ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INDUSTRIALES Y DE TELECOMUNICACIÓN

Titulación :

INGENIERO DE TELECOMUNICACIÓN

Título del proyecto:

SEPARACIÓN DE FUENTES SONORAS ARMÓNICAS: ESTUDIO COMPARATIVO

Diego Cordón Urbiola

Miroslav Zivanovic Jeremic

Pamplona, 24 de abril de 2012

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Índice general

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Capítulo 1
Introducción

1.1. Objetivos

El objeto de este Proyecto Fin de Carrera es realizar un **estudio comparativo** entre métodos que resuelven la **separación de armónicos solapados** en frecuencia. Se trata del principal problema cuando se necesita extraer las diferentes pistas que conforman un sonido polifónico, pues trabajar sobre una fuente directamente desde la mezcla es virtualmente imposible. La separación de fuentes tiene distintas aplicaciones, por ejemplo, codificación, análisis y manipulación de audio.

- **Codificación**

  Los codificadores de audio están basados en eliminar información de la señal de manera que no se puede percibir esa pérdida por parte de un oyente (los llamados codificadores **perceptivos**). Estos métodos trabajan sobre la señal completa y obtienen unos resultados aceptables; no obstante, extraer las fuentes independientes para trabajar sobre ellas permite un mayor *ratio* de codificación. Por ejemplo, MPEG-4 es un estándar de codificación para audio y vídeo que es empleado en múltiples formatos, como MP3, DVD y TV Digital. Dispone de dos métodos de compresión del audio, AAC (*Advanced Audio Coding*) y un método paramétrico que modela las señales como suma de sinusoides independientes y sus armónicos junto al ruido.

- **Análisis**

  Analizar señales polifónicas es complicado porque la existencia de dos o más sonidos que suenan a la vez afecta a las mediciones que se llevan a cabo. Una manera de solucionar ese problema es realizar un paso previo que consista en aislar las distintas fuentes y luego analizarlas por separado.

- **Manipulación**

  La separación de las fuentes permite una manipulación más eficiente; pueden ser suprimidas, desplazadas temporalmente o editadas de manera independiente. Puede ser empleada también para eliminar ruido, útil para mejorar la calidad de grabaciones antiguas.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Para este proyecto se han analizado 3 métodos que persiguen el mismo objetivo aplicando diferentes conceptos. El primero soluciona el problema **filtrando el espectro** de la señal que se pretende separar; el segundo emplea la **Modulación en Amplitud Conjunta** para estimar los parámetros necesarios para la extracción; el tercero aplica funciones de **máxima verosimilitud** para estimar las partes (fuentes independientes) a partir del todo (mezcla polifónica). Debido a los orígenes conceptualmente distintos, se han seguido tres líneas de investigación diferentes en las que se ha estudiado el funcionamiento interno de cada método para finalmente implementar tres algoritmos capaces de llevar a cabo la separación de fuentes.

El estudio consiste en evaluar la capacidad de cada uno de los métodos de extraer las pistas de una mezcla polifónica; cada mezcla está formada por:

- Una pista **fija** que corresponde a una nota musical concreta.
- Una pista **variable** que corresponde a otra nota musical elegida de manera que el **porcentaje de armónicos solapados** entre ambas sea conocido. Por lo tanto, hay tantas mezclas como pistas variables.

Las señales de audio creadas se emplean para medir la capacidad de separación de cada método conforme aumenta el porcentaje de armónicos solapados respecto del total de armónicos existentes. El resultado es un gráfico que informa de la calidad de la separación del mejor al peor caso; para ello, se comprueba la energía de cada pista original con su estimación para cada uno de los tres métodos. Para entender los resultados y el diseño de cada uno de ellos, el proyecto comienza con un capítulo en el que se explican los conceptos musicales y físicos necesarios para entender el origen de los armónicos solapados. También se explica la necesidad de aplicar ventanas en el procesado digital de señales y cómo eso afecta al solapamiento; así se orienta acerca de la problemática que supone la separación de los armónicos solapados y se ayuda a entender los procedimientos que siguen los métodos aquí descritos para tratar de resolver el problema.

1.2. Motivación

Como se ha indicado en el apartado anterior, la separación de fuentes se puede aplicar en multitud de escenarios con distintos objetivos finales. Sin embargo, un usuario que necesite hacer una separación de fuentes no tiene por qué tener conocimientos sobre la materia. Ésto, unido al hecho de que no existe ningún estudio comparativo de métodos de separación recientes, constituye un *handicap* grande que llevará al usuario a obtener malos resultados debido a su desconocimiento.

Uno de los objetivos de este proyecto es obtener unos resultados **objetivos** que permitan **comparar** la calidad de los métodos estudiados. Se trata de tres de los métodos más recientes y por ello **no existe** un estudio que evalúe y dictamine su rendimiento. Estos resultados tienen un doble propósito: por un lado, ponerlos a prueba en las mismas condiciones, para dar una visión realista de su calidad; por otro, que su interpretación **pueda ayudar** a un usuario inexperto a elegir el mejor método para su escenario de trabajo.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Capítulo 2

Fundamentos de la separación de fuentes sonoras

En este capítulo se introducen los conceptos básicos de la música actual occidental, para más adelante explicar el modelado matemático que se le aplica. De esta manera se justifica la aparición de los armónicos solapados.

2.1. Fundamentos musicales

Los sonidos que integran cualquier composición musical se representan en una partitura, de manera similar al uso de las letras para representar el habla humana. La correcta lectura de una partitura requiere interpretar adecuadamente varios elementos.

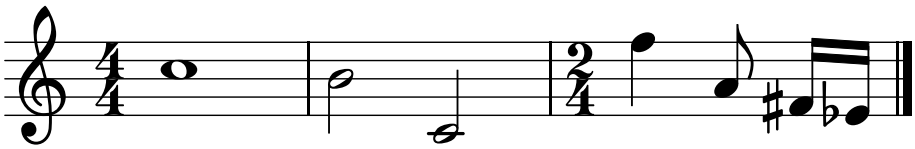


Figura 2.1: *Notación musical habitual.*

Evento o nota musical

Consiste en un sonido con un determinado *pitch* o frecuencia fundamental, que se mantiene activo en un intervalo temporal. Para determinar el valor físico (los Hz) que corresponde a cada evento, se emplea una **escala musical** formada por 7 notas sin alteración -signos que modifican la entonación, esto es, el *pitch*- y 5 con alteración. Existen dos alteraciones: el sostenido (♯) y el bemol (♭). La primera sube un semitono[1] a la nota y la segunda lo baja.

Existen dos formas de nombrar las notas. Según el sistema de notación latino, las 7 notas se representan mediante *Do - Re - Mi - Fa - Sol - La - Si*, y las alteraciones son *Do*♯ o *Re*♭, *Re*♯ o *Mi*♭, *Fa*♯ o *Sol*♭, *Sol*♯ o *La*♭ y *La*♯ o *Si*♭. En el sistema inglés, las siete notas se representan con siete letras, de la A a la G, y las cinco alteraciones se expresan con los sostenidos y bemoles, al igual que en el sistema latino. La tabla 2.1 muestra las equivalencias.

[1] mínima distancia frecuencial entre dos eventos adyacentes. Por ejemplo, A y A♯.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Tabla 2.1: *Sistemas de notación musical*

Sistema latino	*Do*	*Re*	*Mi*	*Fa*	*Sol*	*La*	*Si*
Sistema inglés	*C*	*D*	*E*	*F*	*G*	*A*	*B*

Escala musical

Se indica mediante un número que sigue a la letra de cada nota; por ejemplo, *A*4 quiere decir que se está considerando la nota *A* y la escala musical 4. Esta simbología está referida a un piano de 88 teclas (ver figura 2.2), de manera que *A*4 es equivalente a un *pitch* de 440 Hz. En la música occidental, se emplea la escala **diatónica**, la cual consiste en un conjunto de notas con una separación de 2, 2, 1, 2, 2, 2, 1 semitonos entre una nota y la siguiente.

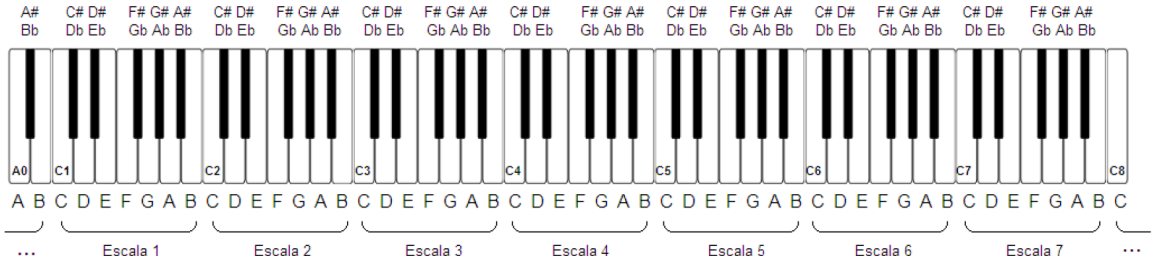


Figura 2.2: *Teclado estándar de 88 teclas.*

Según el convenio utilizado en este tipo de música, los eventos musicales son ordenados utilizando una escala de frecuencia logarítmica, puesto que el oído humano presenta un comportamiento de este tipo. Por lo tanto, la frecuencia fundamental de cada evento es:

$$f_0 = 440 \times 2^{\frac{n}{12}}, \quad n = -48, \ldots, 0, \ldots, 39 \tag{2.1}$$

Los límites de n hacen referencia al teclado estándar de 88 teclas. En este contexto, los valores en frecuencia asociados de *C*4 a *A*4 (sin las alteraciones), que son las empleadas en la fase experimental de este proyecto, ya pueden determinarse y se muestran en la tabla 2.2.

Tabla 2.2: *Frecuencias de las notas C4 a A4.*

Nota musical	*C*4	*D*4	*E*4	*F*4	*G*4	*A*4
Valor en Hz	261.6	293.7	329.6	349.2	392.0	440.0

Pentagrama

El pentagrama consiste en cinco líneas paralelas con cuatro espacios, utilizado para escribir los signos musicales. Según el convenio de la música occidental, para interpretar la composición musical se sigue un flujo temporal de izquierda a derecha, y el *pitch* de cada evento se representa según la posición vertical que ocupa.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Clave

Es un símbolo usado en notación musical, necesario para asociar los eventos musicales con la posición en el pentagrama de los símbolos que los representan. Existen tres claves distintas (figura 2.3), que llevan el nombre de la nota que designan en el pentagrama.

(a) (b) (c)

Figura 2.3: *Tipos de claves musicales. (a) Clave de Sol; (b) Clave de Fa; (c) Clave de Do.*

Duración

La duración se representa mediante figuras musicales. Cada figura tiene una duración relativa respecto a las demás; normalmente se establece como duración de referencia la figura negra (ver tabla 2.3).

Tabla 2.3: *Duración de los eventos musicales.*

Tipo de figura	Redonda	Blanca	**Negra**	Corchea	Semicorchea
Duración relativa	4	2	**1**	1/2	1/4
Símbolo musical	𝅝	𝅗𝅥	♩	♪	𝅘𝅥𝅯

Compás

Es la entidad de la métrica musical, el cual divide la pieza musical en fragmentos de igual duración que facilita su lectura. Los fragmentos están delimitados por unas líneas verticales que cortan el pentagrama, denominadas líneas divisorias. Los compases pueden ser binarios, ternarios o cuaternarios, según el número de tiempos que contengan. El tipo de compás se indica al principio del pentagrama, tras la clave, mediante un cociente. El **numerador** indica el número de tiempos que tendrá el compás; el **denominador** indica la unidad de tiempo, esto es, la figura que ocupa un tiempo del compás. Por ejemplo, en un compás $\frac{3}{4}$ se indica que cada compás tendrá 3 pulsos y el 4 indica que la figura unidad será la negra, luego cada compás tendrá tres negras.

2.2. Fundamentos físicos

De cara al análisis de señales musicales de audio, la **frecuencia fundamental** del sonido es una propiedad de interés. Si una señal $x(t)$ se repite cada cierto intervalo de tiempo, este

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

intervalo se denomina periodo fundamental T_0, se mide en segundos y satisface $x(t) = x(t+T_0)$ para todo t. Se define la frecuencia fundamental como la inversa de T_0:

$$f_0[\text{Hz}] = \frac{1}{T_0}$$

Según el Teorema de Fourier, toda señal periódica $x(t)$ se puede descomponer mediante una suma de sinusoides $s_k(t)$ relacionadas armónicamente entre sí:

$$s_k(t) = \text{Re}\left(a_k \cdot e^{jk2\pi f_0 t}\right)$$

Cada sinusoide $s_k(t)$ se define como la parte real de una exponencial compleja, la cual podría ser representada mediante una función coseno con tres parámetros: amplitud ($|a_k|$), frecuencia (f_0) y fase (ϕ_k):

$$x(t) = \sum_{k=-\infty}^{\infty} s_k(t) = \sum_{k=-\infty}^{\infty} a_k \cdot e^{jk2\pi f_0 t} \tag{2.2}$$

El representar una señal $x(t)$ como exponenciales complejas es útil porque este tipo de funciones son autofunciones de los sistemas lineales e invariantes en el tiempo (LTI), luego **toda combinación lineal de sinusoides a la entrada produce a la salida la misma combinación** con un escalado en la amplitud compleja; es decir, a_k y ϕ_k son modificadas por el sistema, no siéndolo la frecuencia fundamental f_0.

Si se considera que la señal $x(t)$ es real, con valor medio nulo y $a_k = |a_k| \cdot e^{j\phi_k}$, ocurre que $x(t) = x^*(t)$, luego $a_k = a^*_{-k}$ y la expresión (2.2) queda:

$$\begin{aligned} x(t) &= \sum_{k=-\infty}^{\infty} a_k \cdot e^{jk2\pi f_0 t} = \sum_{k=-\infty}^{-1} a_k \cdot e^{jk2\pi f_0 t} + \sum_{k=1}^{\infty} a_k \cdot e^{jk2\pi f_0 t} = \cdots \\ &= 2\sum_{k=1}^{\infty} |a_k| \cos\left(2\pi k f_0 t + \phi_k\right) \end{aligned} \tag{2.3}$$

En esta expresión, cada componente $kf_0 = f_k$, $k \in \mathbb{N}$ se denomina **armónico** y su frecuencia es un **múltiplo entero** de la frecuencia fundamental. Sin embargo, debido a que las señales musicales existentes en el mundo son no estacionarias a largo plazo, es decir, su media μ_x y autocorrelación R_x son variantes con el tiempo, se consideran pequeños intervalos de tiempo denominados **tramas** o **frames** (decenas de milisegundos de duración), dentro de los cuales la señal es -o al menos puede considerarse- estacionaria.

