitud, queda como trabajo a futuro.

Para levantar el archivo y descomprimirlo se utilizó la librería de audio *BASS* (versión 2.2.0.4) para Windows y MAC OSX [25]. Trabaja con archivos MP3, MP2, MP1, OGG, WAV, AIF y otros, mediante la adición de complementos. Esta librería también reproduce los archivos por lo tanto se podrá emplear en la futura implementación de la segunda etapa.

Gran parte del procesamiento de los datos se realiza mediante un conjunto de clases de *NewMat C++ Matrix Library* [26], también empleada para realizar la Transformada de Fourier.

10 Validación

10.1 Resumen

Una vez finalizada la etapa de diseño, se procedió a validar la herramienta por medio de una nueva base de canciones (*conjunto de validación*[21]). Este capítulo está dedicado a presentar los resultados obtenidos considerando los criterios de evaluación ya vistos en el capítulo 8.

10.2 Resultados

Los resultados corresponden al desempeño de la herramienta tal como se implementó finalmente, mediante el método de extracción de características *MFCC*. Como se recordará, en una primera instancia se aplica la Identificación de estribillo (IE) y de no lograrse la identificación con éste, recurre a la Identificación de fragmento representativo (IFR).

Las tablas y gráficas muestran los porcentajes de canciones que presentan una validación exitosa según cada criterio. Primeramente se compararan los resultados para la devolución del resumen de la canción y luego para todas sus repeticiones. Seguidamente se analizan los resultados, comprando con los obtenidos para el *conjunto de entrenamiento.*

Identificación de resumen

COBERTURA	POCENTAJE CANCIONES
MAYOR A 70	82
MENOR A 70	6
FALSO	12

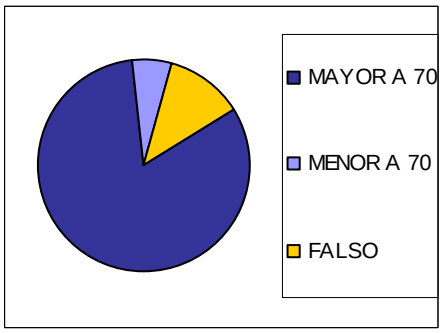


Tabla 10-1: Porcentaje de canciones según cobertura.

[21] Se explica en 8.2.

COMBINACIÓN	POCENTAJE CANCIONES
70/50	68
OTROS	20
FALSO	12

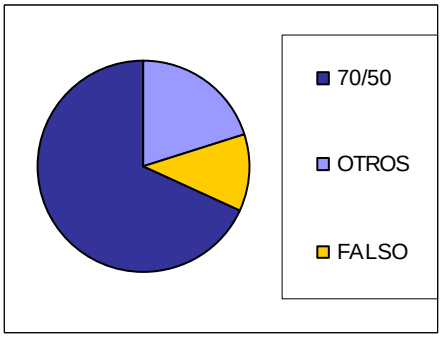


Tabla 10-2: Porcentaje de canciones según combinación.

MEDIDA F1	POCENTAJE CANCIONES
MAYOR A 70	64
MENOR A 70	24
FALSO	12

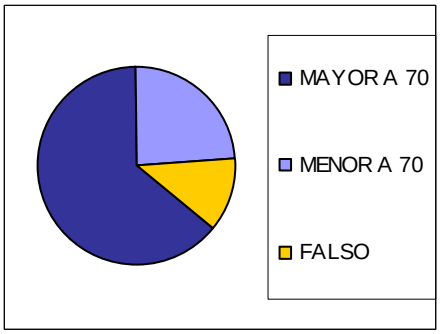


Tabla 10-3: Porcentaje de canciones según medida F1.

Identificación de las repeticiones del estribillo

COBERTURA	POCENTAJE CANCIONES
MAYOR A 70	78
MENOR A 70	10
FALSO	12
NINGUNO	0

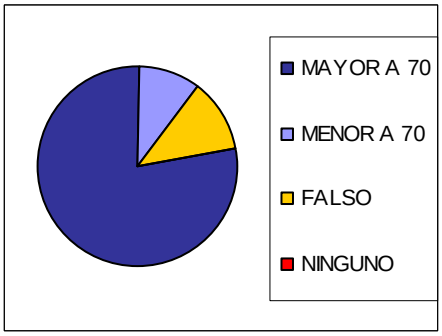


Tabla 10-4: Porcentaje de canciones según cobertura.

COMBINACIÓN	POCENTAJE CANCIONES
70/50	66
OTROS	22
FALSO	12
NINGUNO	0

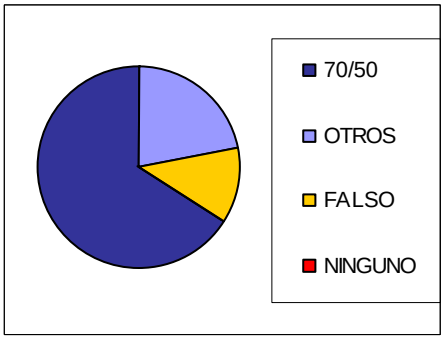


Tabla 10-5: Porcentaje de canciones según combinación.

MEDIDA F1	POCENTAJE CANCIONES
MAYOR A 70	56
MENOR A 70	32
FALSO	12
NINGUNO	0

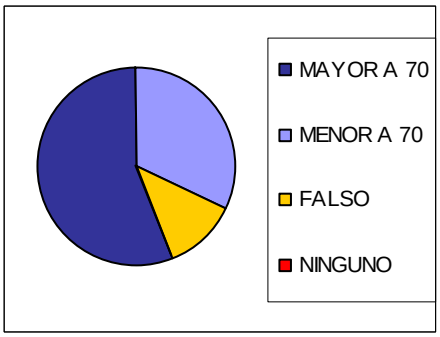


Tabla 10-6: Porcentaje de canciones según medida F1.

10.3 Conclusiones

En un análisis general, se destaca el buen desempeño de la herramienta desarrollada. Tomando como referencia la cobertura, no se evidencia gran diferencia entre los resultados para resumen y repeticiones. Por lo tanto se infiere que en aquellas canciones en que se logra una identificación exitosa del resumen, es factible que las repeticiones se encuentren también.

Lo mismo se deriva del análisis de las respectivas precisiones (criterios de combinación y F1). Es decir, de hallarse un resumen válido por la combinación, se espera que las respectivas repeticiones sean tanto acertadas (cobertura) como precisas.

Cabe destacar que si bien en la herramienta final se aplica el método IFR, éste no fue necesario, por lo que no influencia el análisis. Puede por tanto inferirse que la comparación entre el *conjunto de entrenamiento* y el *conjunto de validación* se realizará para idénticas condiciones de procesamiento.

A partir de los resultados anteriores para cobertura y los analizados para el *conjunto de entrenamiento*[22] se desprende que los porcentajes de éxito en la identificación de resumen son similares (84% frente a el 82% actual). Esto verifica en principio la independencia de la base de datos con respecto a los parámetros.

Los análisis para la combinación de cobertura y precisión, y F1, evidencian mínimas diferencias (por ejemplo en combinación un 73 contra el 68% en el actual). En las repeticiones, por el contrario, se obtuvieron para este grupo mejores resultados que para el de *entrenamiento*. Debido a lo limitado de las bases de canciones utilizadas, estas diferencias no pueden tomarse como concluyentes.

Para ambos grupos, los porcentajes de éxitos alcanzados permiten concluir que el sistema implementado es aplicable a cualquier conjunto de canciones que cumplan los requerimientos antes establecidos.

[22] Sección 8.8

11 Conclusiones Generales

11.1 Resumen

En este capítulo se expone el análisis global del trabajo así como las posibles mejoras a realizar.

11.2 Conclusiones generales

Como primera conclusión se observa que los objetivos iniciales del proyecto fueron cumplidos satisfactoriamente. La herramienta implementada identifica y reproduce el resumen de la canción, y en los casos que posean estribillo, sus repeticiones.

El porcentaje de detecciones exitosas, 82%, se considera un resultado muy bueno, que avala la configuración elegida en la determinación de los numerosos parámetros analizados.

Al enfrentarse a la identificación del resumen, se observó la complejidad del universo de canciones. Debido a esto debió estudiarse previamente las patologías presentes en el conjunto analizado para posteriormente resolverlas.