En resumen, como toda señal $x(t)$ puede ser descompuesta en múltiples sinusoides relacionadas entre sí por la frecuencia fundamental, el sonido producido por un instrumento musical puede desglosarse de esta manera. El resultado de este modelado matemático combinado con la teoría musical establecida actualmente, explicada en la sección 2.1, crea el problema de los armónicos solapados en frecuencia. Si se parte de la tabla 2.2 y se calculan las frecuencias de los armónicos para esas frecuencias (tabla 2.4), se observa que varios valores de f_k están próximos.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Tabla 2.4: *Frecuencias fundamentales y sus armónicos.*

Nota musical	*C*4	*D*4	*E*4	*F*4	*G*4	*A*4
$k = 1$ (f_0)	261.6	293.7	329.6	349.2	392.0	440.0
$k = 2$	523.2	587.4	659.2	698.40	784.0	880.0
$k = 3$	784.8	881.1	988.8	1047.6	1176.0	1320.0
$k = 4$	1046.4	1174.8	1318.4	1396.8	1568.0	1760.0
$k = 5$	1308.0	1468.5	1648.0	1746.0	1960.0	2200.0
$k = 6$	1569.6	1762.2	1977.6	2095.2	2352.0	2640.0
$k = 7$	1831.2	2055.9	2307.2	2444.4	2744.0	3080.0
$k = 8$	2092.8	2349.6	2636.8	2793.6	3136.0	3520.0
$k = 9$	2354.4	2643.3	2966.4	3142.8	3528.0	3960.0
$k = 10$	2616.0	2937.0	3296.0	3492.0	3920.0	4400.0

Por ejemplo, el segundo armónico de la nota *D*4 suena a 881.1 Hz y el primero de *A*4 suena a 880 Hz. Si se representa el espectro de una señal de audio formada por dos fuentes distintas que reproducen esas notas al unísono, el resultado conlleva pérdida de información si las frecuencias fundamentales están próximas. Lo siguiente es, pues, determinar cuándo se produce o no esa pérdida, definiendo un **umbral de decisión** cuyo valor depende de **cómo se haya enventanado** la señal.

El efecto del enventanado

En la sección anterior se ha explicado la necesidad de dividir las señales de audio en intervalos o tramas de unos pocos milisegundos de duración, para poder asumir la periodicidad de $x(t)$; sin embargo, con una duración de 10 ms no se puede aproximar la Transformada de Fourier de una señal sinusoidal a una función $\delta(f)$. Realmente, una trama de $x(t)$, es la propia señal multiplicada por un pulso centrado en un instante t y de anchura temporal τ, de forma que ese pulso viene desde $t = -\infty$ y la "recorre" hasta $t = \infty$, construyendo así las tramas $x_m(t)$:

$$x_m(t) = x(t) \cdot \Pi\left(\frac{t - i\tau}{2\tau}\right), \quad i = \ldots, -4, -2, 0, 2, 4, \ldots \tag{2.4}$$

Es necesario modificar cada trama $x_m(t)$ de manera que su espectro sea lo más parecido a una delta de Dirac. El procedimiento seguido es:

1. Convertir $x_m(t)$ a $x_m[n]$ mediante un proceso de muestreo. El **Teorema de Nyquist** obliga a que la velocidad a la que se te toman las muestras sea al menos doble de la componente frecuencial máxima de $x_m(t)$. Como el oído humano no percibe sonidos de más de 20 kHz, se toma $f_s = 44.1$ kHz (muestras/s). Antes de muestrear se "pasa" $x_m(t)$ por un filtro paso bajo con ancho de banda $BW = f_s/2$ para evitar el **aliasing**.

2. Multiplicar $x_m[n]$ por una **ventana** de igual longitud $w[n]$. El resultado es $x_m[n] \cdot w[n] = v_m[n]$.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

3. Calcular la DFT para finalmente obtener $V_m(k) = X_m(k) * W(k)$.

El objetivo de multiplicar por una ventana es doble:

- Que $W(k)$ se parezca lo más posible a un impulso para que $V_m(k)$ reproduzca fielmente la respuesta en frecuencia de $x_m[n]$.
- Que $w[n]$ suavice las transiciones bruscas que se producen al multiplicar $x(t)$ por el pulso Π.

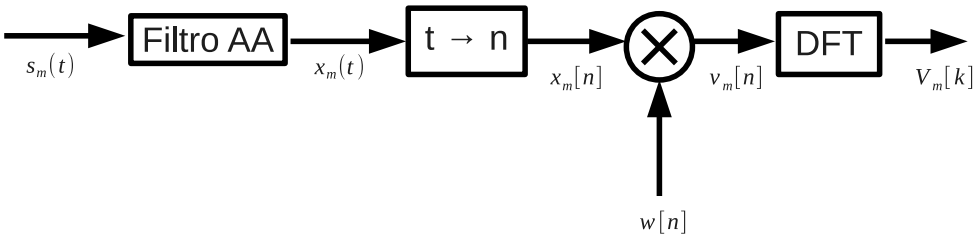


Figura 2.4: *Pasos del proceso del análisis de Fourier en tiempo discreto de una señal en tiempo continuo.*

Ambos requerimientos entran en conflicto, por lo que se ha de buscar una solución de compromiso que deriva en la existencia de distintas ventanas (figura 2.5).

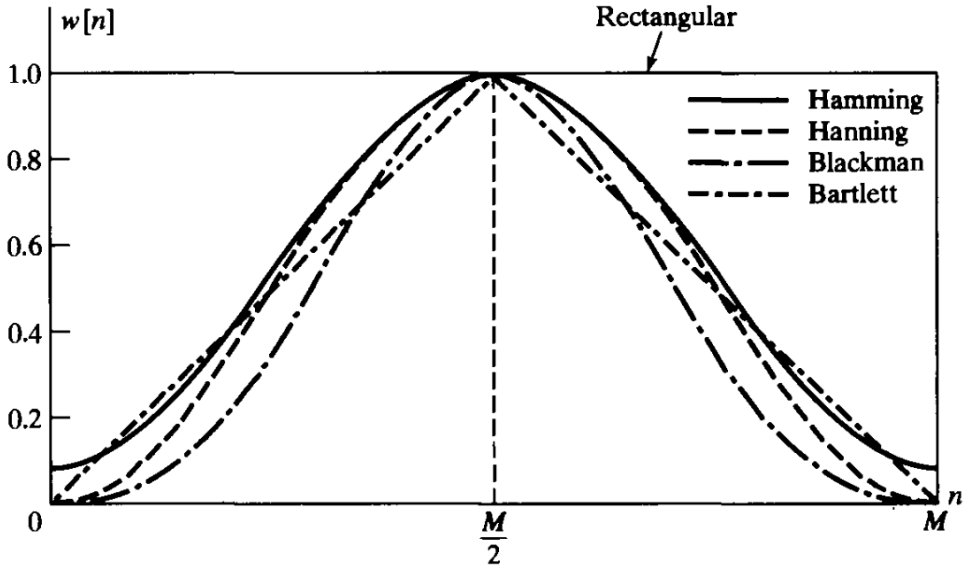


Figura 2.5: *Ventanas comúnmente utilizadas.*

En resumen, al enventanar se busca acercar las propiedades de $x_m[n]$ a las de $x(t)$, es decir, que el espectro obtenido sea como el de una función periódica e infinita, cuando en realidad se está trabajando con una señal finita.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Efecto del enventanado en una señal armónica Sea una señal en tiempo continuo formada por la suma de dos componentes sinusoidales:

$$s_m(t) = A_0 \cos(2\pi f_{01} t + \phi_0) + A_1 \cos(2\pi f_{02} t + \phi_1), \quad -\infty < t < \infty. \tag{2.5}$$

Suponiendo un proceso de muestreo sin solapamiento ni errores en la cuantificación, la señal en tiempo discreto resultante es:

$$x_m[n] = A_0 \cos(2\pi f_0 n + \phi_0) + A_1 \cos(2\pi f_0 n + \phi_1) \tag{2.6}$$

La trama enventanada $v_m[n]$ se puede desarrollar en función de exponenciales complejas y utilizar la propiedad de desplazamiento en frecuencia. Entonces, $v_m[n]$ queda:

$$\begin{aligned} v_m[n] = &\frac{A_0}{2} w[n] e^{j\phi_0} e^{j2\pi f_{01} n} + \frac{A_0}{2} w[n] e^{-j\phi_0} e^{-j2\pi f_{01} n} \\ &+ \frac{A_1}{2} w[n] e^{j\phi_1} e^{jw\pi f_{01} n} + \frac{A_1}{2} w[n] e^{-j\phi_1} e^{-j2\pi f_{02} n} \end{aligned} \tag{2.7}$$

La Transformada de Fourier de la secuencia enventanada es:

$$\begin{aligned} V_m(f) = &\frac{A_0}{2} e^{j\phi_0} W\left(e^{j2\pi(f-f_{01})}\right) + \frac{A_0}{2} e^{-j\phi_0} W\left(e^{j2\pi(f+f_{01})}\right) \\ &+ \frac{A_1}{2} e^{j\phi_1} W\left(e^{j2\pi(f-f_{02})}\right) + \frac{A_1}{2} e^{-j\phi_1} W\left(e^{j2\pi(f+f_{02})}\right) \end{aligned} \tag{2.8}$$

De acuerdo con la ecuación (2.8), la Transformada de Fourier está formada por el espectro de la ventana replicado en las frecuencias $\pm f_{01}$ y $\pm f_{02}$ y escalado por las amplitudes de las exponenciales complejas que forman la señal. Si esto se extiende a una señal armónica real, por cada fuente independiente habrá:

- Un pico que represente la frecuencia fundamental.
- Tantos picos como armónicos tenga la señal, posicionados en los múltiplos de f_0 y escalados en amplitud.
- Lóbulos laterales alrededor de los picos que afectarán al resto del espectro.

Umbral de decisión empleado El umbral empleado en los métodos que se describen en este proyecto es proporcional a la **resolución frecuencial** de la ventana, f_b:

$$f_b = \frac{f_s}{N}[\text{Hz}] \tag{2.9}$$

siendo f_s la frecuencia de muestreo y N el número de puntos de la ventana. En este proyecto se ha utilizado como umbral $1.5 f_b$ pues **es el valor empleado** en los artículos que proponen los métodos estudiados. Es decir, si la diferencia entre las frecuencias de dos armónicos es menor de $1.5 f_b$, ambos armónicos se consideran solapados y es necesario recuperar sus amplitudes y frecuencias mediante cualquiera de las técnicas descritas en el capítulo 3.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

♣ Ejemplo de enventanado

La figura 2.6 muestra $|W(f)|$ para una ventana rectangular de longitud 64, y en las figuras 2.7 a 2.10 se muestra $V_m(f)$ para distintos valores de $f_{02} - f_{01}$ con $f_s = 10$ kHz, $\phi_0 = \phi_1 = 0$, $A_0 = 1$ y $A_1 = 0.75$.

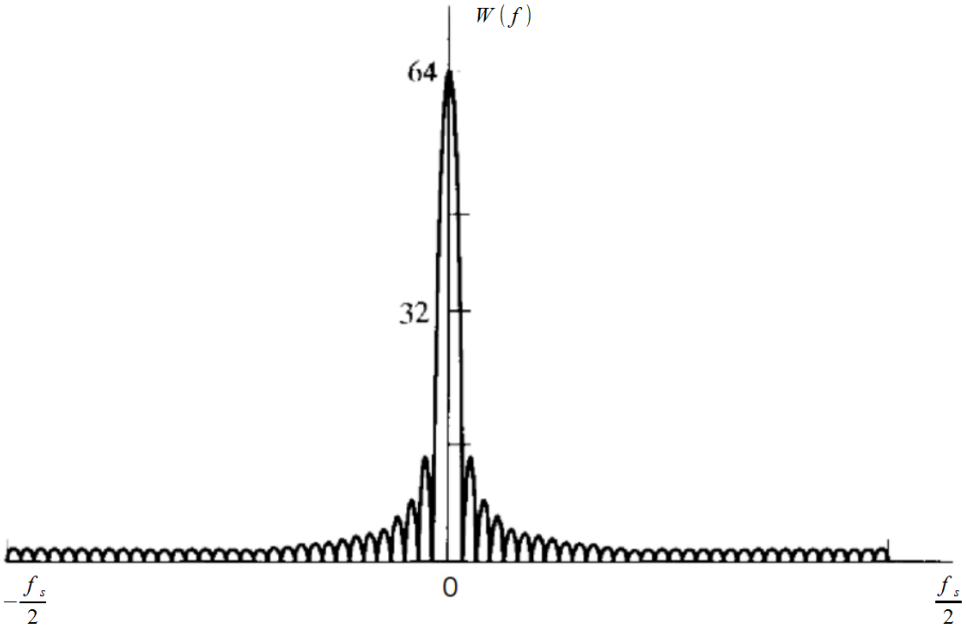


Figura 2.6: *Transformada De Fourier de una ventana rectangular.*

En la figura 2.7, $f_{02} = 14.7$ kHz y $f_{01} = 7.35$ kHz. La expresión (2.8) sugiere que, al no haber solapamiento entre las réplicas de $W(f)$ en f_{01} y f_{02}, habrá un pico de altura $32A_0$ en f_{01} y de $32A_1$ en f_{02} (pues $W(f)$ tiene un pico de altura 64), que se aprecian de forma evidente.

A partir de la figura 2.8, las frecuencias se van acercando progresivamente:

- En la figura 2.8, hay mucho solapamiento entre las réplicas de la ventana en f_{01} y f_{02}; aunque están diferenciados los dos picos, la amplitud del espectro en $f = f_{01}$ está afectada por la amplitud de la señal sinusoidal en la frecuencia f_{02} y viceversa, debido a la dispersión espectral que introduce la ventana. Esta interacción se denomina **leakage**, por lo que es necesario utilizar una ventana que afecte lo menos posible al espectro $V_m(f)$ resultante (la ventana empleada en este caso no es la más adecuada).
- En la figura 2.9, al no sumarse en fase los lóbulos laterales generados por la ventana, se reducen las amplitudes de los picos. Se ha producido más leakage pero siguen *resolviéndose* los picos de los tonos.
- En la figura 2.10, el solapamiento entre las ventanas espectrales en f_{01} y f_{02} es tan grande que los dos picos se han fundido en uno sólo; es decir, empleando una ventana rectangular no es posible *resolver* en el espectro dos frecuencias tan juntas.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

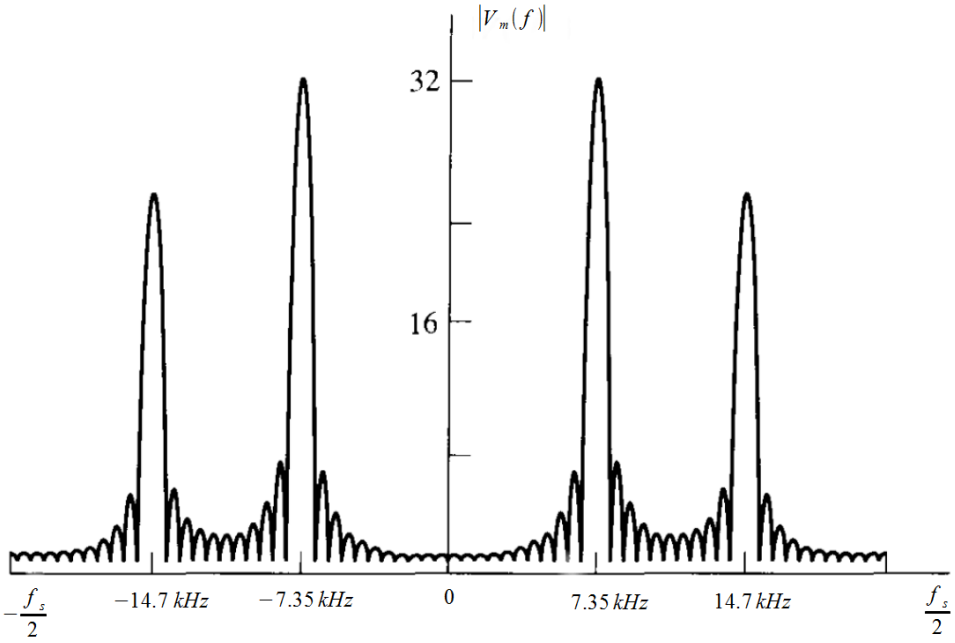


Figura 2.7: *Transformada de Fourier de dos tonos enventanados para* $f_{02} - f_{01} = 7.35$ kHz.