Una de las limitantes observadas del sistema es que al enfrentarse a canciones con más de un estribillo posible, elige uno de ellos y descarta el otro. Esa información ignorada es en algunos casos un estribillo considerado válido y aunque la herramienta lo detecta, no se utiliza. Una alternativa para solucionar este problema es entregar más de una opción de resumen, y que el usuario decida cuál es la correcta.

Respecto al método IE, se subraya que se optimizó para la devolución del resumen y no de sus repeticiones. De todas formas se obtiene una noción de la estructura con las repeticiones encontradas, pero se debería profundizar en el análisis de la detección desde este enfoque.

La utilización del método IFR brindó robustez a la herramienta, ya que permite que siempre se entregue un resumen, solucionando los casos en que el método IE no logre identificarlo. Sin embargo, en su implementación actual, no se detectan repeticiones.

11.3 Mejoras a futuro

Las mejoras están enfocadas al perfeccionamiento de los algoritmos, a mejorar su calibración y a considerar contextos más diversos para su aplicación.

La música polifónica presenta la dificultad de estar compuesta por la superposición de distintas fuentes, por lo que el análisis de cada una de ellas por separado podría mejorar el desempeño. En la búsqueda de repeticiones del estribillo, se cumplen las hipótesis antes nombradas de que se mantiene la melodía y la letra. Debido a lo anterior, logrando la separación de dichas fuentes, permitiría buscar la similitud solamente en la voz, sin la influencia de los instrumentos.

La complejidad de las técnicas de separación de fuentes de audio, implican una investigación aparte, quedando este estudio como una optimización pendiente.

Con respecto al procedimiento utilizado, en la determinación de los parámetros de *MFCC* se realizaron numerosas comparaciones entre las distintas configuraciones. Sin embargo todavía quedan parámetros por analizar, como la forma de los filtros utilizados, el espaciamiento entre ellos, etc. La medida de distancia para calcular la matriz de similitud es otro parámetro a investigar, ya que se mantuvo la usada en la bibliografía consultada tanto en *MFCC* como en VC.

En este trabajo se utilizó la matriz de similitud propia, con la que se compararon instantes de tiempo de una misma canción. En un futuro la matriz de similitud podría emplearse para la comparación de estribillos de distintas canciones y de esa forma encontrar temas similares. Para esto se pueden considerar distintos criterios a analizar, como el género musical, el ritmo, artista, etc.

Anteriormente se indicó que ya en la actualidad existen grupos de investigación que trabajan sobre el audio comprimido (MP3), adaptando a los parámetros de compresión técnicas ya aplicadas (*MFCC*). De lograrse esta implementación se ahorraría el costo computacional de descomprimir el archivo.

Puede esgrimirse que las técnicas aplicadas en el presente trabajo para el tratamiento del audio y el posterior procesamiento de imágenes, pueden ser aprovechadas para diversas aplicaciones. Resta por tanto profundizar en cada una, a fin de lograr las adaptaciones necesarias.

APÉNDICES

A.Base de canciones.

a. Conjunto de entrenamiento de 100 canciones

About A Girl – Nirvana
Aeroplane - Red hot chilli peppers
All because of you - U2
All you need is love - The Beatles
Amnesia - Chumbawamba
Antes - Jorge Drexler
Arranca Corazones - Ataque 77
Brimful of Asha - Cornershop (Remix)
Bring Me To Life - Evanescence
California Girls - Beach Boys
Cant stand losing you - The Police
Cant take my eyes off you - David Williams
Caress Me Down - Sublime
Come As You Are - Nirvana
Creep - Radiohead
Crua Chan - Sumo
Diez años despues – Los Rodriguez
Don´t Call Me White - NOFX
Don´t Know Why - Norah Jones
Don´t speak - No Doubt
Dyer Maker - Sheryl Crow
El Hombre del Piano - Ana Belen
Every Morning – Sugar Ray
Fast Car - Tracy Chapman
Feelin´ the Same Way - Norah Jones
Fly Away - Lenny Kravitz
Friday - The Cure
Guantanamera - Wyclef Jean feat. Lauryn Hill & Celia Cruz
Head over feet - Alanis Morisette
Hello, I love you - The Doors
Hey Ya - Outkast
Hollywood - Madonna
Home sweet home – Los Pericos
House - Alanis Morisette
I said a little pray for you - Aretha Franklin
I Think I´m Paranoid - Garbage
I Wanna Be Sedated - The Ramons
I Will Survive - Gloria Gaynor
Imagine - John Lennon
In The Cold, Cold, Night - The White Stripes
In the Sun - Blondie
Into the Deep - Kula Shaker
Lagrimas de Diamante - Paulinho Moska
Last Nite - The Strokes

Light my fire - The Doors
Like A Prayer - Madonna
Like a virgin - Madonna
Linger - The Cranberries
Loosing my religion - REM
Mamma mia - Abba
Mary Popins y el deshollinador - Fabiana Cantilo
Mi perro dinamita - Los Redondos
My immortal - Evanescence
No Rain - Blind Melon
Nobody Knows You When You´re Down & Out - Eric Clapton
NoMan´sWoman - Sinead O´Connor
Northern star - Hole
Nothing Else Matters - Metallica
Oleada - Julieta Venegas
Palabras mas palabras menos - Los Rodriguez
Reggeaton - Daddy Yankee ft Pitbull - Gasolina(remix)
Rock Around The Clock - Elvis Presley
Santeria - Sublime
Se me acabo la fuerza - Mana
Shes got her ticket - Tracy Chapman
Shinny Happy People - REM
Should I Stay or Should I Go - The Clash
Sleep to dreem - Fiona Apple
So quiet - Bjork
Solo le Pido a Dios - Ana Belen
Start me up - The Rolling Stones
Strong enough - Sheryl Crow
Summer of 69 - Bryan Adams
Sweet Baby - Macy Gray
Sweet Home Alabama - Urge Overkill
Tearjerker - Red hot chilli peppers
Tequila Sunrise - Cypress Hill
Thank u - Alanis Morisette
Thank You - Dido
The Man Who Sold The World - Nirvana
The New Pollution - Beck
There she goes - Bob Marley
Time - Blind Melon
Torn - Natalie Imbruglia
Twist and Shout - The Beatles
Un pacto para vivir - Bersuit Vergarabat
Universe - Blur
Ve con El - El cuarteto de nos
Vertigo - U2
Vira Vira - Mamonas asesinas
Vuelve a ti - Miranda
Waiting For Your Love - Los Pericos
When I Grow Up - Garbage
Where is the love - Black Eyed Peas

Where It´s At - Beck
Wind of Change - Scorpions
Woo hoo - Blur
X Offender - Blondie
Yesterday - The Beatles
You gotta be - Des´ree

b. Conjunto de validación de 50 canciones

American Woman - Lenny Kravitz
Back In Black - ACDC
Billie Jean - Michael Jackson
Black Dog - Led Zeppelin
California Dreamin - The Mamas And The Papas
Californication - Red Hot Chilli Peppers
Candombe de la aduana - Niquel
Candombe para Gardel - Ruben Rada
Como brillaba tu alma - No Te Va Gustar
Cotton Fields - Creedence
Could You Be Loved - Bob Marley
Cry Baby - Janis Joplin
De Musica Ligera - Soda Estereo
Devolve la bolsa - Bersuit Vergarabat
Dyer Maker - Led Zeppelin
El Dia Que me quieras - Carlos Gardel
Everybody Hurts - R.E.M.
Funky Town - Boney M
Girl, youll be a woman soon - Urge Overkill
I try - Macy Gray
In my place - ColdPlay
In The Name of Love - U2
La Flaca - Jarabe de Palo
Lança Perfume - Rita Lee
Living next door to Alice - Smokie
Maria - Blondie
Money For Nothing - Dire Straits
Mujer amante - Rata Blanca
Need You Tonight - INXS
Nothing Compares To You - Sinead O Connor
Ojalá - Silvio Rodriguez
Ojos Así - Shakira
Por Dentro - La Vela Puerca
Proud Mary - Creedence
Que ves - Divididos
Rasguña Las Piedras - Sui Generis
Rivers Of Babylon - Boney M
Satisfaction - The Rolling Stones
Sin Documentos - Los Rodriguez
Spinning around - Kyllie Minogue
Stereotypes - Blur
Sweet Child of mine - Guns n Roses
Te quiero asi - Chichi Peralta
Tears in Heaven - Eric Clapton
The Great Pretender - Platers

B.Resultados de la encuesta realizada sobre la detección manual de estribillos.