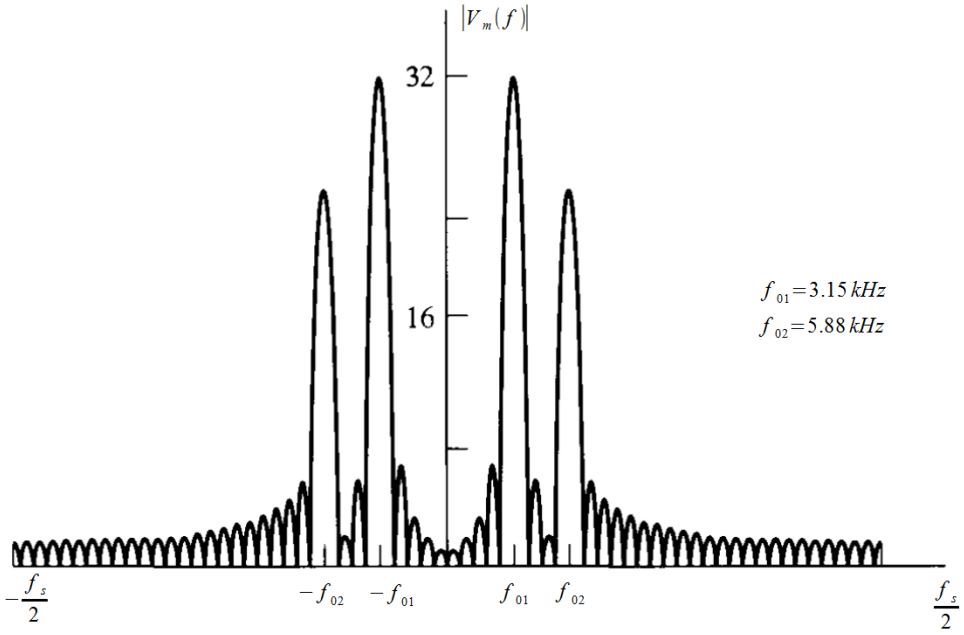


Figura 2.8: *Transformada de Fourier de dos tonos enventanados para* $f_{02} - f_{01} = 2.73$ kHz.

La última figura del ejemplo muestra el problema que trata este proyecto: al emplear una ventana se ha producido pérdida de información pues ya no es posible extraer las amplitudes y fases directamente de $V_m(f)$. Si bien empleando otro tipo de ventana es posible resolver frecuencias más próximas, ésto tiene un límite y **va a ser inevitable que aparezcan armónicos solapados en frecuencia**. Además, la validez de los métodos descritos en el siguiente capítulo está condicionada a que las señales $x(t)$ que son muestreadas para obtener las secuencias dis-

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

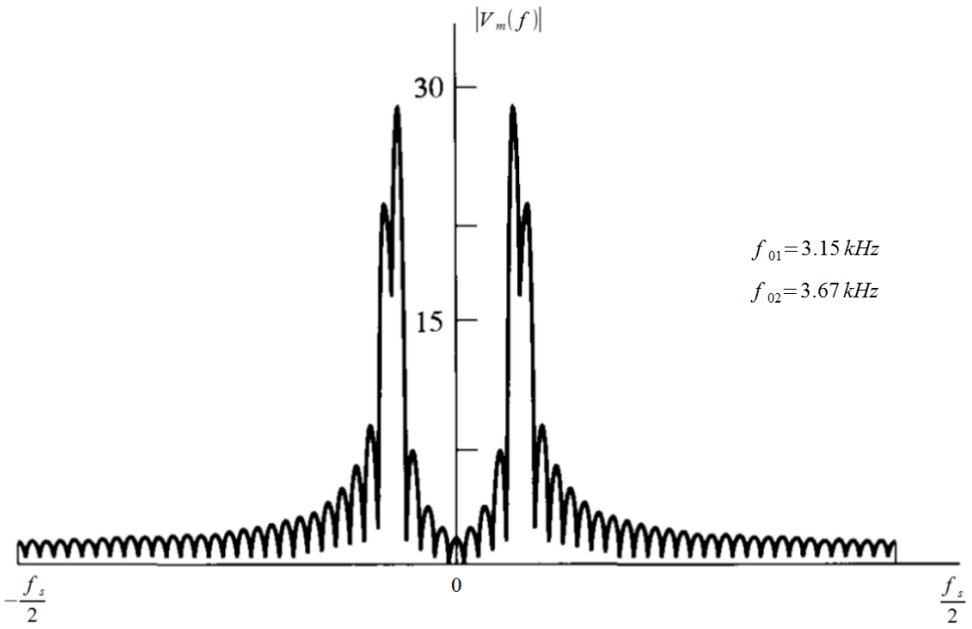


Figura 2.9: *Transformada de Fourier de dos tonos enventanados para* $f_{02} - f_{01} = 520$ Hz.

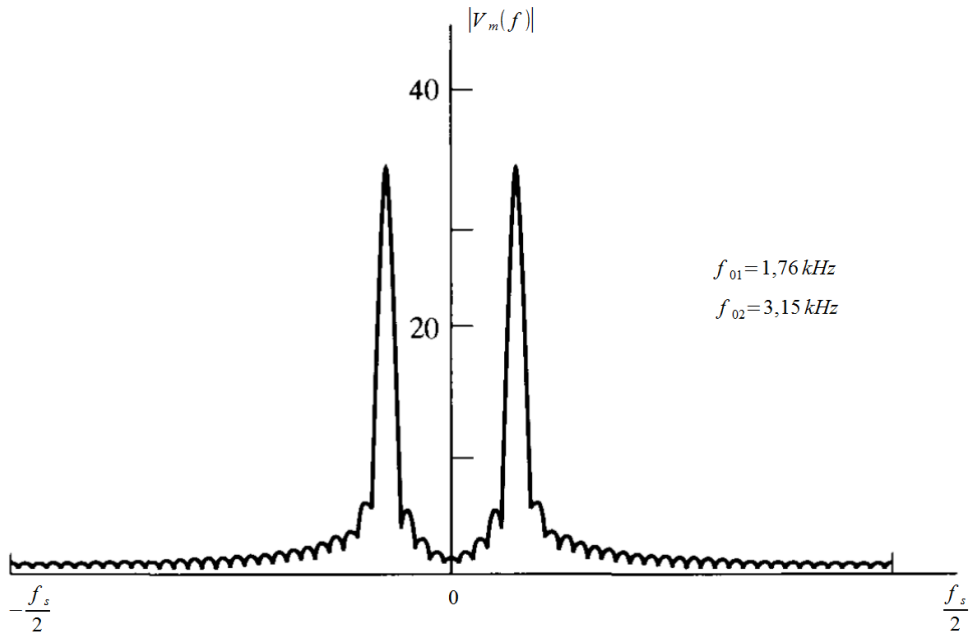


Figura 2.10: *Transformada de Fourier de dos tonos enventanados para* $f_{02} - f_{01} = 1.39$ kHz.

cretas $x[n]$ puedan ser modeladas como una suma de sinusoides, es decir, como en la expresión (2.3). A una señal $x(t)$ se le puede aplicar el Teorema de Fourier si:

- Es absolutamente integrable sobre cualquier periodo:

$$\int_T |x(t)|dt < \infty$$

- Tiene un número finito de máximos y mínimos en cualquier periodo.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

- Tiene un número finito de discontinuidades en cualquier periodo.

Se asume que las señales empleadas para realizar el estudio cumplen estas condiciones.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Capítulo 3

Métodos de separación de armónicos solapados

3.1. Filtrado espectral

Esta sección describe el primer método estudiado en este proyecto. Comienza con una introducción teórica en la que se explica que filtrar en el dominio discreto equivale a multiplicar secuencias de valores, y continúa con el desarrollo de los filtros y un esquema del funcionamiento interno del programa.

Al final se muestra una tabla que resume los parámetros más relevantes que condicionan el buen funcionamiento del método, con los valores óptimos empleados.

Introducción

Dada una secuencia $x[n]$ finita formada por N muestras ($n = 0, \ldots, N-1$), la Transformada Discreta de Fourier (DFT, *Discrete Fourier Transform*) es una **correspondencia** entre los N valores de $x[n]$ y N coeficientes de sus armónicos en el espectro. La DFT se conoce también como **ecuación de síntesis**:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N}, \quad k = 0, 1, \ldots, N-1 \tag{3.1}$$

La inversa de la DFT es la **ecuación de análisis**, y permite recuperar la señal temporal a partir de su espectro:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] e^{j2\pi kn/N}, \quad n = 0, 1, \ldots, N-1 \tag{3.2}$$

$X[k]$ es una secuencia finita, formada por N puntos que pueden interpretarse como N muestras del espectro continuo $X(\Omega)$ correspondiente a una secuencia finita, no periódica, formada por el mismo número de muestras. Estas muestras son equidistantes y corresponden a los sucesivos valores de $\Omega = 2\pi k/N$ con $k = 0, 1, \ldots, N-1$. La discretización de $X(\Omega)$ se debe a que $x[n]$ es periódica:

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

$$x[n - mN] = x[n] \quad \forall m \in \mathbb{Z}$$

Esta propiedad se deduce de las ecuaciones de análisis y síntesis. Para demostrar la periodicidad de $x[n]$:

$$\begin{aligned} x[n - mN] &= \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{j2\pi k(n-mN)/N} = \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{j2\pi kn/N}e^{-j2\pi kmN/N} \\ &= \frac{1}{N}\sum_{k=0}^{N-1}\Big(X[k]e^{j2\pi kn/N}\cdot\big\{\cos(2\pi km) - j\sin(2\pi km)\big\}\Big) = \frac{1}{N}\sum_{k=0}^{N-1} X[k]e^{j2\pi kn/N} \\ &= x[n] \end{aligned} \tag{3.3}$$

Se puede comprobar que$X[k]$ también es periódica de igual manera.

Fast Fourier Transform

Las ecuaciones de análisis y síntesis son herramientas computacionales con un papel fundamental en el procesamiento digital de señales, donde se incluyen las señales de audio, que son con las que trabaja este proyecto. Esta importancia se debe en gran parte a la existencia de algoritmos eficientes para el cálculo de la DFT (análisis) y DFT^{-1} (síntesis), conocidos como **FFT** (Transformada Rápida de Fourier, *Fast Fourier Transform*).

Filtros en tiempo discreto

Los filtros en tiempo discreto son secuencias de puntos y se caracterizan con una **respuesta impulsional** $h[n]$, es decir, la salida que el sistema devuelve si la señal de entrada es un impulso: $x[n] = \delta[n] \Rightarrow y[n] = h[n]$. Filtrar en el dominio discreto equivale a la operación de **convolución**:

$$y[n] = x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] \tag{3.4}$$

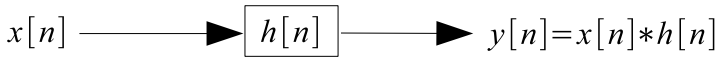


Figura 3.1: *Esquema de un proceso de filtrado en tiempo discreto.*

Una de las propiedades de la Transformada de Fourier es que una convolución en el dominio temporal se convierte en un producto en el dominio transformado:

$$y[n] = x[n] * h[n] \rightleftharpoons Y[k] = X[k]H[k] \tag{3.5}$$

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Por otro lado, se ha visto que la FFT devuelve una secuencia finita de puntos. Entonces, es posible filtrar $x[n]$ multiplicando su espectro por la **función de transferencia** del filtro y tras aplicarle una DFT inversa, se obtiene una $\hat{x}[n]$ modificada.

El método que se describe a continuación se basa en el filtrado de los picos espectrales correspondientes a los armónicos de las distintas fuentes que conforman una señal, manteniendo intactos los pertenecientes a la fuente i y eliminando el resto. La secuencia que se consigue así, al pasarla al dominio temporal, corresponderá a la pista i-ésima de la mezcla. Para ello, solo hay que multiplicar los puntos de $X[k]$ y $H[k]$ uno a uno y después antitransformar; es el diseño de los filtros lo que determina la calidad de la separación.

Desarrollo

Como se ha justificado en la sección 2.2, se comienza dividiendo la señal original en tramas de N muestras y se analiza cada trama individualmente. Empleando el umbral de decisión, establecido en $1.5f_b = 1.5f_s/N$, se clasifican los armónicos de cada fuente como solapados o no solapados. Estos armónicos son localizables, pues se conoce el *pitch* de cada señal y se ha demostrado que están en frecuencias que son múltiplos enteros de f_0. Es decir, un armónico de la fuente m centrado en $k_i f_{0m}$ ($k_i > 0$) está solapado si existe otro armónico de otra fuente n centrado en $k_j f_{0n}$ ($k_j > 0$), tal que $|k_i f_{0m} - k_j f_{0n}| < 1.5f_b$.

♣ Ejemplo

La imagen 3.2 es una combinación de dos eventos musicales: *C*4 y *D*4. Para un tamaño de ventana de 2048 muestras y frecuencia de muestreo de 44.1 kHz, el umbral se establece en 32 Hz, lo que lleva a la existencia de 2 armónicos solapados entre los 10 primeros, además de que en este caso también hay solapamiento en los picos de los *pitches*. Si se extiende a 20 armónicos, resulta un 25 % de solapamiento. La tabla 3.1 muestra las frecuencias en las que caen los picos.

Tabla 3.1: *Frecuencias de los primeros 10 armónicos de C4 y D4. Los solapados están resaltados en negrita.*

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
*C*4	**261.6**	523.2	784.8	1046.4	1308.0	1569.6	1831.2	2092.8	**2354.4**	**2616.0**
*D*4	**293.7**	587.4	881.1	1174.8	1468.5	1762.2	2055.9	**2349.6**	**2643.3**	2937.0

Una vez determinada la condición para cada armónico de cada fuente, comienza la fase de filtrado. Para ello, se construyen dos filtros según exista o no solapamiento:

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

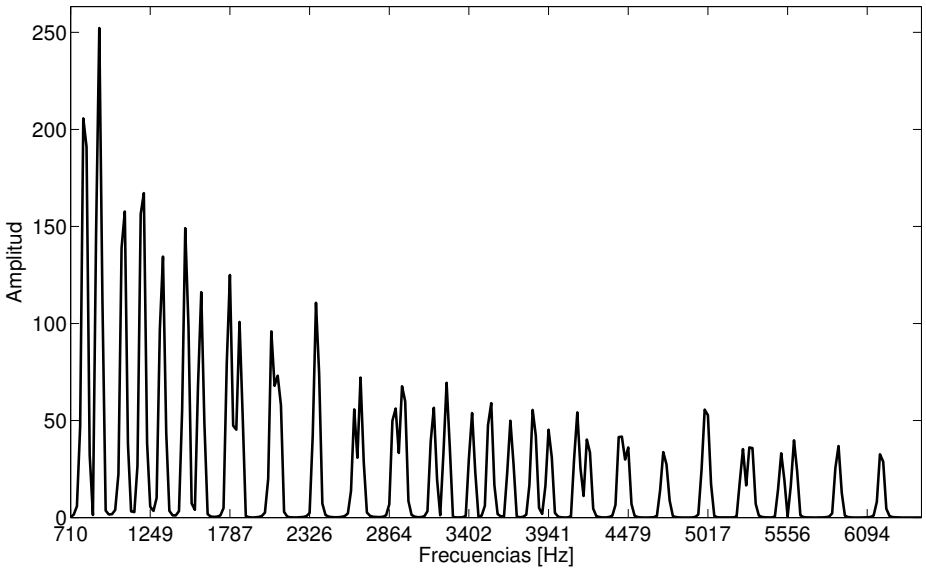


Figura 3.2: *Espectro de una trama enventanada en la que se reproducen simultáneamente las notas C4 y D4.*

Filtrado de armónicos no solapados

En este caso, se calcula la muestra de $X[k]$ sobre la que se centra el pico υ en el espectro y se denota k_υ^c. La frecuencia asociada a k_υ^c es:

$$f_\upsilon = k_\upsilon^c \frac{f_s}{N} \Rightarrow k_\upsilon^c = \frac{N f_\upsilon}{f_s} \tag{3.6}$$

Como $X[k]$ son muestras de un espectro continuo $X(\Omega)$, es posible que k_υ^c no sea un valor entero -lo es sólo si f_υ es múltiplo de $\frac{f_s}{N}$-.

El filtro ha de extraer la amplitud y fase del pico sin modificarlos, para luego emplear esa información al reconstruir la señal $\hat{x}[n]$ a la que corresponde. Por lo tanto, el módulo de su función de transferencia para los armónicos no solapados es la unidad. Para completar su diseño hay que decidir el rango de muestras que ha de "capturar", pues en el resto su módulo sera nulo. Si se considera k_υ^c como el valor máximo del pico, éste está ubicado entre dos mínimos locales:

- Un mínimo local "por la izquierda": $k_\upsilon^l < k_\upsilon^c$.
- Un mínimo local "por la derecha": $k_\upsilon^r > k_\upsilon^c$.

El rango de muestras en el que la función de transferencia no se anula es $k_\upsilon^l \leq k \leq k_\upsilon^r$. Estos valores, a diferencia de k_υ^c, siempre son valores enteros, pues se obtienen directamente de la secuencia $X[k]$; el valor central k_υ^c se emplea únicamente para indicar el punto donde comenzar la búsqueda.

♣ **Ejemplo**

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

El espectro de la figura 3.3 muestra 6 armónicos y el pico de la frecuencia fundamental. El máximo local del último armónico corresponde a $k_v^c = 95.91$, y los mínimos locales más próximos son $k_v^l = 90$ y $k_v^r = 97$, luego el diseño del filtro es:

$$H^p(k) = \begin{cases} 1 & \text{si } 90 \leq k \leq 97 \\ 0 & \text{resto} \end{cases}$$

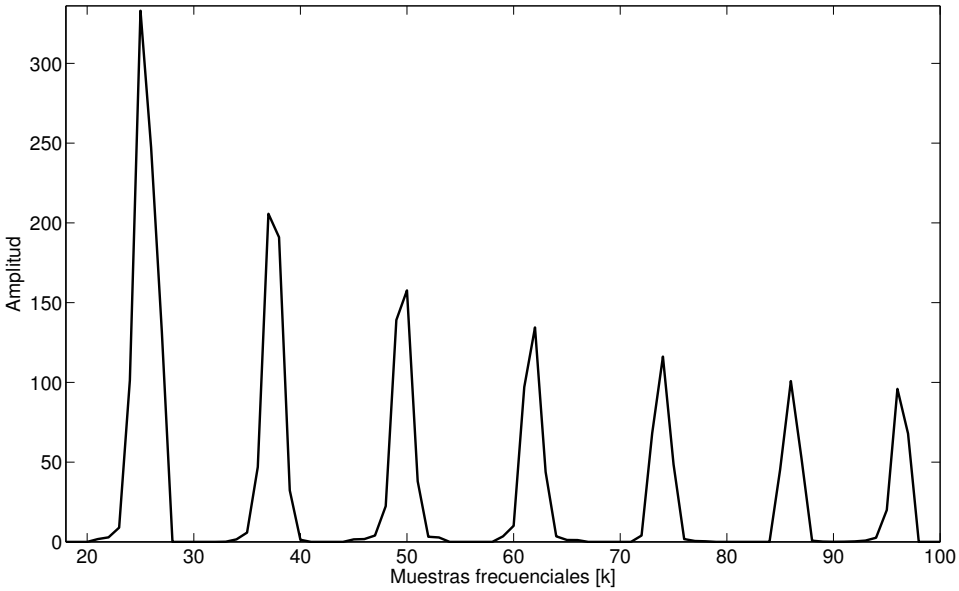


Figura 3.3: *Espectro de una trama enventanada con varios armónicos.*

Filtrado de armónicos solapados

Cuando se produce solapamiento, existen M armónicos de M fuentes mezclados en el intervalo $k^l \leq k \leq k^r$. De manera similar al apartado anterior, k^l y k^r representan los límites inferior y superior del conjunto de muestras en el que se produce el solapamiento. Hay que repartir la energía de esas muestras para obtener M picos distintos construyendo M filtros; cada uno de esos filtros mantiene la energía de un armónico y elimina la del resto. Para su diseño se tiene en cuenta tres parámetros:

- **Amplitud del armónico**

  Para "repartir" la energía disponible en la mezcla de manera proporcional a la que contenía cada uno de los M armónicos por separado, es necesario conocer la amplitud A_m^p (la amplitud del m-ésimo armónico de la p-ésima fuente) de cada uno de ellos. Para ello, se realiza una **interpolación lineal** entre las amplitudes de los armónicos más cercanos que no solapan y se obtiene una estimación $\hat{A}_m^p$.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

- **Frecuencia del armónico**

  Como se construye un filtro por armónico, se requiere el valor de su frecuencia f_m^p para centrarlo. Las distintas frecuencias pueden tener hasta $1.5 f_b$ Hz de diferencia, por lo que no es aceptable centrar el filtro en la muestra k con mayor energía, sino que lo adecuado es calcularla empleando (3.6).

- **Forma de la ventana**

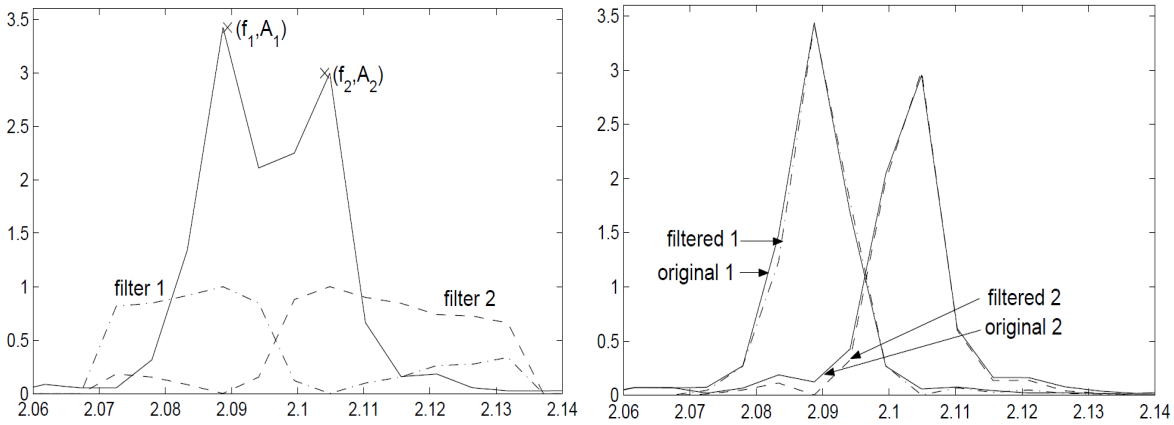
  El espectro de la ventana empleada $\mathcal{F}_h(f)$ aparece replicado y centrado en cada una de las f_m^p, y en la zona de solapamiento se produce la suma de todos ellos, como se muestra en la figura 2.10 de la página 16. Combinando ésto con la amplitud estimada $\hat{A}_m^p$ y la frecuencia f_m^p, se tiene una aproximación del espectro que cada armónico genera por separado.

El filtro $\hat{H}^p(k)$ que se propone es:

$$\hat{H}^p(k) = A_m^p \cdot |\mathcal{F}_h(|f_k - f_m^p|)| \tag{3.7}$$

Esta expresión se normaliza dividiendo $\hat{H}^p(k)$ entre el total de filtros:

$$H^p(k) = \frac{\hat{H}^p(k)}{\sum_{q \in Q} \hat{H}^q(k)} \quad k^l \leq k \leq k^r \tag{3.8}$$



(a) Solapamiento entre armónicos y forma de los filtros $H^p(k)$.

(b) Diferencia entre armónicos originales y estimados.

Figura 3.4: *Resultado de separar dos armónicos solapados mediante filtrado espectral. El eje horizontal representa la frecuencia en* kHz *y el vertical la amplitud en* dB.

Por último, hay que determinar k^l y k^r. Como en el caso del filtrado de armónicos no solapados, se localizan los mínimos locales alrededor de la muestra con el máximo y se emplean esos valores para hacer los M filtros; el intervalo así calculado representará prácticamente toda la energía de cada armónico.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Reconstrucción de las señales

Una vez diseñados los filtros para ambos casos, el proceso de separación de las fuentes sigue los siguientes pasos:

1. Tomar una trama $x[n]$, multiplicarla por la ventana $h[n]$ y calcular su FFT: $\mathcal{F}_{\mathbf{x}\cdot\mathbf{h}}$. El número de pistas de la mezcla y sus *pitches* se asumen conocidos, además de despreciar los armónicos de orden 21 en adelante, por aportar muy poco energéticamente.
2. Crear un vector $\mathbf{v}_i$ del mismo tamaño de la trama.
3. Multiplicar $\mathcal{F}_{\mathbf{x}\cdot\mathbf{h}}$ por la unidad en los rangos de muestras donde haya armónicos no solapados correspondientes a la i-ésima fuente. El resto del espectro se anula.
4. Guardar el resultado obtenido en $\mathbf{v}$.
5. Multiplicar$\mathcal{F}_{\mathbf{x}\cdot\mathbf{h}}$ por la expresión (3.8) en los rangos de muestras donde haya armónicos solapados correspondientes a la i-esima fuente. El resto del espectro se anula.
6. Sumar el resultado obtenido a $\mathbf{v}$.
7. Repetir los pasos 2 a 6 para cada fuente i de las M de la mezcla.
8. Aplicar la FFT inversa a los M vectores $\mathbf{v}_i$ para obtener las $\hat{x}_i[n]$.

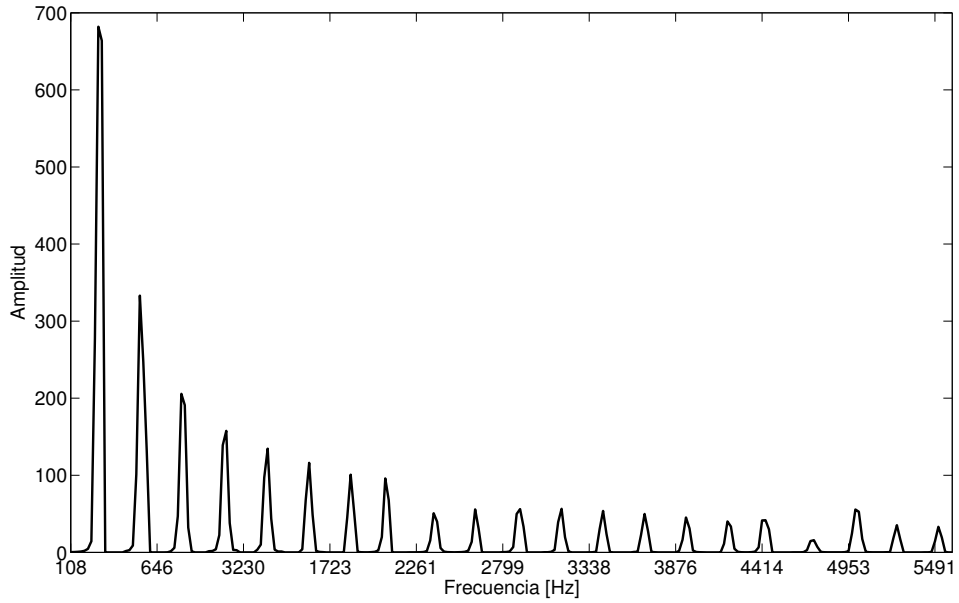
♣ **Ejemplo**

La figura 3.2 muestra el espectro de una señal que es la suma de dos sonidos, *C*4 y *D*4. El resultado de separar las fuentes mediante filtrado espectral se muestra en las figuras 3.5(a) y 3.5(b).