Esta encuesta se realizó para verificar que los estribillos extraídos manualmente para la validación de los métodos coincidieran con las opiniones de un grupo 20 personas.

Cada persona encuestada escuchó los estribillos detectados manualmente de la base de 50 canciones. La encuesta no fue realizada con las canciones completas porque hubiera demandado demasiado tiempo de los encuestados; de todas formas estuvieron a su disposición en caso de que las requirieran.

Se clasificaron los fragmentos de la siguiente forma:

1. Estribillo correcto
2. Estribillo correcto pero demasiado corto
3. Estribillo correcto pero demasiado largo
4. Estribillo dudoso
5. Estribillo incorrecto
6. La canción no tiene estribillo

	1	2	3	4	5	6
American Woman - Lenny Kravitz	9	7		4		
Back In Black - ACDC	11		6			3
Billie Jean - Michael Jackson	18		2			
Black Dog - Led Zeppelin	12			2	2	4
California Dreamin - The Mamas And The Papas	20					
Californication - Red Hot Chilli Peppers	20					
Candombe de la aduana - Niquel	20					
Candombe para Gardel - Ruben Rada	20					
Como brillaba tu alma - No Te Va Gustar	20					
Cotton Fields - Creedence	13		4	3		
Could You Be Loved - Bob Marley	20					
Cry Baby - Janis Joplin	20					
De Musica Ligera - Soda Estereo	20					
Devolve la bolsa - Bersuit Vergarabat	20					
Dyer Maker - Led Zeppelin	20					
El Dia Que me quieras - Carlos Gardel	16		1			3
Everybody Hurts - R.E.M.	15	4		1		
Funky Town - Boney M	20					
Girl, youll be a woman soon - Urge Overkill	20					
I try - Macy Gray	20					
In my place - ColdPlay	20					
In The Name of Love - U2	20					
La Flaca - Jarabe de Palo	20					

Lança Perfume - Rita Lee	20					
Living next door to Alice - Smokie	20					
Maria - Blondie	20					
Money For Nothing - Dire Straits	20					
Mujer amante - Rata Blanca	20					
Need You Tonight - INXS	15			4	1	
Nothing Compares To You - Sinead O Connor	20					
Ojalá - Silvio Rodriguez	17	3				
Ojos Así - Shakira	20					
Por Dentro - La Vela Puerca	20					
Proud Mary - Creedence	13		7			
Que ves - Divididos	16		4			
Rasguña Las Piedras - Sui Generis	20					
Rivers Of Babylon - Boney M	20					
Satisfaction - The Rolling Stones	12	4		4		
Sin Documentos - Los Rodriguez	20					
Spinning around - Kyllie Minogue	20					
Stereotypes - Blur	20					
Sweet Child of mine - Guns n Roses	20					
Te quiero asi - Chichi Peralta	20					
Tears in Heaven - Eric Clapton	16			3		1
The Great Pretender - Platers	18				2	
The Man Who Sold The World – David Bowie	20					
Vasos Vacios - Los Fabulosos Cadillacs	20					
We are The Champions - Queen	20					
Wonderwall - Oasis	8		12			
Y nos dieron las 10 - Joaquín Sabina	20					

Los porcentajes de respuestas para las distintas clasificaciones se pueden ver en la Tabla B-1. Más de un 90% de respuestas coincidieron con la elección realizada al identificar el estribillo manualmente en las canciones del conjunto de validación.

Debido a esto se puede concluir que el método de evaluación está efectivamente analizando si lo que detecta la herramienta es el fragmento memorable de la canción.

1	90,90%
2	1,80%
3	3,60%
4	2,10%
5	0,50%
6	1,10%

Tabla B-1: Porcentajes de respuestas para las 6 clasificaciones.

C.Estadísticas

a. Estadísticas MFCC

Se presentan a continuación las distintas configuraciones de parámetros analizadas y las estadísticas obtenidas según los criterios de la sección 8.4.

	Rango de frecuencias	N_{mfcc}	Uso del 1er coeficiente	Datos ventana
Opción 1	20 a 3000	12	Si	1s, solap 50%
Opción 2	20 a 3000	12	No	1s, solap 50%
Opción 3	20 a 3000	12	Si	2s, solap 50%
Opción 4	20 a 3000	12	Si	0.5s, solap 50%
Opción 5	20 a 3000	12	Si	0.2s, solap 50%
Opción 6	20 a 3000	6	Si	1s, solap 50%
Opción 7	20 a 3000	30	Si	1s, solap 50%
Opción 8	20 a 4000	12	Si	1s, solap 50%
Opción 9	20 a 8000	12	Si	1s, solap 50%
Opción 10	20 a 8000	12	No	1s, solap 50%

- **Resumen, análisis 1: cobertura**

La siguiente tabla muestra los porcentajes de canciones que tienen una cobertura del 100%, mayor al 90, 80 y 70%, así como también los casos que se encuentra un estribillo falso y los que no se encuentra ningún resumen (ninguno). Se analiza la cobertura promedio dentro de las canciones que encuentran algo del estribillo real.

PORCENTAJE COBERTURA	op1	op2	op3	Op4	op5	op6	op7	op8	op9	op10
100	56	39	56	48	39	38	49	62	57	42
MAYOR A 90	74	60	66	67	51	53	69	74	72	59
MAYOR A 80	81	65	72	73	57	61	74	80	79	63
MAYOR A 70	84	69	77	77	59	64	79	83	83	70
MENOR A 70	6	16	9	14	19	5	12	3	5	19
PROMEDIO	94	85	91	88	85	91	89	96	94	85
FALSO	7	14	13	6	13	9	8	7	8	7
NINGUNO	3	1	1	3	9	22	1	7	4	4

Tabla C-1: Análisis de cobertura de resumen para distintas opciones de MFCC con IE

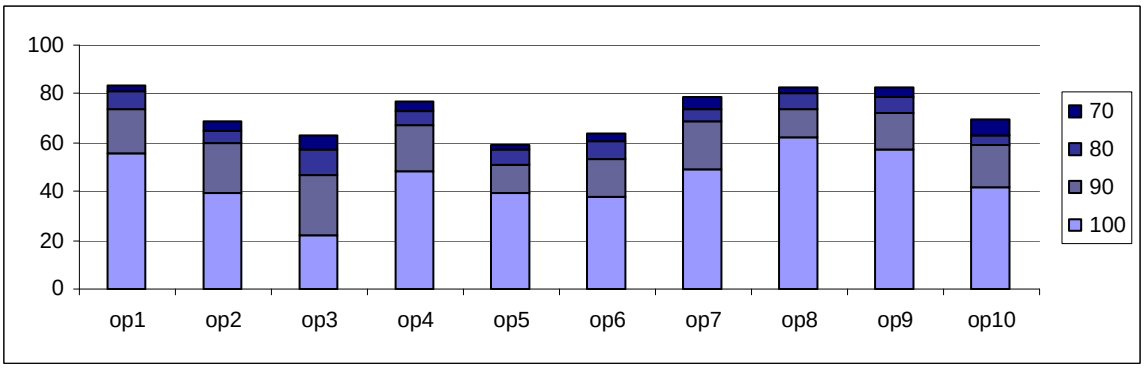


Figura C-1: Análisis de cobertura de resumen para distintas opciones de MFCC con IE

- **Resumen, análisis 2: combinación de cobertura y precisión**

Se analiza la combinación de un 70% de cobertura junto con una precisión del 70 y 50% para el resumen.