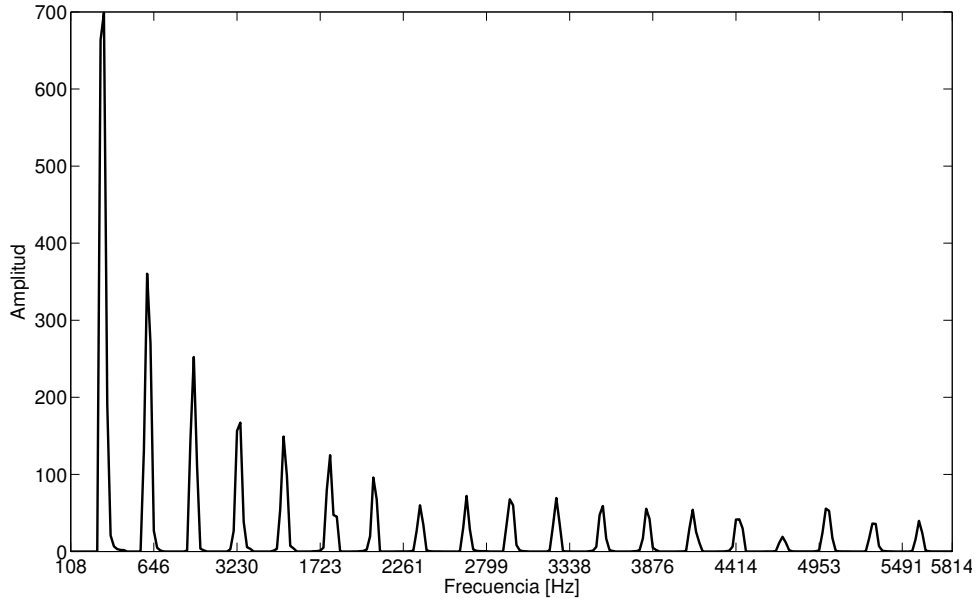
Valores empleados

Se muestra en la tabla 3.2 los valores por defecto introducidos en el código de este algoritmo.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira



(a) Evento musical *C*4 extraído.



(b) Evento musical *D*4 extraído.

Figura 3.5: *Fuentes armónicas extraídas mediante filtrado espectral.*

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Tabla 3.2: *Resumen de los parámetros empleados en el método del filtrado espectral*

Parámetro	**Valor**	**Dimensiones**	**Símbolo**
Tipo de ventana	Hamming		$h[n]$
Frecuencia de muestreo	44100	Hz	f_s
Tamaño de la ventana	2048	muestras	N
Resolución frecuencial de la ventana	f_s/N	Hz	f_b
Tamaño de las tramas temporales	2048	muestras	N
Hop size de la STFT	1792	muestras	
Umbral de decisión para determinar solapamiento entre armónicos	$1.5f_b$	Hz	

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

3.2. Modulación en amplitud conjunta

Esta sección describe el segundo método estudiado en este proyecto. La introducción teórica justifica la utilidad de la Modulación en Amplitud Conjunta para la separación de armónicos solapados, y en el desarrollo se explican las herramientas matriciales con las que se logra la separación.

Al final se muestra una tabla que resume los parámetros más relevantes que condicionan el buen funcionamiento del método, con los valores óptimos empleados.

Introducción

Cuando se genera un sonido con un instrumento musical, la estructura física del mismo provoca que surjan *armónicos* acompañando al tono principal, lo que da riqueza al sonido. El conjunto del tono y sus armónicos conforman el *timbre* del instrumento, y hacen que sea posible distinguir uno de otro, aunque estén reproduciendo sonidos de la misma frecuencia fundamental F_i. Es la distribución de amplitudes de los armónicos lo que permite diferenciarlos.

Las frecuencias f_i de esos armónicos son múltiplos enteros de la fundamental. Aunque en la práctica esto no es del todo exacto, se puede asumir que:

$$f_i^{h_i} \approx h_i F_i \tag{3.9}$$

siendo h_i el orden del armónico: $h_i = 2, 3, 4 \ldots$

Por otro lado, es inevitable que las señales sufran modificaciones en su frecuencia, amplitud y fase mientras suenan. Estas variaciones se conocen como *modulaciones*, y matemáticamente se expresan como una función $x(t)$ llamada *moduladora*. Son divididas en dos tipos:

- **Modulaciones angulares**

  Afectan a la fase y a la frecuencia de la señal. Se denomina PM (*Phase Modulation*, modulación de fase) en el caso de la fase y FM (*Frequency Modulation*, modulación de frecuencia) en el de la frecuencia. Sus expresiones matemáticas demuestran que hay poca diferencia entre ambas modulaciones; de hecho, no es posible determinar si una señal recibida tiene modulación en fase o frecuencia si no se dispone de más información. Se denominan angulares porque modifican el *pitch* de la señal.

  El efecto musical que identifica una modulación de este tipo se conoce como *vibrato*. Se da de forma natural entre los instrumentos musicales acústicos, incluida la voz de un cantante. Tiende a enriquecer el espectro del sonido producido.

- **Modulaciones lineales**

  Afectan a la amplitud de la señal, pero manteniendo su frecuencia constante. Se conocen como modulaciones AM (*Amplitude Modulation*). Auditivamente se perciben como una variación rápida en la amplitud de la señal. El efecto musical que realiza una modulación de este tipo es el *trémolo*.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

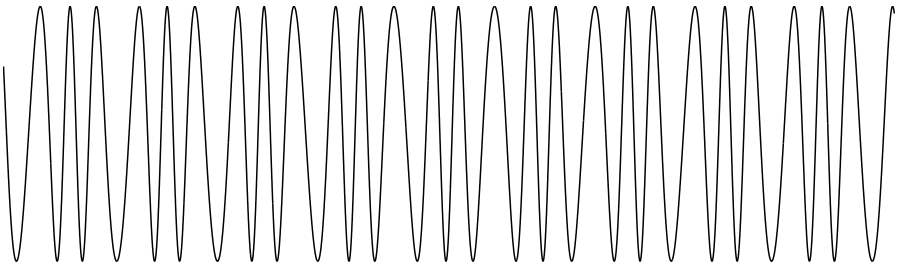


Figura 3.6: *Ejemplo de una señal con modulación angular. La señal original es un tono, y la moduladora $x(t)$ es otro tono de menor frecuencia. La señal resultante, $s(t)$, presenta variaciones en su frecuencia. Como se ha explicado, no es posible determinar si esta señal ha sido modulada en fase o en frecuencia.*

Matemáticamente, si un instrumento reproduce un sonido $s(t)$, su amplitud sufre una variación temporal determinada por $x(t)$, llamada señal *moduladora*:

$$s(t) = A_c[1 + \mu x(t)]\cos(\omega_c t) \tag{3.10}$$

donde μ es el índice de modulación, y cuantifica la variacion de la amplitud del tono, A_c. De esta manera, si $\mu = 0$ no existe modulación y si $\mu = 1$, $s(t)$ se anula (no suena).

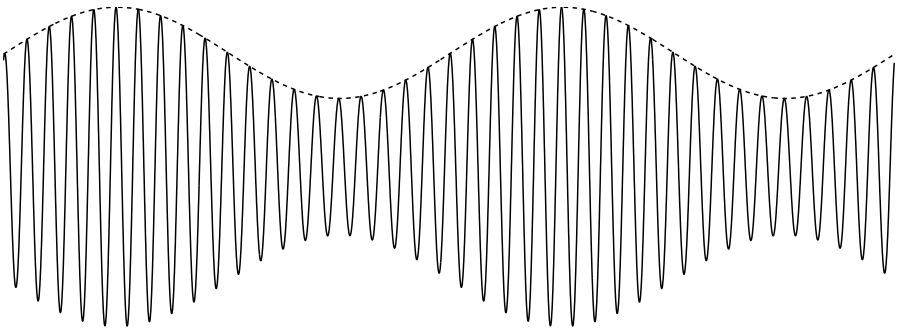


Figura 3.7: *Ejemplo de señal modulada en AM. En este caso, la señal original es una sinusoide cuya amplitud es modificada por la moduladora $x(t)$, que también es una sinusoide de menor frecuencia, y conforma $s(t)$, lo que realmente se oye.*

Para el método aquí descrito se ha estudiado si es posible aplicar las propiedades de las modulaciones al ámbito de la separación de armónicos solapados. Para obtener la amplitud y la fase de un armónico que solapa, se ha buscado la forma de estimar esos parámetros observando cómo evolucionan esos valores en los armónicos no solapados generados por la misma fuente. La herramienta principal que han empleado es la Transformada de Fourier de Tiempo Corto (*Short Time Fourier Transform*, STFT en inglés). Dividiendo la señal en tramas o *frames* de 4096 muestras (habiendo tomado 44100 muestras por segundo al digitalizarla), estudiaron la variación en amplitud de los armónicos mientras dura la señal.

Así, se comprueba que es complicado relacionar la amplitud de un armónico i con la de otro j cuando suenan simultáneamente, puesto que esa relación variará en función del tipo

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

de instrumento, de la forma en la que es tocado o de cómo fue construido. Sin embargo, la evolución de la amplitud de ese mismo i-ésimo armónico es muy similar a la que sigue el j-ésimo; es lo que se conoce como **Modulación en Amplitud Conjunta** (CAM, *Common Amplitude Modulation*). Al expresarlos como $s_i(t)$ y $s_j(t)$, siguiendo la notación empleada en (3.10), las moduladoras $x_i(t)$, $x_j(t)$ de ambos tonos resultan ser similares. Entonces, hay que determinar si es correcto predecir la evolución en amplitud de un armónico, sabiendo cómo lo hace otro que pertenece a la misma fuente. La figura 3.10 en la página 38 ilustra el efecto de esta modulación.

Demostración de la validez de la Modulación en Amplitud Conjunta

Asumiendo que en cada trama m en la que se trabaja las amplitudes de los armónicos se mantienen constantes, el modelo sinusoidal de una señal s se expresa:

$$x_s^{(m)}[n] = \sum_{h_i=1}^{H_i} a_i^{h_i}(m) \cos\left(2\pi f_i^{h_i}(m) n T_n + \phi_i^{h_i}(m)\right) \tag{3.11}$$

$a_i^{h_i}$ y $f_i^{h_i}$ son la amplitud y frecuencia del armónico i-ésimo de s, mientras que $\phi_i^{h_i}$ es su fase **al comienzo de la trama** m. Finalmente, H_i expresa el número total de armónicos de la fuente s y $T_n = 1/f_s$ es el periodo de muestreo, en segundos. Hay que aclarar que H_i sólo puede ser sabido de antemano si x_s ha sido sintetizada mediante software, empleando la expresión (3.11). Si la señal fuese real, H_i sería un valor tal que $H_i F_i(m) < f_s/2$ y $(H_i + 1)F_i(m) \geq f_s/2$. Sin embargo, en la práctica, la energía aportada por los armónicos mayores es tan baja que limitar la búsqueda hasta los 20 primeros es suficiente[1].

La Modulación en Amplitud Conjunta va a ser empleada para estimar las amplitudes de los armónicos que están solapados en frecuencia, a lo largo de las tramas que conforman la señal. Para ello, se define:

$$r_{m^* \to m}^{h_i} = \frac{a_i^{h_i}(m)}{a_i^{h_i}(m^*)} \tag{3.12}$$

Es decir, $r_{m^* \to m}^{h_i}$ refleja el cambio en la amplitud del armónico h_i desde la trama m^* a la m. Para comprobar que la suposición tomada es correcta, se comparan los armónicos con el de mayor energía de la fuente, denotado como h_i^*. Así, el coeficiente de correlación entre h_i^* y otro armónico cualquiera g_i entre las tramas m_0 y m_1 se calcula como:

$$C(h_i^*, g_i) = \frac{\sum_{l=m_0}^{m_1} r_{m^* \to l}^{h_i^*} r_{m^* \to l}^{g_i}}{\sqrt{\left(\sum_{l=m_0}^{m_1} \left(r_{m^* \to l}^{h_i^*}\right)^2\right)\left(\sum_{l=m_0}^{m_1} \left(r_{m^* \to l}^{g_i}\right)^2\right)}} \tag{3.13}$$

[1] Del conjunto de armónicos que integran un sonido, el oído humano es capaz de percibir hasta el armónico 16

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

A partir de esta expresión, se obtiene un *box plot* con el que se comprueba lo que ya se intuía en la figura 3.10: **sí existe correlación** y **es mayor cuanto menor es la diferencia de energías entre los armónicos**.

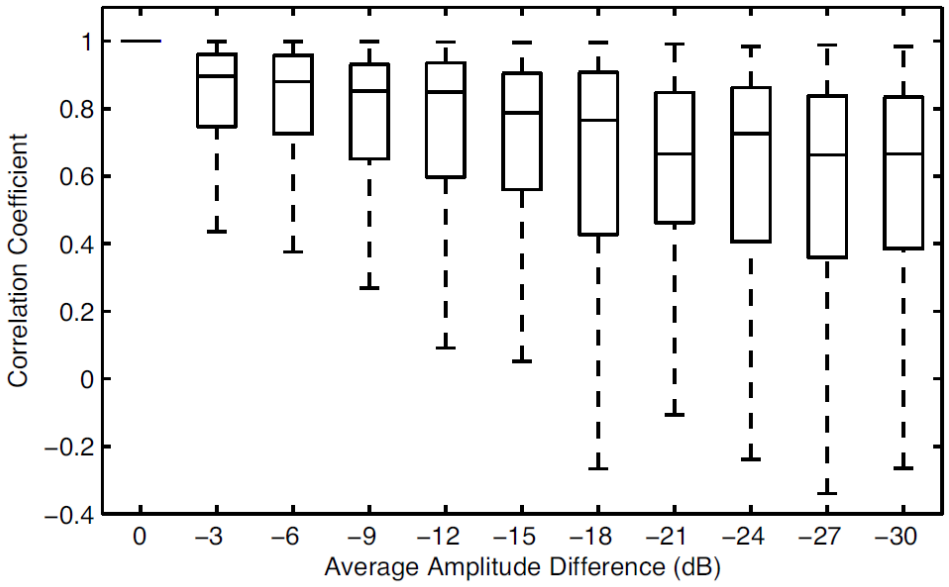


Figura 3.8: *Box plots de los coeficientes de correlación $C(h_i^*, g_i)$ obtenidos entre el armónico de mayor energía y el resto, en función de la diferencia de amplitudes entre éstos.*

Estimación de la fase

Empleando la Modulación en Amplitud Conjunta es posible estimar las amplitudes de los armónicos que solapan en frecuencia. Sin embargo, este resultado por sí mismo está incompleto, pues a cada amplitud le acompaña una fase que determina el resultado final del solapamiento.

Por ejemplo, suponiendo que haya dos armónicos cuyas frecuencias son tan cercanas que solapan en frecuencia, sus amplitudes van a ser a_1 y a_2; pero al sumarse el resultado va a ser $a = |a_1 + a_2 e^{i\Delta\phi}|$, siendo $\Delta\phi = \phi_2 - \phi_1$ la fase relativa de los armónicos. Si $\Delta\phi = 0$, entonces $a = |a_1 + a_2|$, pero si $\Delta\phi = \pi$, $a = |a_1 - a_2|$, dando un valor final muy diferente. Queda claro que la fase no puede ser obviada si se quiere conseguir un buen resultado.

Para estimarla de manera aproximada, se relaciona la frecuencia instantánea de cada armónico al inicio de una trama, con la misma al inicio de la trama inmediatamente anterior:

$$\phi_i^{h_i}(m+1) - \phi_i^{h_i}(m) = \Delta\phi_i^{h_i}(m) \approx 2\pi f_i^{h_i}(m) T_m \tag{3.14}$$

Recordando que la frecuencia de un armónico es múltiplo de la fundamental, esta expresión puede expresarse también como:

$$\Delta\phi_i^{h_i}(m) = 2\pi f_i^{h_i}(m) T_m = 2\pi h_i F_i(m) T_m \tag{3.15}$$

T_m expresa el avance de la STFT (Transformada de Fourier de Tiempo Corto, *Short Time Fourier Transform*) a lo largo de la señal. Cada trama temporal es de 2048 muestras, que tras

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

enventanarla se obtiene su Transformada de Fourier. Para elegir las siguientes 2048 muestras de la siguiente trama, este parámetro indica las muestras nuevas que se van a analizar, en este caso 1024 muestras. Como se expresa en segundos, $T_m = 1024/f_s \approx 2.3$ ms.

La figura 3.9 muestra la estimación conseguida con la expresión (3.15) junto con la variación de fase real, de una señal real. La aproximación conseguida es bastante precisa, aunque no sea capaz de predecir las variaciones "rápidas".

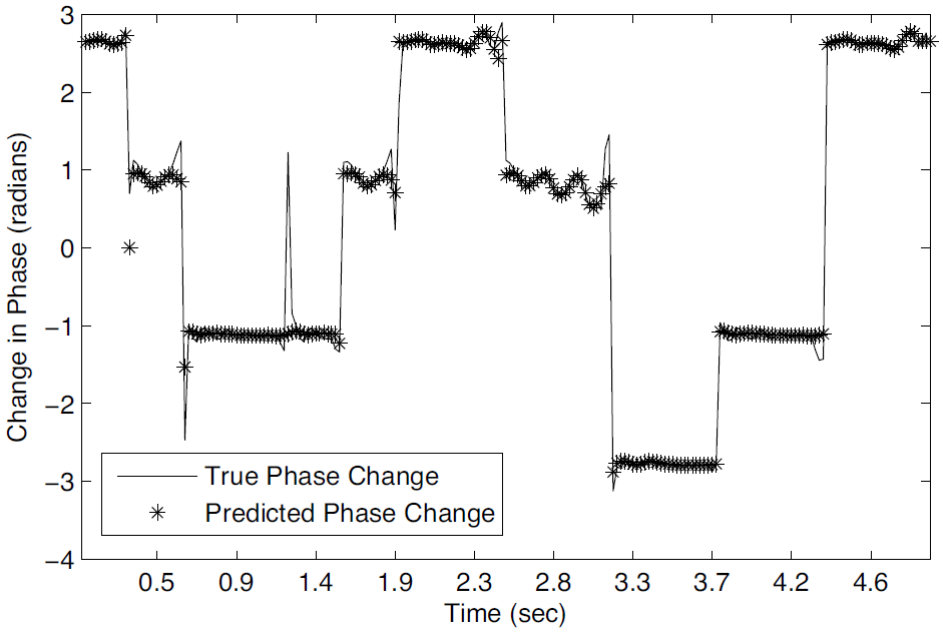


Figura 3.9: *Fase real y estimada del primer armónico durante 5 segundos del sonido de una flauta.*

Desarrollo

Como se ha explicado en el apartado 3.2, la amplitud puede ser estimada teniendo en cuenta la Modulación en Amplitud Conjunta. La fase también es obtenida a partir del *pitch* de la fuente, que se asume conocido. Además, las frecuencias de los armónicos se consideran múltiplos exactos de las frecuencias fundamentales.

El proceso se divide en tres partes. Primero se estudia el efecto del enventanado desde el punto de vista del solapamiento de armónicos; las características de la ventana empleada (forma y longitud) determinarán las regiones donde éste se produce y las muestras frecuenciales k a las que afecta. Después, se propone analizar estos solapamientos mediante matrices para falicitar los cálculos y poder por último extraer la información de las distintas fuentes por separado y así reconstruirlas independientemente.

También se muestran las ecuaciones consideradas "fundamentales", que son las que necesita resolver el algoritmo para llevar a cabo la separación.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Efectos del enventanado

Cada fuente independiente que se se busca separar puede expresarse como en la ecuación (3.11). Ahí están reflejados los dos parámetros que este algoritmo se encarga de calcular en caso de ser necesario (cuando hay solapamiento): la amplitud $a_i^{h_i}(m)$ y la fase $\phi_i^{h_i}(m)$. Es necesario pues determinar cuándo se produce o no se produce solapamiento. Esto va a depender de la **resolución frecuencial** de la ventana:

$$f_b = \frac{f_s}{N}$$

siendo N el tamaño muestral de cada trama. f_b es un valor clave a partir del cual se definen los parámetros que estudian el solapamiento.

Emplear una ventana $w[n]$ va a afectar al espectro de la trama a analizar: cuanto menor anchura tiene el lóbulo principal, mayor resolución frecuencial, es decir, se pueden distinguir armónicos que están más juntos (en comparación con el límite de otras ventanas); pero a mayor resolución, menor amplitud relativa de los lóbulos secundarios. Esto implica un compromiso, pues se busca que el espectro de la ventana sea lo más parecido a una función *delta* -$\delta(f)$- (ver capítulo 2.2). En el método aquí descrito se emplea una ventana de **Hanning**.

Una vez enventanada la trama m, se calcula la DFT:

$$X_i(m,k) = \sum_{h_i=1}^{H_i} \frac{a_i^{h_i}(m)}{2} \left(e^{j\phi_i^{h_i}(m)} W\left(kf_b - f_i^{h_i}(m)\right) + e^{-j\phi_i^{h_i}(m)} W\left(kf_b + f_i^{h_i}(m)\right) \right) \quad (3.16)$$

Aquí, W es la Transformada de Fourier en Tiempo Discreto (TFTD, *Discrete Time Fourier Transform*) de la ventana de análisis $w[n]$. Es sencillo llegar a este resultado, pues la Transformada de Fourier de una sinusoide es conocida. Además, si se tiene en cuenta que enventanar en el dominio temporal es convolucionar en el frecuencial, el resultado es la Transformada de Fourier de la ventana replicada en las frecuencias $\pm 2\pi f_i^{h_i}$ para cada armónico h_i, además de en $\pm F_i$, y escalada por las amplitudes $a_i^{h_i}(m)$ que forman la señal. Por último, el espectro de la ventana es continuo mientras que el de x_i no lo es, por lo que W sólo toma valores en las muestras frecuenciales k.

Una vez determinado f_b, se define un umbral θ_1 que se utiliza para determinar si una muestra k está relacionada en la trama m con el armónico h_i:

$$|kf_b - f_i^{h_i}(m)| < \theta_1 \quad (3.17)$$

Con este umbral se obtiene un conjunto de muestras $\mathbf{K}_i^{h_i}$ que contienen energía perteneciente a ese armónico en la trama m. Si este armónico está solapado, esas muestras también contendrán energía de otro, mientras que si no solapa, se utilizarán en la parte final del algoritmo, para reconstruir la señal individual de la fuente que la ha generado. Para determinar finalmente si dos armónicos solapan en frecuencia, se emplea un criterio similar al de (3.17), definiendo otro umbral θ_2. Así, un armónico h_i se solapa con otro armónico h_p en la trama m si se cumple:

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

$$|f_i^{h_i}(m) - f_p^{h_p}(m)| < \theta_2 \tag{3.18}$$

Tanto θ_1 como θ_2 son valores en Hz.

Como se conocen los *pitches* de todas las fuentes que conforman la señal, mediante (3.9) se calculan las muestras k en las que están centradas los distintos armónicos y se utiliza el umbral θ_1 para hallar los conjuntos de muestras asociadas a cada armónico.

♣ **Ejemplo de solapamiento**

Sea una señal formada por dos fuentes x_a y x_b cuyos *pitches* son, respectivamente, 202 Hz y 305 Hz. Para simplificar las cuentas, se considera que están formadas por el tono principal más sus 5 primeros armónicos; entonces, la condición de (3.17) permite calcular los conjuntos de muestras de cada uno de ellos. Mediante una función auxiliar programada para tal efecto, que emplea el parámetro θ_2, se ha obtenido que el segundo armónico de x_a (606 Hz) solapa con el primero de x_b (610 Hz), puesto que $610 - 606 = 4 < 44100/4096 = f_b \approx 10$ Hz (θ_2 siempre va a ser mayor que f_b). Es decir, la ventana empleada no tiene suficiente resolución. Conocida esa región de solapamiento, se combinan los conjuntos de muestras de ambos armónicos. Para 606 Hz se tienen las muestras 61, 62, 63 y 64 y para 610 Hz 60, 61, 62 y 63. Entonces, el conjunto final es $\mathbf{D} = \mathbf{K}_i^{h_i} \bigcup \mathbf{K}_p^{h_p} = \{60, 61, 62, 63, 64\}$.

Análisis matricial de los armónicos solapados

Para facilitar los cálculos del algoritmo, se emplea una notación matricial que permita obtener los valores de los armónicos solapados ($a_i^{h_i}$ y $\phi_i^{h_i}$) "de una vez": se construyen dos matrices que forman un sistema lineal, cuya solución es un vector columna con los parámetros deseados:

$$\mathbf{RS} = \mathbf{Z} \tag{3.19}$$

La matriz $\mathbf{S}$ va a contener los parámetros buscados. Si se busca resolver P armónicos solapados, será de la forma:

$$\mathbf{S} = \begin{pmatrix} S_{i_1}^{h_{i_1}}(m_0) \\ \vdots \\ S_{i_P}^{h_{i_P}}(m_0) \end{pmatrix} \tag{3.20}$$

donde las amplitudes $a_i^{h_i}$ y fases $\phi_i^{h_i}$ se han agrupado en un solo parámetro:

$$S_i^{h_i}(m_0) = \frac{a_i^{h_i}(m_0)}{2} e^{j\phi_i^{h_i}}(m_0) \tag{3.21}$$

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Entonces, se construyen las matrices $\mathbf{Z}$ y $\mathbf{R}$, definidas sólo para los valores de k donde haya **solapamiento espectral** y **entre dos tramas** m_0 **y** m; es decir, hay que construir matrices y resolver el sistema $\mathbf{RS} = \mathbf{Z}$ para cada conjunto $\mathbf{D}$ distinto obtenido (como en el ejemplo anterior).

- **Construcción de Z**

  Suponiendo que la mezcla de las I fuentes ha sido lineal, el espectro final es:

$$Z(m,k) = \sum_{i=1}^{I} X_i(m,k) \tag{3.22}$$

  Esta expresión puede "transformarse" aplicando las aproximaciones de la Modulación en Amplitud Conjunta y de la estimación de la fase a partir del *pitch*:

$$Z(m,k) = \sum_i \underbrace{\frac{a_i^{h_i}(m_0)}{2} e^{j\phi_i^{h_i}(m_0)}}_{1} \cdot \underbrace{r_{m_0 \to m}^{h_i} e^{j\sum_{l=m_0}^{m} \Delta\phi_i^{h_i}(l)}}_{2} \underbrace{\cdot W(kf_b - h_i F_i(m))}_{3} \tag{3.23}$$

  El primer término coincide con la expresión (3.21), que es lo que se quiere calcular.

  El segundo término incluye la variación del armónico h_i entre la trama m_0 a m -al igual que en la expresión (3.12)- junto con la estimación del valor de la fase calculado empleando la ecuación (3.15) a lo largo de las tramas. La Modulación en Amplitud Conjunta permite aproximar $r_{m_0 \to m}^{h_i}$ por la variación en amplitud de otro armónico h_i^* que no solape:

$$r_{m_0 \to m}^{h_i} \approx r_{m_0 \to m}^{h_i^*} \tag{3.24}$$

  Gracias a esas aproximaciones, sólo es desconocido $S_i^{h_i}(m_0)$.

  El tercer término refleja la influencia de la ventana empleada, W. Entonces, la matriz $\mathbf{Z}$ es:

$$\mathbf{Z} = \begin{pmatrix} Z(m_0,k_0) \\ \vdots \\ Z(m_0,k_1) \\ \vdots \\ Z(m,k) \\ \vdots \\ Z(m_1,k_0) \\ \vdots \\ Z(m_1,k_1) \end{pmatrix} \tag{3.25}$$

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

- **Construcción de R**

  R es la combinación de los términos 2 y 3 de (3.23):

$$R_{i,j}(m,k) = r_{m_0 \to m}^{h_i^*} e^{j \sum_{l=m_0}^{m} \Delta\phi_i^{h_i}(l)} W(kf_b - h_i F_i(m)) \tag{3.26}$$

  siendo i la fuente entre las I posibles que conforman la mezcla y j un armónico solapado de los P existentes. La matriz queda:

$$\mathbf{R} = \begin{pmatrix} R_{i_1}(m_0,k_0) & \cdots & R_{i_P}(m_0,k_0) \\ \vdots & & \vdots \\ R_{i_1}(m_0,k_1) & \cdots & R_{i_P}(m_0,k_1) \\ \vdots & & \vdots \\ R_{i_1}(m,k) & \cdots & R_{i_P}(m,k) \\ \vdots & & \vdots \\ R_{i_1}(m_1,k_0) & \cdots & R_{i_P}(m_1,k_0) \\ \vdots & & \vdots \\ R_{i_1}(m_1,k_1) & \cdots & R_{i_P}(m_1,k_1) \end{pmatrix} \tag{3.27}$$

  Una vez construidas las matrices **R** y **Z**, **S** se obtiene:

$$\mathbf{S} = (\mathbf{R}^H \mathbf{R})^{-1} \mathbf{R}^H \mathbf{X} \tag{3.28}$$

  Nota: H significa la traspuesta conjugada de la matriz.

Resolviendo este sistema se obtienen las amplitudes y fases de **todos los armónicos solapados** en la trama inicial m_0. Calcularlos a lo largo de toda la señal ahora es sencillo, no hay más que multiplicar esos valores por el segundo término de la ecuación (3.23):

$$S_i^{h_i}(m) = S_i^{h_i}(m_0) r_{m_0 \to m}^{h_i^*} e^{j \sum_{l=m_0}^{m} \Delta\phi_i^{h_i}(l)} \tag{3.29}$$

siendo m la trama de la señal en la que se están obteniendo.

Es importante ver que para poder realizar estos cálculos es imprescindible conocer $r_{m_0 \to m}^{h_i^*}$, razón por la que antes hay que extraer las amplitudes de los armónicos no solapados. Esta información es fácil de obtener mediante mínimos cuadrados:

$$a_i^{h_i^*}(m) = \frac{2 \times \sum_k |Z(m,k)| \cdot |W(kf_b - h_i^* F_i(m))|}{\sum_k |W(kf_b - h_i^* F_i(m))|^2} \tag{3.30}$$

Esta expresión se aplica a cada conjunto de muestras k asociadas a un armónico, luego si hay T armónicos no solapados en la trama m, se hará T veces.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Sintetización de fuentes por separado

Una vez obtenidas las amplitudes y fases de los armónicos solapados, el último paso es reconstruir las señales originales $x_i[n]$ que forman la mezcla $z[n]$. El proceso es distinto para las muestras frecuenciales k asociadas a armónicos solapados y las asociadas a los no solapados.