PORCENTAJE COMBINACIÓN	op1	op2	Op3	Op4	op5	Op6	op7	op8	op9	op10
70/70	47	37	18	34	19	28	36	47	37	40
70/50	60	45	49	45	27	51	51	60	55	50

Tabla C-2: Análisis de combinación de cobertura y precisión de resumen para distintas opciones de MFCC con IE

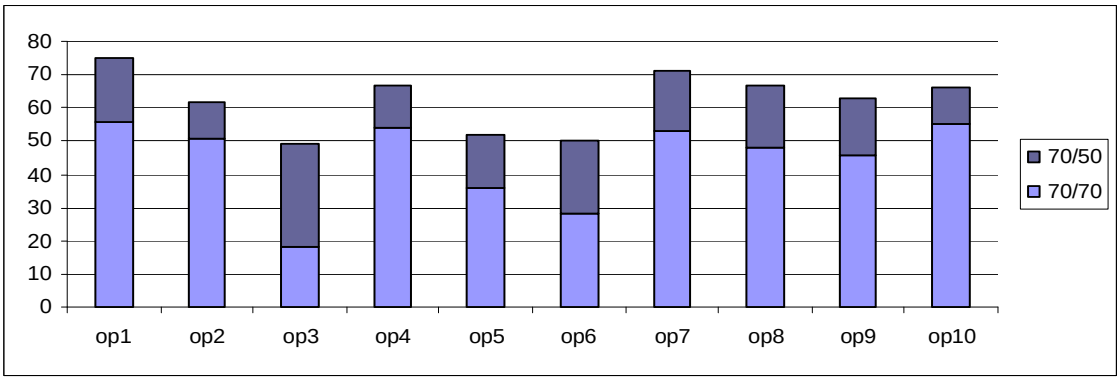


Figura C-2: Análisis de combinación de cobertura y precisión de resumen para distintas opciones de MFCC con IE

- **Repeticiones, análisis 1: cobertura**

Se presentan los porcentajes de canciones con una cobertura en las repeticiones del resumen encontrado del 100%, mayor al 90, 80 y 70%, así como también los casos que se encuentran todos los estribillos falsos y los que no se encuentra ningún resumen (ninguno). Se analiza la cobertura promedio del total de las repeticiones del resumen dentro de las canciones que encuentran algo del estribillo real.

PORCENTAJE COBERTURA	op1	op2	op3	Op4	op5	op6	op7	op8	op9	op10
100	24	12	22	10	9	23	21	26	24	20
MAYOR A 90	51	31	47	36	24	41	40	52	50	40
MAYOR A 80	63	46	57	47	31	51	49	61	59	46
MAYOR A 70	68	50	63	53	35	58	58	67	66	52
MENOR A 70	26	38	29	41	47	15	35	22	25	38
PROMEDIO	83	73	80	74	68	81	76	80	82	75
FALSO	3	11	7	3	9	5	6	4	5	6
NINGUNO	3	1	1	3	9	22	1	7	4	4

Tabla C-3: Análisis de cobertura de repeticiones para distintas opciones de MFCC con IE

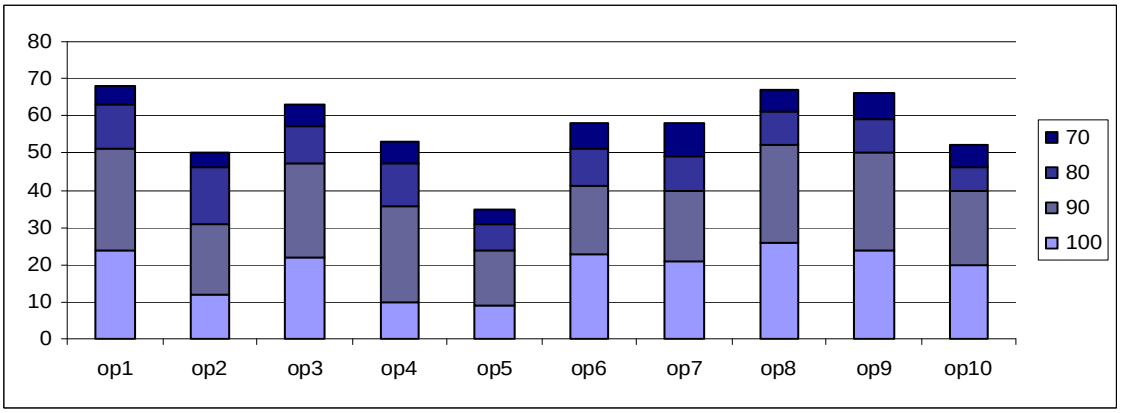


Figura C-3: Análisis de cobertura de repeticiones para distintas opciones de MFCC con IE

- **Repeticiones, análisis 2: combinación de cobertura y precisión**

Se analiza la combinación de un 70% de cobertura junto con una precisión del 70 y 50% para las repeticiones del resumen.

PORCENTAJE COMBINACIÓN	op1	Op2	op3	op4	op5	op6	op7	op8	op9	op10
70/70	47	37	25	34	19	28	36	47	37	40
70/50	60	45	45	45	27	51	51	60	55	50

Tabla C-4: Análisis de combinación de cobertura y precisión de repeticiones para distintas opciones de MFCC con IE

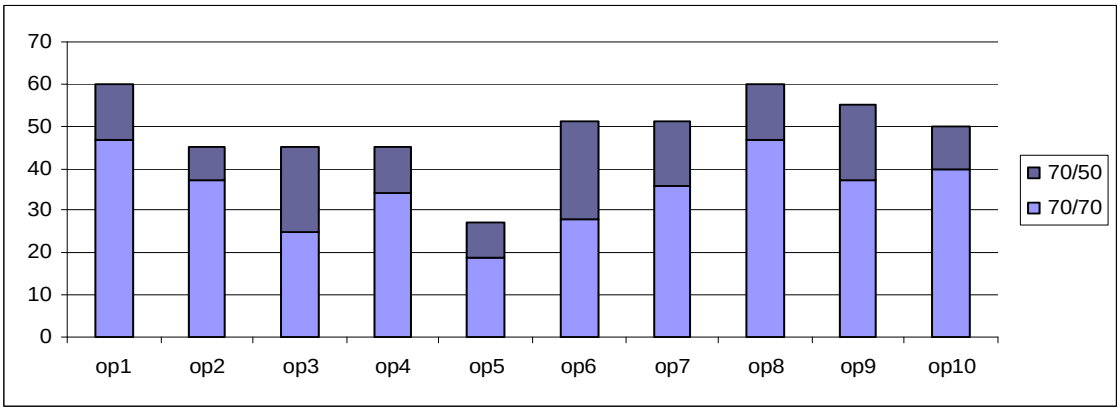


Figura C-4: Análisis de cobertura y precisión de repeticiones para distintas opciones de MFCC con IE

b. Estadísticas VC

	Rango de frecuencias	Datos ventana
Opción 1	130 a 8000	1s, solap 50%
Opción 2	130 a 4000	1s, solap 50%
Opción 3	130 a 8000	0.2s, solap 50%
Opción 4	130 a 8000	0.5s, solap 50%

- **Resumen, análisis 1: cobertura**

PORCENTAJE COBERTURA	op1	op2	op3	Op4
100	37	37	33	38
MAYOR A 90	47	45	41	50
MAYOR A 80	60	59	46	59
MAYOR A 70	67	67	54	61
MENOR A 70	15	13	23	18
PROMEDIO	84	86	77	84
FALSO	15	16	21	18
NINGUNO	3	4	2	3

Tabla C-5: Análisis de cobertura de resumen para distintas opciones de VC con IE

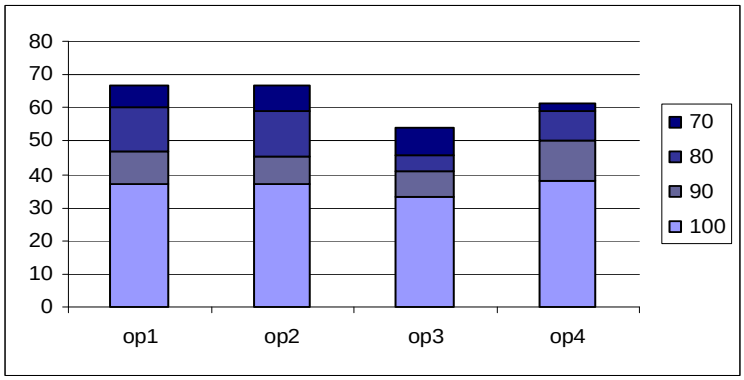


Figura C-5: Análisis de cobertura de resumen para distintas opciones de VC con IE

- **Resumen, análisis 2: combinación de cobertura y precisión**

PORCENTAJE COMBINACIÓN	op1	op2	op3	op4
70/70	45	46	29	36
70/50	55	56	43	53

Tabla C-6: Análisis de combinación de cobertura y precisión de resumen para distintas opciones de VC con IE