Para el primer caso, la DFT de cada trama m se obtiene:

$$\hat{X}_i^o(m,k) = S_i^{h_i}(m)W(kf_b - h_iF_i(m)) \tag{3.31}$$

En el segundo caso, la información se extrae directamente de la mezcla $Z(m,k)$:

$$\hat{X}_i^{no}(m,k) = Z(m,k) \tag{3.32}$$

Finalmente, la estimación final de la fuente i-ésima es $\hat{X}_i = \hat{X}_i^{no} + \hat{X}_i^o$. Aplicando la DFT inversa, se obtienen las distintas $\hat{x}_i[n]$.

Valores empleados

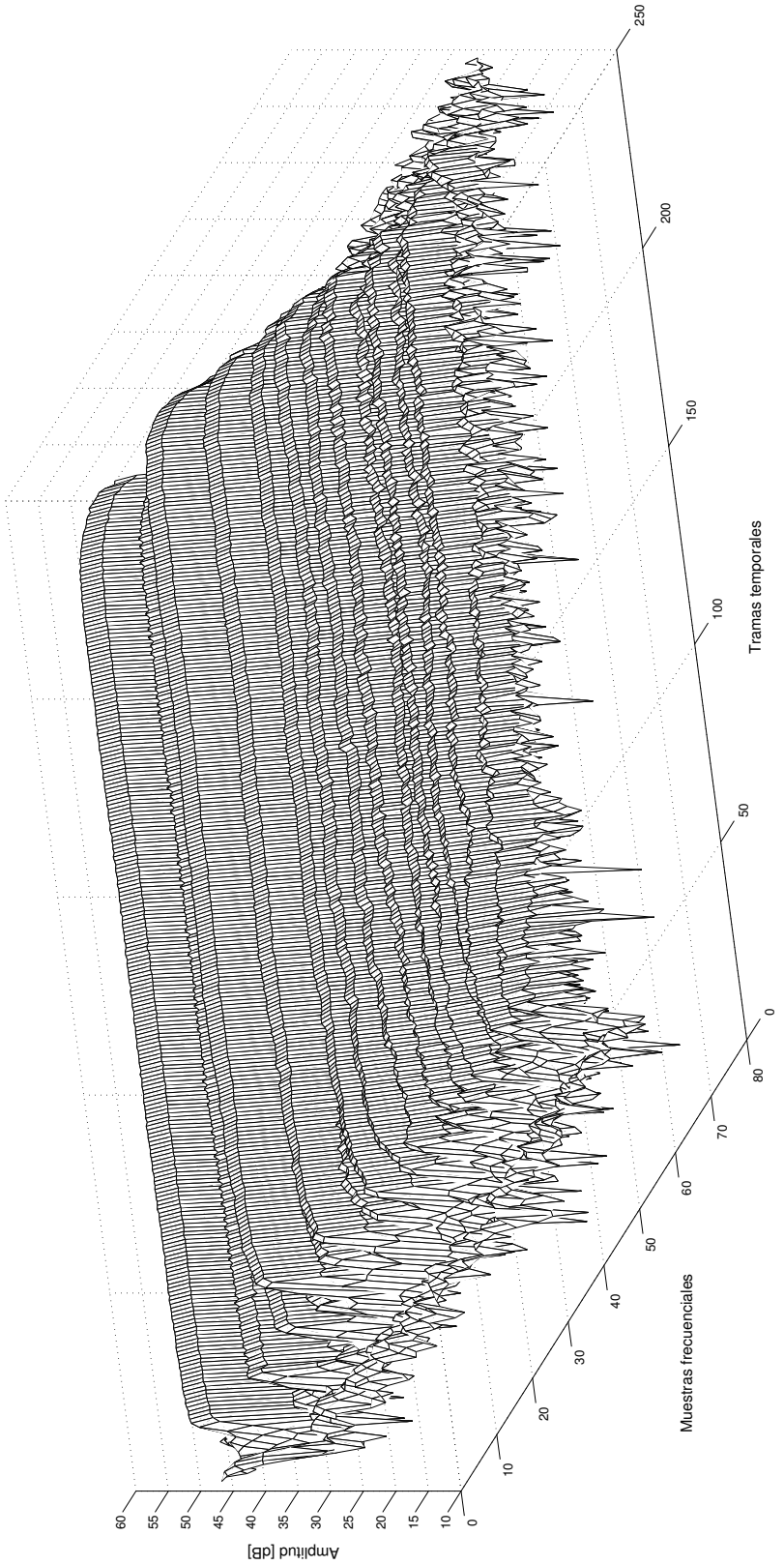
Se muestra en la tabla 3.3 los valores por defecto introducidos en el código de este algoritmo.

Tabla 3.3: *Resumen de los parámetros empleados en el método de la Modulación en Amplitud Conjunta*

Parámetro	**Valor**	**Dimensiones**	**Símbolo**
Tipo de ventana	Hanning		$W(f), w[n]$
Frecuencia de muestreo	44100	Hz	$f_s, 1/T_n$
Tamaño de la ventana	2048	muestras	N
Resolución frecuencial de la ventana	f_s/N	Hz	f_b
Tamaño de las tramas temporales	2048	muestras	L
Hop size de la STFT	1024	muestras	T_m
Umbral de decisión para asociar una muestra k con un armónico h_i	$2.5f_b$	Hz	θ_1
Umbral de decisión para determinar solapamiento entre armónicos	$1.5f_b$	Hz	θ_2

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Figura 3.10: *Evolución de varios armónicos generados por un clarinete que suena durante* 5 s*. Se ve como desde el inicio al final del sonido sus amplitudes siguen la misma envolvente, aunque escalada en amplitud. Como se ha visto en la sección 3.2, ésto puede utilizarse para resolver el problema de los armónicos solapados en frecuencia.*



upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

3.3. Modelado armónico

Esta sección describe el tercer método estudiado en este proyecto. En la introducción teórica se describen los modelos armónicos con algunos ejemplos; el desarrollo muestra el proceso de aprendizaje de los mismos y cómo se aplican conceptos de estadística para realizar la separación de las fuentes.

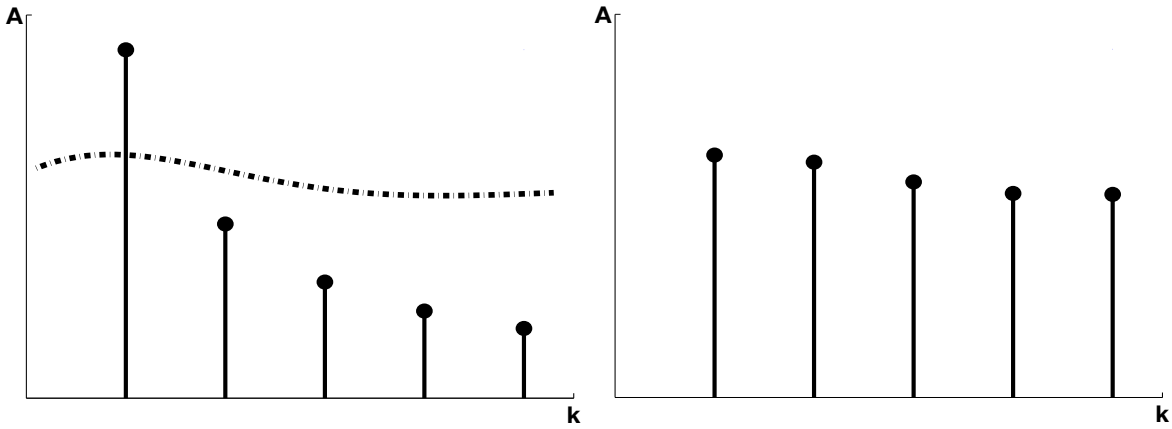
Al final se muestra una tabla que resume los parámetros más relevantes que condicionan el buen funcionamiento del método, con los valores óptimos empleados.

Introducción

Modelos AHS

Cada fuente que conforma una señal tiene diferente *timbre*: las amplitudes de sus distintos armónicos siguen un "patrón" distinto. Este patrón es generado en dos fases:

1. Una **vibración** generada por una cuerda; por ejemplo, un violín o las cuerdas vocales de un cantante.
2. Una **resonancia** producida por el cuerpo del instrumento.



(a) Ejemplo de espectro de vibración creado por un instrumento y de espectro de resonancia (línea discontinua).

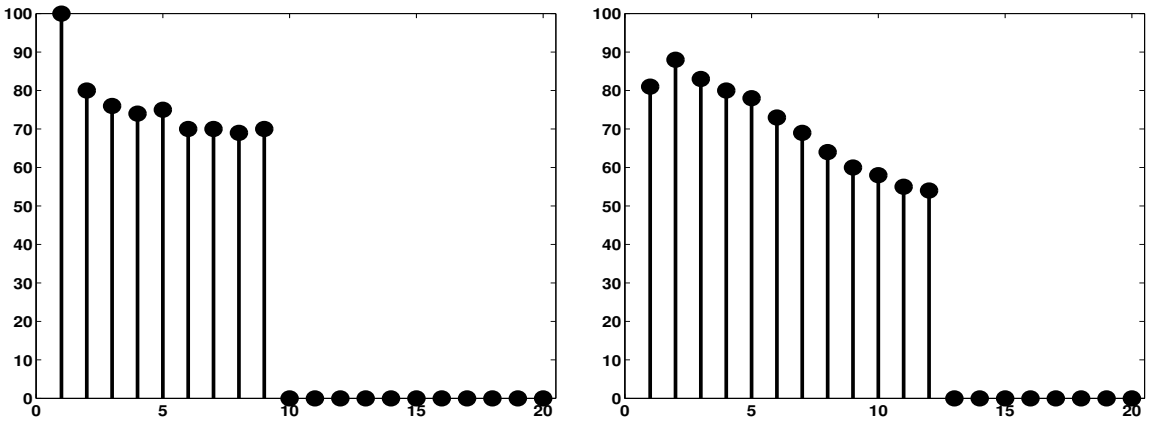
(b) Resultado final producto del espectro de vibración de un instrumento y su espectro de resonancia. Así es como se "oye".

Figura 3.11: *Muestra de como se genera un sonido armónico. El eje horizontal representa las muestras frecuenciales y el eje vertical la amplitud.*

La señal resultante es el producto de los armónicos generados y del "perfil espectral" que genera la caja de resonancia. Cada instrumento presenta un perfil único que genera un "patrón" distinto del resto (figura 3.11), denominado **estructura armónica media** (AHS, *Average Harmonic Structure*). Conocer esa información para cada fuente independiente es necesario para poder realizar su correcta extracción (las figuras 3.12(a), 3.12(b), 3.13(a) y 3.13(b) muestran

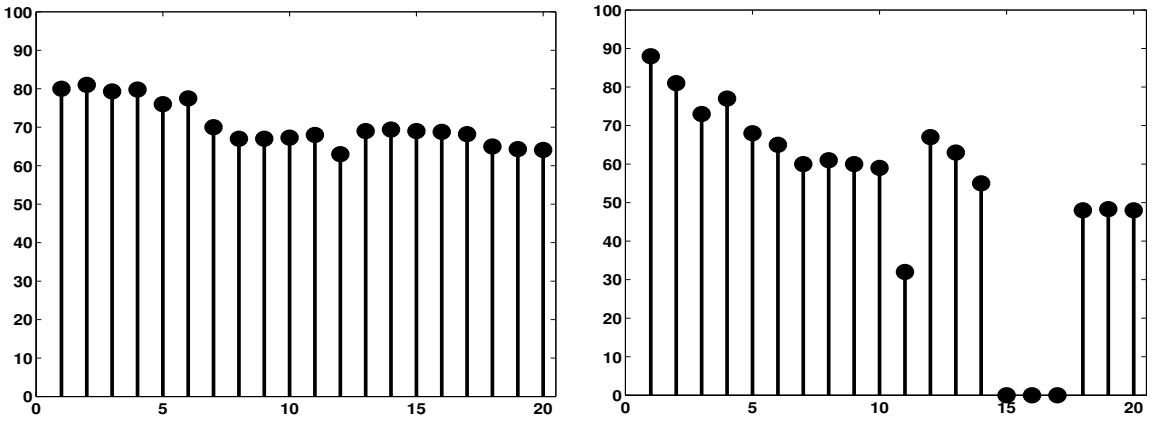
upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

ejemplos de estructuras armónicas). El término *Average* hace referencia a que los valores obtenidos son una media de los que se extraen a lo largo de todas las tramas, pues un cambio de *pitch* implica una variación en las amplitudes de los armónicos. Esto no es un problema puesto que está estudiado que estos valores no varían mucho aunque sí lo haga la frecuencia fundamental, cosa que no ocurre en el caso de la voz: los armónicos generados por la voz humana **sufren mucha dispersión**, lo que evita que una pista de esas características pueda extraerse de igual manera que el resto. No obstante, es posible extraer una -y como mucho una- pista de este tipo, aunque con ciertas deficiencias. Además, hay que resaltar que a diferencia de los dos métodos explicados en las secciones 3.2 y 3.1, éste no se ha enfocado en resolver el problema de los armónicos solapados en frecuencia específicamente.



(a) Estructura armónica típica de un violín (b) Estructura armónica típica de una trompeta

Figura 3.12: *Estructuras armónicas de dos instrumentos para un tono/pitch concreto en una trama. El eje horizontal corresponde al número de armónicos y el vertical a sus amplitudes en* dB.



(a) Estructura armónica de una voz masculina (b) Estructura armónica de una voz femenina

Figura 3.13: *Estructuras armónicas de las voces de dos cantantes para un tono/pitch concreto en una trama. El eje horizontal corresponde al número de armónicos y el vertical a sus amplitudes en* dB.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Desarrollo

Este método consta de dos fases. Primero es necesario un **proceso de aprendizaje** que realiza un análisis de la señal a partir del cual se extraen los picos espectrales significativos, que son agrupados en **clusters**, tantos como fuentes que forman la señal. Éstos clusters contienen pares amplitud-frecuencia que se emplean para obtener las estructuras armónicas. El siguiente paso es usar esa información para clasificar los picos espectrales de cada trama como pertenecientes a una u otra fuente y así construir los espectros por separado. Para llevarlo a cabo, se emplea el concepto estadístico de **máxima verosimilitud**, que aplicado a este contexto, va a permitir clasificar los distintos picos existentes en las tramas como pertenecientes a una u otra fuente. Finalmente, se obtienen las señales en el dominio temporal mediante la Transformada Inversa de Fourier.

Aprendizaje de los modelos AHS

1. **Detección de picos**

   Los picos significativos, que corresponden a los armónicos de las fuentes, son máximos locales de la trama, aunque pueden aparecer más picos producidos por combinaciones de lóbulos secundarios surgidos debido al enventanado y/o ruido. Éstos son desechados tras un proceso de filtrado como el que se muestra en la figura 3.14: tras realizar un filtrado Gaussiano a lo largo de la trama (línea punteada), los máximos significativos han de estar 8 dB por encima (línea gruesa), a la vez que no pueden tener una diferencia mayor de 50 dB respecto al máximo de la trama (línea de puntos). Estos umbrales son valores elegidos intuitivamente y no son determinantes en el resultado final, siendo el filtrado lo que realmente elimina los picos "innecesarios". El resultado de esta detección es un conjunto de $f_1, \ldots, f_K$ frecuencias y $A_1, \ldots, A_K$ amplitudes.