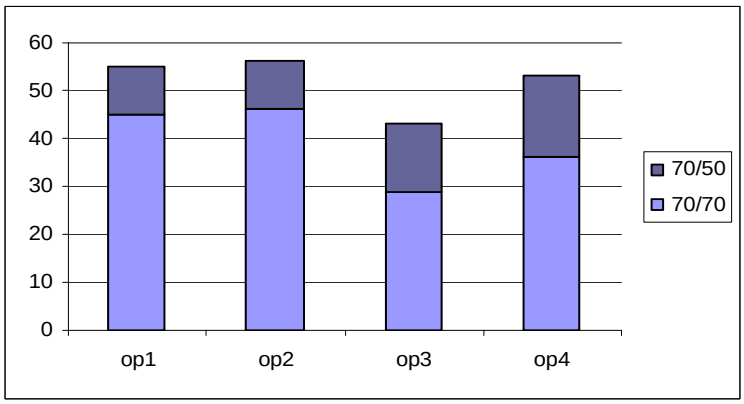


Figura C-6: Análisis de combinación de cobertura y precisión de resumen para distintas opciones de VC con IE

- **Repeticiones, análisis 1: cobertura**

PORCENTAJE COBERTURA	op1	op2	op3	Op4
100	13	10	1	9
MAYOR A 90	28	24	10	19
MAYOR A 80	36	39	17	32
MAYOR A 70	45	48	24	38
MENOR A 70	42	38	60	49
PROMEDIO	69	70	57	65
FALSO	10	10	14	10
NINGUNO	3	4	2	3

Tabla C-7: Análisis de cobertura de repeticiones para distintas opciones de VC con IE

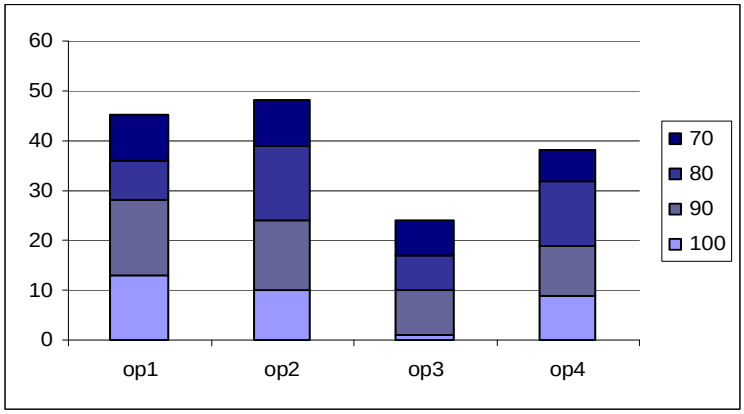


Figura C-7: Análisis de cobertura de repeticiones para distintas opciones de VC con IE

- **Repeticiones, análisis 2: combinación de cobertura y precisión**

PORCENTAJE COMBINACIÓN	op1	op2	op3	Op4
70/70	26	31	10	18
70/50	35	40	19	29

Tabla C-8: Análisis de combinación de cobertura y precisión de repeticiones para distintas opciones de VC con IE

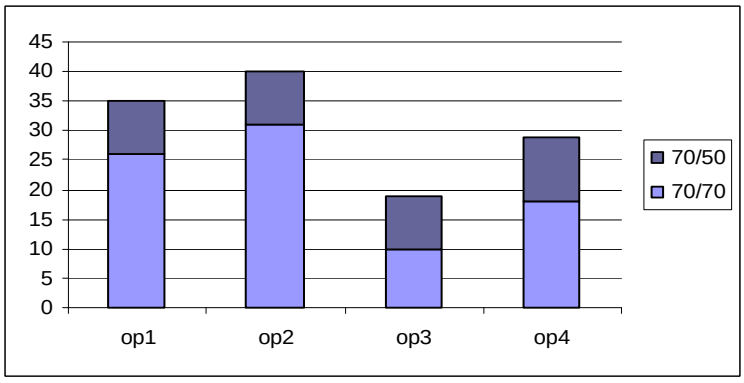


Figura C-8: Análisis de cobertura y precisión de repeticiones para distintas opciones de VC con IE

c. Estadísticas MFCC contra VC

- **Resumen, análisis 1: cobertura**

PORCENTAJE COBERTURA	MFCC	VC
100	56	37
MAYOR A 90	74	45
MAYOR A 80	81	59
MAYOR A 70	84	67
PROMEDIO	94	86
FALSO	7	16
NINGUNO	3	4

Tabla C-9: Análisis de cobertura de resumen para las configuraciones óptimas de MFCC vs. VC con IE

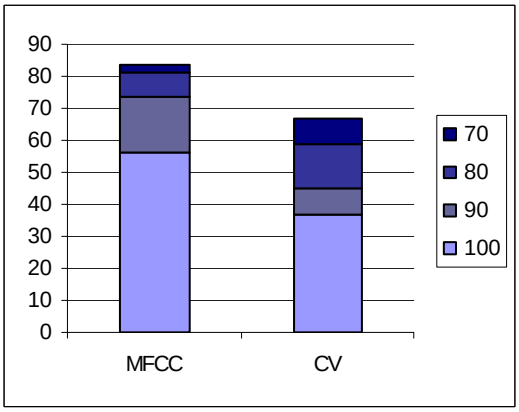


Figura C-9: Análisis de cobertura de resumen para las configuraciones óptimas de MFCC vs. VC con IE

- **Resumen, análisis 2: combinación de cobertura y precisión**

PORCENTAJE COMBINACIÓN	MFCC	VC
70/70	56	46
70/50	75	56

Tabla C-10: Análisis de combinación de cobertura y precisión de resumen para las configuraciones óptimas de MFCC vs. VC con IE

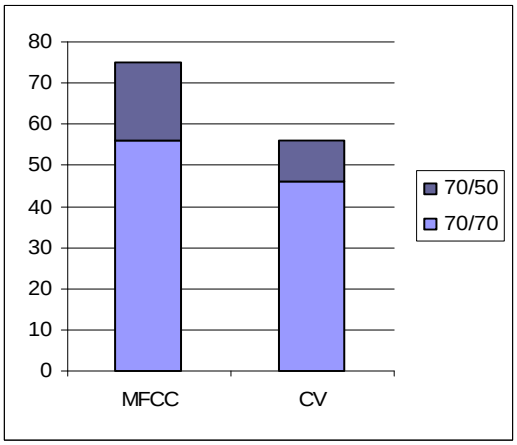


Figura C-10: Análisis de combinación de cobertura y precisión de resumen para las configuraciones óptimas de MFCC vs. VC con IE

- **Repeticiones, análisis 1: cobertura**

PORCENTAJE COBERTURA	MFCC	VC
100	24	10
MAYOR A 90	51	24
MAYOR A 80	63	39
MAYOR A 70	68	48
PROMEDIO	83	70
FALSO	3	10
NINGUNO	4	4

Tabla C-11: Análisis de cobertura de repeticiones para las configuraciones óptimas de MFCC vs. VC con IE

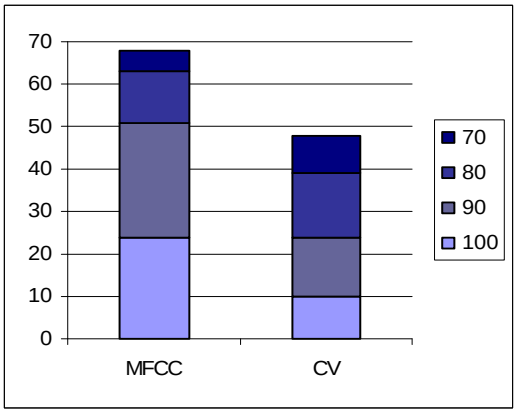


Figura C-11: Análisis de cobertura de repeticiones para las configuraciones óptimas de MFCC vs. VC con IE

- **Repeticiones, análisis 2: combinación de cobertura y precisión**

PORCENTAJE COMBINACIÓN	MFCC	VC
70/70	47	31
70/50	60	40

Tabla C-12: Análisis de combinación de cobertura y precisión de repeticiones para las configuraciones óptimas de MFCC vs. VC con IE

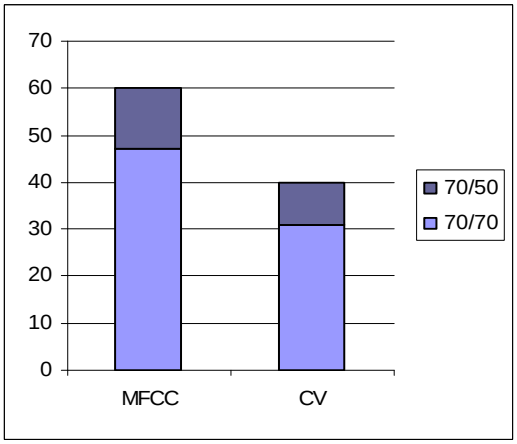


Figura C-12: Análisis de cobertura y precisión de repeticiones para las configuraciones óptimas de MFCC vs. VC con IE

d. Análisis de parámetros para IFR

Se analizan las siguientes opciones para determinar el largo óptimo y el vector de pesos.

	Largo	Vector de peso
Opción 1	20	15%
Opción 2	25	15%
Opción 3	30	15%
Opción 4	25	10%
Opción 5	30	20%

- **Resumen, análisis 1: cobertura**

PORCENTAJE COBERTURA	Op1	op2	op3	op4	op5
100	7	15	23	11	19
MAYOR A 90	12	21	30	19	23
MAYOR A 80	22	33	38	30	33
MAYOR A 70	27	43	50	64	64
PROMEDIO	60	68	71	66	67
FALSO	30	24	22	26	23

Tabla C-13: Análisis de cobertura para IFR

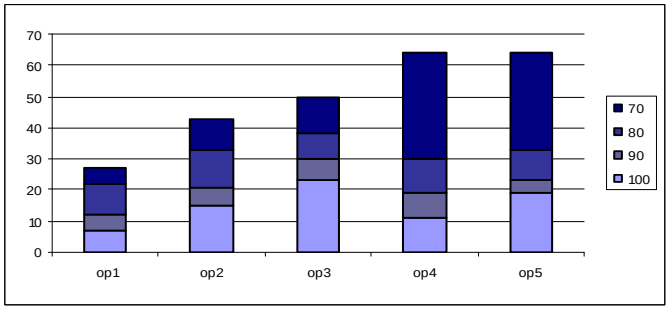


Figura C-13: Análisis de cobertura para IFR

- **Resumen, análisis 2: combinación de cobertura y precisión**

PORCENTAJE COMBINACIÓN	op1	op2	op3	op4	op5
70/70	20	22	14	12	12
70/50	24	37	31	33	26

Tabla C-14: Análisis de cobertura y precisión para IFR

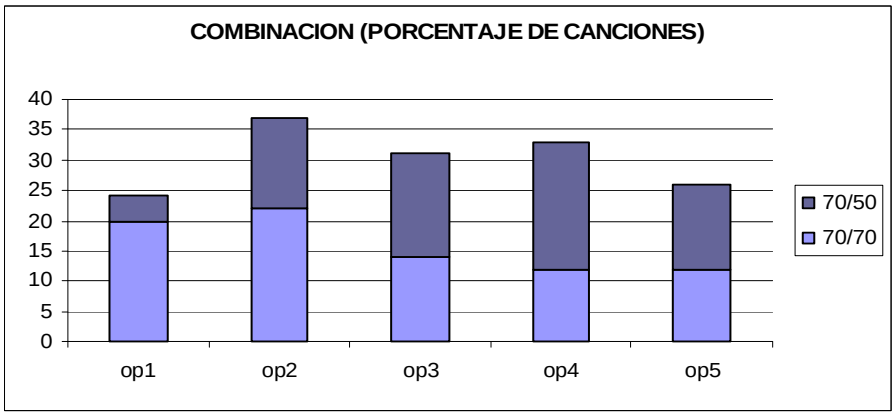


Figura C-14: Análisis de cobertura y precisión para IFR

e. Estadísticas de comparación de IE con IFR (MFCC óptima)

- **Resumen, análisis 1: cobertura**

PORCENTAJE COBERTURA	IE	IFR
100	56	11
MAYOR A 90	74	19
MAYOR A 80	81	30
MAYOR A 70	84	64
PROMEDIO	94	66
FALSO	7	24
NINGUNO	3	0

Tabla C-15: Análisis de cobertura en resumen de IE vs. IFR

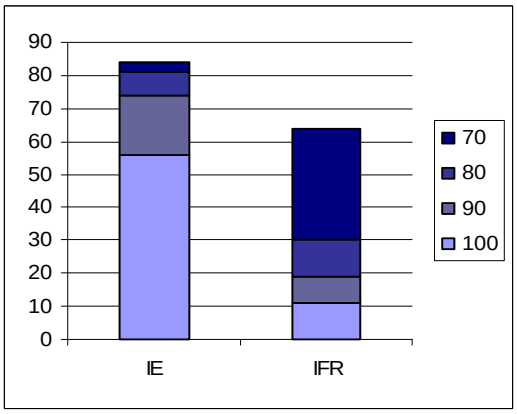


Figura C-15: Análisis de cobertura en resumen de IE vs. IFR

- **Resumen, análisis 2: combinación de cobertura y precisión**

PORCENTAJE COMBINACIÓN	IE	IFR
70/70	56	12
70/50	75	33

Tabla C-16: Análisis de combinación de cobertura y precisión en resumen de IE vs. IFR

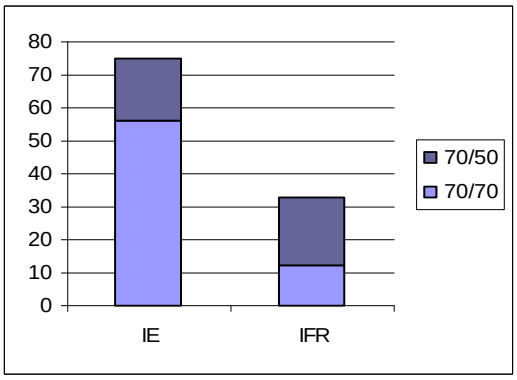


Figura C-16: Análisis de combinación de cobertura y precisión en resumen de IE vs. IFR

D.Procesamiento de Imágenes por filtros espaciales (convolución).

a. Filtrado espacial

Los filtros espaciales tienen como objetivo modificar la contribución de determinados rangos de frecuencias a la formación de la imagen. El término espacial se refiere al hecho de que el filtro se aplica directamente a la imagen y no a una transformada de la misma, es decir, el nivel de gris de un píxel se obtiene directamente en función del valor de sus vecinos [23].

Los filtros espaciales pueden clasificarse basándose en su linealidad: *filtros lineales* **y** *filtros no lineales***. A su vez los** *filtros lineales* **pueden clasificarse según las frecuencias** que dejen pasar: los ***filtros paso bajo*** atenúan o eliminan las componentes de alta frecuencia a la vez que dejan inalteradas las bajas frecuencias; los ***filtros paso alto*** atenúan o eliminan las componentes de baja frecuencia con lo que agudizan las componentes de alta frecuencia; los ***filtros paso banda*** eliminan regiones elegidas de frecuencias intermedias.

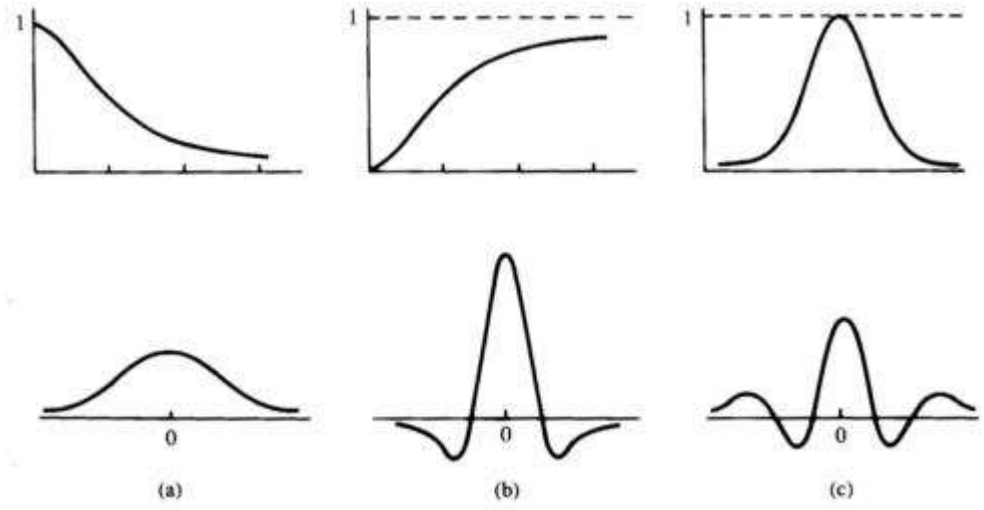


Figura D-1: Arriba: secciones de filtros en frecuencia con simetría circular. Abajo: secciones correspondientes a filtros espaciales. (a) Filtro paso bajo. (b) Filtro paso alto. (c) Filtro paso banda.

La forma de operar de los filtros lineales es por medio de la utilización de máscaras que recorren toda la imagen centrando las operaciones sobre los píxeles que se encuadran en la región de la imagen original que coincide con la máscara y el resultado se obtiene mediante una computación (suma de convolución) entre los píxeles originales y los diferentes coeficientes de las máscaras.

Los ***filtros espaciales no lineales*** también operan sobre entornos. Sin embargo, su operación se basa directamente en los valores de los píxeles en el entorno en consideración. Unos ejemplos de filtros no lineales habituales son los filtros mínimo, máximo y de mediana que son conocidos como ***filtros de rango***. El filtro de mediana tiene un efecto de difuminado de la imagen, y permite realizar una eliminación de ruido de forma eficaz, mientras que el filtro de máximo se emplea para buscar los puntos más brillantes de una imagen produciendo un efecto de ***erosión***, y el filtro de mínimo se emplea con el objetivo contrario, buscar los puntos más oscuros de una imagen produciendo un efecto de ***dilatación***.

Otra clasificación de los filtros espaciales puede hacerse basándose en su finalidad, y así tenemos los ***filtros de realce*** (Sharpening) para eliminar zonas borrosas o ***filtros de suavizado*** (Smoothing) para difuminar la imagen. también tenemos los filtros diferenciales que se componen de varios tipos de máscaras (Laplaciano, Prewitt, Sobel, etc.), y se utilizan para la ***detección de bordes***. El proceso de detección de bordes se basa en realizar un incremento del contraste en las zonas donde hay una mayor diferencia entre las intensidades, y en una reducción de éste donde no tenemos variación de intensidad.

b. Convolución

El tratamiento de imágenes más empleado y conocido, es el tratamiento espacial también conocido como ***convolución***. Las convoluciones discretas son muy usadas en el procesado de imagen para el suavizado de imágenes, el afilado de imágenes, detección de bordes, detección de ciertas estructuras y otros efectos. Mediante este proceso se calcula el valor de un determinado punto en función de su valor y del valor de los puntos que le rodean, aplicando una simple operación matemática en función de la cual se obtendrá un valor resultante para el punto en cuestión.

La operación de la convolución puede representarse como la siguiente operación:

$$g(x,y) = f(x,y) * a(x,y)$$

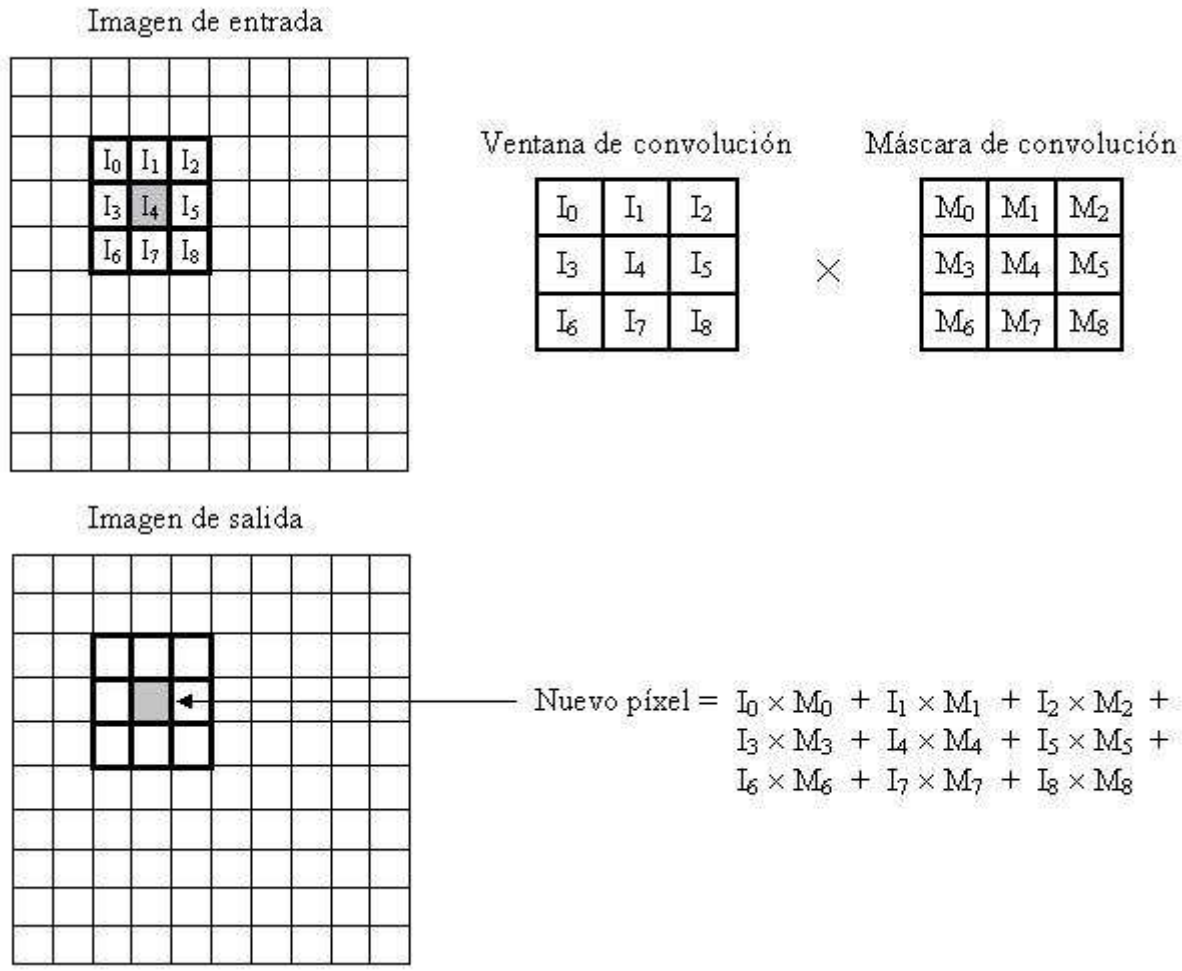
donde ***a(x,y)*** es la función respuesta al impulso del filtro a aplicar, ***f(x,y)*** es la imagen de entrada y ***g(x,y)*** es la imagen filtrada. Las expresiones matemáticas para el caso bidimensional son las siguientes:

Caso continuo $$g(x,y) = f(x,y) \otimes a(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\chi,\zeta)\, a(x-\chi, y-\zeta)\, d\chi\, d\zeta$$

Caso discreto $$g[m,n] = f[m,n] \otimes a[m,n] = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} f[j,k]\, a[m-j, n-k]$$

La operación matemática en que consiste la convolución es simplemente una suma ponderada de píxeles en el vecindario del píxel fuente. Los pesos son determinados por una pequeña matriz llamada ***máscara de convolución*** o ***elemento estructurante***, que determina los coeficientes a aplicar sobre los puntos de una determinada área. La posición del valor central de la matriz se corresponde con la posición del píxel de salida.

Una ventana deslizante, llamada ***ventana de convolución***, se centra en cada píxel de una imagen de entrada y genera nuevos píxeles de salida. Para aplicar la máscara a esa zona se multiplican los valores de los puntos que rodean al píxel que estamos tratando por su correspondiente entrada o coeficiente en la máscara y luego se suman esos productos. El resultado es el nuevo valor para el píxel central, tal y como se puede ver en el siguiente ejemplo para un elemento estructurante cuadrado.



E.Medida F1

Una medida de validación extensamente empleada es la "medida F_1". Fue introducida por C.J. Rijsbergen y establece la integración de la precisión (p) y cobertura (c) de los resultados. Su definición es:

$$F_1 = \frac{2 \cdot c \cdot p}{c + p}$$

Los términos de precisión y cobertura representan a distintas variables según la aplicación, de forma general se refieren a:

- Cobertura: habilidad de un sistema de presentar todos los resultados relevantes.
- Precisión: habilidad de un sistema de presentar únicamente resultados relevantes.

La medida F_1 demuestra ser un promedio armónico (H), cuya definición para n muestras $x_1 \mathrm{K}\ x_n$ es:

$$\frac{1}{H} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{x_i}$$

Se aprecia que es el recíproco del promedio aritmético (promedio estándar). Aplicando lo anterior a la precisión (p) y cobertura (c), se tiene:

$$H = \frac{1}{\frac{1}{2}\left(\frac{1}{c} + \frac{1}{p}\right)} = \frac{2}{\frac{p}{c \cdot p} + \frac{c}{c \cdot p}} = \frac{2 \cdot c \cdot p}{c + p} = F_1$$

Puede aplicarse un vector de pesos al inverso de los valores, de modo de darle mayor influencia a alguna de las variables:

$$F_\beta = \frac{1}{\alpha \cdot \frac{1}{c} + (1-\alpha) \cdot \frac{1}{p}} = \frac{p \cdot c}{\alpha \cdot p + (1-\alpha) \cdot c} = \frac{\frac{1}{1-\alpha} \cdot p \cdot c}{\frac{\alpha}{1-\alpha} \cdot p + c} \quad \begin{matrix} = \\ 6\,4\,7\ 4\,8 \\ \beta^2 = \frac{\alpha}{1-\alpha} \end{matrix} \quad \frac{(\beta^2 + 1) \cdot p \cdot c}{\beta^2 \cdot p + c}$$

F.Contenido del CD

Los requerimientos para utilizar los archivos incluidos en el CD se detallan a continuación:

- **Sistema operativo:** Windows XP/2000/98/95
- **Memoria RAM:** 256 MB o superior
- **Lectora de CD**
- **Acrobat Reader**
- **MATLAB 5.3 o superior con toolbox de imágenes.**

Se encuentran las siguientes carpetas:

- **Base de canciones**: Contiene dos subcarpetas con las canciones del conjunto de entrenamiento y las del conjunto de validación respectivamente. Además se incluye un archivo en excel con los tiempos de inicio y fin de los estribillos identificados manualmente.

- **Documentación:** En un archivo PDF se puede encontrar el presente trabajo junto con una carpeta conteniendo los archivos de las referencias bibliográficas.

- **Programas:** Contiene tres subcarpetas:

    - IARC INTERFAZ GRÁFICA MATLAB: Son los códigos en MATLAB de la implementación de la herramienta completa tal como se definió finalmente. Para utilizarlos se debe copiar la carpeta a cualquier directorio y desde allí, correr en MATLAB la función resumen.m.

    - IARC INTERFAZ GRÁFICA MATLAB (utilizando VC): Se pueden encontrar los códigos en MATLAB de la herramienta implementada con el método de extracción de características Vectores de Croma. La forma de utilizar estos códigos es igual a la del punto anterior.

    - IARC IMPLEMENTACIÓN C++: Se encuentran los códigos en C++. Incluyen el procesamiento de la señal de audio hasta calcular la matriz de similitud. El programa ejecutable es *estribillo.cpp*, primero crea un objeto de la clase *cancion* (archivos cancion.h y cancion.cpp) el cual contiene las muestras de la cancion. Para esto emplea la librería de audio BASS (versión 2.2.0.4). Luego el ejecutable crea un objeto de la clase mfcc (archivos mfcc.cpp y mfcc.h), para posteriormente crear la matriz de similitud. Las operaciones con matrices y la FFT se realizan con la librería NewMat C++ (versión 10).

Referencias

[1] Alan V. Oppenheim, Ronald W. Schafer. (1999) “Tratamiento de Señales en Tiempo Discreto”.

[2] Steven W. Smith. (1999) “The Scientist and Engineer's Guide to Digital Signal Processing”.
http://www.dspguide.com

[3] Malcom Slaney. (1998) “Auditory Toolbox: A MatlabToolbox for Auditory Modeling Work”.
http://rvl4.ecn.purdue.edu/~malcolm/interval/ 1998-010/AuditoryToolboxTechReport.pdf

[4] Terri Kamm, Hynek Hermansky, Andreas G. Andreou. (1997) “Learning the Mel-scale and Optimal VTN Mapping”.
http://www.clsp.jhu.edu/ws97/acoustic/reports/KHAMel.pdf

[5] Beth Logan. (2000) “Mel Frequency Cepstral Coefficients for Music Modeling”.
http://ciir.cs.umass.edu/music2000/papers/logan_abs.pdf

[6] Lie Lu, Muyuan Wang, Hong-Jiang Zhang. (2004) “Repeating Pattern Discovery and Structure Analysis from Acoustic Music Data”.
http://research.microsoft.com/asia/dload_files/group/mc/2004/4-MIR04_MusicStruc.pdf

[7] D. Van Steelant, B. De Baets, H. De Meyer, M. Leman, J. P. Martens, L. Clarisse, M. Lesaffre. (2002) “Discovering Structure and Repetition in Musical Audio”. http://www.ipem.ugent.be/MAMI/Public/Papers/VanSteelantetAlEurofuse2002.pdf

[8] Masataka Goto. (2003) “A Chorus-Section Detecting Method For Musical Audio Signals”. http:// staff.aist.go.jp/m.goto/PAPER/ICASSP2003goto.pdf

[9] Matthew Cooper, Jonathan Foote. (2002) “Automatic Music Summarization via Similarity Analysis”.
http://www.fxpal.com/people/cooper/Papers/ISMIR02-1.pdf

[10] Logan, B. and Chu, S. (2000) “Music Summary using Key Phrases”.
http://www.hpl.hp.com/techreports/Compaq-DEC/CRL-2000-1.pdf

[11] Chilton, Dr. E. (1999) “Speech Analisis”.

[12] Josep Martí Roca. (2003) “Situación Actual de las Tecnologías del Habla”.
http://www.ia.csic.es/Sea/publicaciones/4372kb003.pdf

[13] Federico Miyara. (1999) “La voz humana”.
http://www.eie.fceia.unr.edu.ar/~acustica/biblio/fonatori.pdf

[14] Adam Berenzweig, Beth Logan, Daniel P.W. Ellis, Brian Whitman. (2003) “A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures”. http://alumni.media.mit.edu/~bwhitman/ismir03-sim.pdf

[15] Wim D'haes, Xavier Rodet. (2003) "Discrete Cepstrum Coefficients as Perceptual Features".
http://mediatheque.ircam.fr/articles/textes/Dhaes03b/

[16] Bee-Suan Ong, Perfecto Herrera. (2004) "Computing Structural Descriptions of Music through the Identification of Representative Excerpts from Audio Files".
http://ww.iua.upf.es/mtg/ismir2004/graduateschool/people/Suan/OngBeeSuan.pdf

[17] Xi Shao, Changsheng Xu, Ye Wang , Mohan S Kankanhalli. (2004) "Automatic Music Summarization in Compressed Domain".
http://www.comp.nus.edu.sg/~shaoxi/papers/ICASSP04_Shao.pdf

[18] Jonathan Foote. (1999) "Visualizing Music and Audio using Self-Similarity".
http://www.fxpal.com/publications/FXPAL-PR-99-093.pdf

[19] Matthew Cooper, Jonathan Foote. (2003) "Summarizing Popular Music Via Structural Similarity Analysis".
http://www.fxpal.com/publications/FXPAL-PR-03-04.pdf

[20] Jeewoong Ryu, Yumi Sohn, Munchurl Kim. (2002) "MPEG-7 Metadata Authoring Tool"
http://vega.icu.ac.kr/~mccb-lab/publications/Paper/Ryu_ACM_Multimedia2002.pdf

[21] Jean-Julien Aucouturier, Mark Sandler. (2001) "Using Long-Term Structure to Retrieve Music: Representation and Matching".
http://www.csl.sony.fr/downloads/papers/uploads/aucouturier-01b.pdf

[22] Jason D. M. Rennie. (2004) "Derivation of the F-Measure".
http://people.csail.mit.edu/jrennie/writing/fmeasure.pdf

[23] Adriana Dapena. (2005) "Técnicas de Procesado de la Imagen"
http://www.des.udc.es/~adriana/TercerCiclo/CursoImagen/curso/web/Filtrado_Espacial.htm

[24] Elias Pampalk. Ma Toolbox for Matlab "A collection of similarity measures for audio ".
http://www.ofai.at/~elias.pampalk/ma

[25] Un4seen Developments. BASS audio library
http://www.un4seen.com/

[26] Robert Davies. NewMat C++ Matrix Library
http://www.robertnz.net/nm_intro.htm

[27] Sergi Jordà Puig. (2001) "Audio Digital y Midi. Capítulo 1: Principios de Acústica".
http://www.ccapitalia.net/reso/articulos/audiodigital/pdf/01-PrincipiosAcustica.pdf

[28] Domeltsch Online. Music Theory Online: pitch, temperamento & timbre.
http://www.dolmetsch.com/musictheory27.html