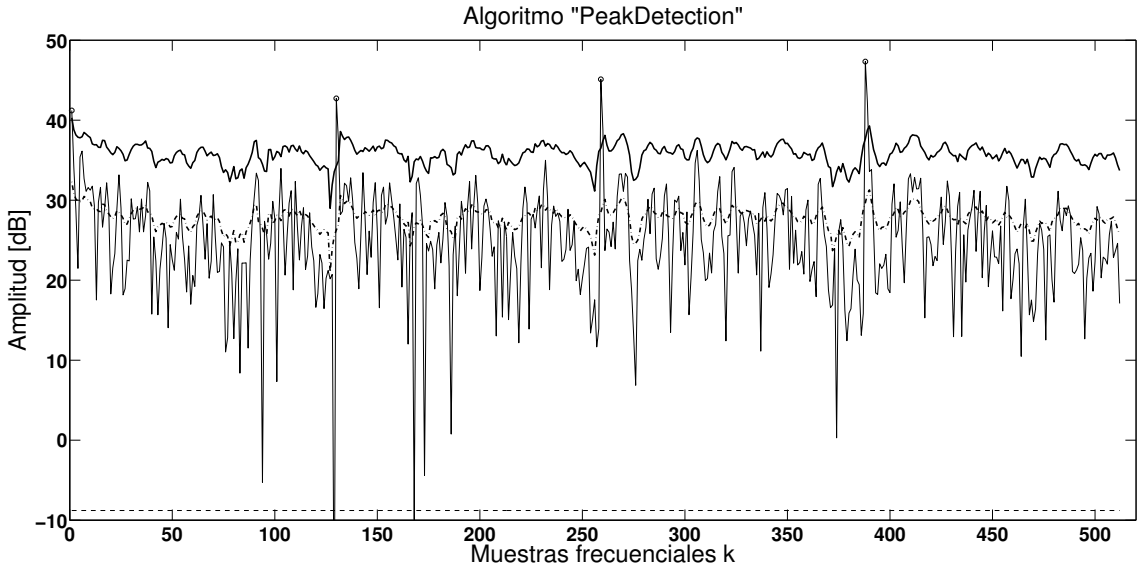


Figura 3.14: *Resultado mostrado por el detector de picos que incluye el algoritmo.*

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

2. **Extracción de la estructura armónica**

Una vez obtenidos los K picos de la trama, éstos corresponden a los distintos *pitches* que la conforman junto a sus respectivos armónicos. Para determinar esas N fuentes (valor que se asume conocido) y sus frecuencias f_0^1, f_0^2, ..., f_0^N se aplica la función de máxima verosimilitud:

$$p(O|f_0^1, f_0^2, \ldots, f_0^N) = p(f_1, f_2, \ldots, f_K|f_0^1, f_0^2, \ldots, f_0^N) = \prod_{i=1}^{K} p(f_i|f_0^1, f_0^2, \ldots, f_0^N) \tag{3.33}$$

En esta expresión, la observación O es la información obtenida de la trama en la que se está trabajando, que equivale a las K frecuencias y amplitudes obtenidas en el paso anterior. Como no es seguro que en cada trama estén presentes las N fuentes, así se encuentra la "combinación" que más se aproxima a la observación.

Para modelar la verosimilitud de un pico $p(f_i|f_0^1, f_0^2, \ldots, f_0^N)$ se calcula la desviación frecuencial de f_i respecto de los armónicos de las N fuentes. Es decir, se trata de averiguar qué armónico de qué fuente es f_i. Hay que recordar que de las N posibles fuentes que pueden sonar simultáneamente, no tienen por qué existir todos sus armónicos. Por ejemplo, en la figura 3.12(a) no hay armónicos por encima del décimo, pero en el caso de la voz humana masculina, se generan hasta 17 armónicos (figura 3.13(a)).

La desviación frecuencial es el mínimo de las N desviaciones que se pueden calcular:

$$d^2(f_i) = \min_j d^2(f_i, f_0^j) \tag{3.34}$$

$$d(f_i, f_0^j) = \frac{f_i/f_0^j - \left[f_i/f_0^j\right]}{\left[f_i/f_0^j\right]} \tag{3.35}$$

En la expresión (3.35), $[\,\cdot\,]$ representa un redondeo al entero más próximo. La función de verosimilitud se calcula:

$$p(f_i|f_0^1, f_0^2, \ldots, f_0^N) = \frac{1}{C_1} \exp\left\{-\frac{d^2(f_i)}{2\sigma_1^2}\right\} \tag{3.36}$$

σ_1^2 representa la desviación estándar, siendo $\sigma_1 = 0.03$ para permitir que f_i pueda oscilar medio semitono alrededor de la frecuencia ideal y aún así ser cosiderada como un armónico, y C_1 es el factor de normalización. La verosimilitud se modela como una distribución Gaussiana de $d(f_i)$.

Este resultado por sí mismo no es suficiente para determinar las distintas frecuencias f_i. Debido al cálculo de un mínimo en la expresión (3.34), conforme se añaden más *pitches* en la expresión (3.33), su resultado incrementará y tenderá a $1/C_1^K$ conforme N tiende a infinito; para encontrar la solución correcta, tras calcular la verosimilitud dada en (3.33) se emplea un parámetro que "penaliza" el valor obtenido proporcionalmente a N, para

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

evitar que el resultado incluya más *pitches* de los que realmente hay. Se denomina **criterio Bayesiano** (BIC, *Bayesian Information Criterion* en inglés):

$$\text{BIC} = \ln p(O|f_0^1, f_0^2, \ldots, f_0^N) - \frac{1}{2}N \ln K \tag{3.37}$$

Así, la solución final de la búsqueda ha de maximizar (3.37).

3. **Agrupación**

   Una vez clasificadas las K frecuencias como *pitches* y sus correspondientes armónicos, es necesario emplear un algoritmo que sea capaz de agrupar las distintas f_j para generar las estructuras armónicas de las fuentes. El resultado será un conjunto de clusters, uno por cada fuente independiente; cada cluster a su vez será un conjunto de pares amplitud-frecuencia. Como se ha establecido que no haya más de 20 armónicos, será un vector 20-dimensional, aunque no se hayan encontrado 20 armónicos por fuente.

Separación de las fuentes

Cada fuente independiente es una señal $s(t)$ armónica, de manera que se puede expresar como:

$$s(t) = \sum_{r=1}^{R} A_r(t) \cos[\theta_r(t)] + e(t) \tag{3.38}$$

$e(t)$ representa el ruido que acompaña a la señal, $A_r(t)$ es la amplitud instantánea, R es el número máximo de armónicos de cada fuente y $\theta_r(t) = \int_0^t 2\pi r f_0(\tau) d\tau$ es la fase instantánea, representando $f_0(\tau)$ la frecuencia fundamental del r-ésimo armónico en el instante τ. Además, R es tomado como 20 por defecto, puesto que los armónicos mayores de ese valor presentan amplitudes tan bajas que se ven absorbidos por los lóbulos laterales de otros armónicos más fuertes; ésto ocurre debido al enventanado que se aplica a cada trama antes de obtener su Transformada de Fourier, tal y como se explica en el apartado 3.2. Por último, se asume que la amplitud instantánea se mantiene constante en una trama temporal (2048 muestras), así $A_r(t)$ pasa a ser A_r^l: la amplitud del r-ésimo armónico en cada trama l. Estos valores son convertidos a unidades de dB:

$$B_r^l = \begin{cases} 20 \log_{10}(A_r^l) & \text{si } A_r^l > 1 \\ 0 & \text{resto} \end{cases} \tag{3.39}$$

La **estructura armónica media** de $s(t)$ ($\bar{\mathbf{B}}$) se obtiene promediando los B_r^l de cada trama:

$$\bar{\mathbf{B}} = [\bar{B}_1, \ldots, \bar{B}_R] \tag{3.40}$$

$$\bar{B}_i = \frac{1}{L_i} \sum_{l=1, B_i^l \neq 0}^{L_i} B_i^l, \quad i = 1, \ldots, R. \tag{3.41}$$

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Además, de manera similar a como se calcula $\bar{\mathbf{B}}$ se obtiene otro parámetro que mide su **variación** a lo largo de la señal y que es empleado más adelante en el proceso de separación de las fuentes. Se trata de la **Inestabilidad Armónica Media**, HSI:

$$HSI = \frac{1}{R}\sum_{i=1}^{R}\left(\frac{1}{L_i}\sum_{l=1,B_i^l\neq 0}^{L^i}(B_i^l - \bar{B}_i)^2\right) \tag{3.42}$$

Así, si la señal original $y(t)$ tiene N fuentes, habrá N estructuras armónicas ($\bar{\mathbf{B}}$) que se calculan a partir de (3.40) y (3.41). Hay que resaltar que **las distintas A_r^l son obtenidas directamente de** $y(t)$ mediante el proceso de aprendizaje, puesto que **no se conocen las** $s_1(t), s_2(t), \ldots, s_N(t)$ por separado.

El proceso final de extracción de las fuentes independientes consta a su vez de tres pasos:

1. Primero, de manera similar a como se hace en (3.33), se emplea la función de máxima verosimilitud para determinar los armónicos que sí han sido generados por f_0. Es decir, conociendo la estructura armónica del instrumento con *pitch* f_0, $\bar{\mathbf{B}}$, se determina si el armónico de amplitud y frecuencia A_i y f_i corresponde a esa fuente:

   $$p(f_i, A_i|f_0, \bar{\mathbf{B}}) = p(f_i|f_0, \bar{\mathbf{B}}) \cdot p(A_i|f_i, f_0, \bar{\mathbf{B}}) = p(f_i|f_0) \cdot p(A_i|f_i, f_0, \bar{\mathbf{B}})$$

   Tanto $p(f_i|f_0)$ como $p(A_i|f_i, f_0, \bar{\mathbf{B}})$ son modeladas como distribuciones Gaussianas. La primera es una distrución de $d(f_i, f_0)$, que representa lo que se "aleja" f_i del valor que debe tener un armónico generado por f_0, mientras que la segunda modela la diferencia entre la amplitud A_i y el valor que teóricamente ha de tomar, que lo proporciona $\bar{\mathbf{B}}$. Entonces, la expresión anterior se calcula:

   $$p(f_i|f_0) \cdot p(A_i|f_i, f_0, \bar{\mathbf{B}}) = \frac{1}{C_2}\exp\left\{-\frac{d^2(f_i, f_0)}{\sigma_1^2}\right\} \times \exp\left\{-\frac{D^2(A_i, \bar{\mathbf{B}})}{\sigma_2^2}\right\} \tag{3.43}$$

   Como se ha indicado en la página 42, $\sigma_1 = 0.03$ y ahora $\sigma_2 = HSI$, que se calcula en (3.42) a partir de la estructura armónica $\bar{\mathbf{B}}$. C_2 es el factor de normalización, similar al de la expresión (3.36).

   Con este primer paso, se eliminan armónicos que no pertenecen a ninguna fuente pero que han sido detectados en el proceso de aprendizaje y considerados como componentes de las estructuras armónicas. Es posible corregir los errores producidos en el proceso de aprendizaje puesto que en la expresión (3.43) se conoce $\bar{\mathbf{B}}$, que contiene información de todas las tramas de la señal, mientras que la expresión (3.36) se calcula condicionada a una trama. Es en este paso donde se realiza la clasificación definitiva de los armónicos y se asignan a la fuente correspondiente.

2. Se comienzan a extraer los armónicos de la señal original trama a trama para construir el espectro de cada fuente por separado. Este paso puede efectuarse de dos formas: en **cascada** o en **paralelo**. Ambas técnicas tienen ciertas ventajas y desventajas:

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

- En paralelo, las fuentes se reconstruyen extrayendo la información desde el espectro original, lo que evita que se produzcan errores encadenados; sin embargo, no permite extraer fuentes con una variación (HSI) alta, como es el caso de la voz humana, pues de una trama a otra se producen grandes fluctuaciones en las amplitudes de los armónicos así como en el número e índice de los armónicos generados, hecho que impide el correcto funcionamiento de este método.
- En cascada, primero se extraen los armónicos de una fuente y se elimina esa información del espectro original, para después extraer y eliminar los armónicos de la siguiente y así sucesivamente. Si se tiene una fuente muy variable, los armónicos restantes corresponderán a esa fuente, que no puede ser extraída implementando un proceso en paralelo. Sin embargo, también contendrá ruido generado por los errores cometidos en las extracciones anteriores.

En este caso se ha implementado el proceso en cascada, para poder extraer no sólo instrumentos, sino también una pista de voz.

3. El último paso consiste en aplicar la Transformada Inversa de Fourier a las tramas espectrales obtenidas, y así se obtienen las fuentes estimadas por separado.

Valores empleados

Se muestra en la tabla 3.4 los valores por defecto introducidos en el código de este algoritmo.

Tabla 3.4: *Resumen de los parámetros empleados en el método del Modelado Armónico*

Parámetro	**Valor**	**Dimensiones**	**Símbolo**
Tipo de ventana	Hamming		
Frecuencia de muestreo	44100	Hz	
Tamaño de la ventana	2048	muestras	L
Tamaño de las tramas temporales	2048	muestras	
Estructura armónica media	$\bar{\mathbf{B}}$		AHS
Inestabilidad armónica			HSI
Número máximo de armónicos	20		R
Fuentes independientes			N
Desviación estándar de la frecuencia	0.03		σ_1
Desviación estándar de la amplitud	HSI de la fuente		σ_2
Hop size de la STFT	1024	muestras	
Zero-padding	2048	muestras	

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Capítulo 4
Fase experimental

Una vez explicados los conceptos y metodología de los tres métodos a analizar, esta sección muestra la calidad de la separación de fuentes de cada uno de ellos. Como se ha explicado en la sección 2.2, la calidad de las señales obtenidas por cada algoritmo va a depender de lo bien que separen los armónicos que solapan en frecuencia.

En la tabla 2.2 en la página 8 se muestran las frecuencias fundamentales de las notas *C*4, *D*4, *E*4, *F*4, *G*4 y *A*4, y en la sección 2.2 se demuestra cómo los armónicos de esas señales están en múltiplos fundamentales de esas frecuencias. Para efectuar la separación de las fuentes, el algoritmo ha de extraer los parámetros de los armónicos, tanto de los que solapan como de los que no. Con esos datos se reconstruye la señal $\hat{x}[n]$, que va a ser una estimación de la original $x[n]$.

Condiciones de simulación

Las señales empleadas son generadas de manera artificial. Se emplean sinusoides que incluyen:

- Bajo nivel de ruido: $SNR \simeq 40$ dB.
- Modulación AM: índice de modulación de 0.5.
- Modulación FM: índice de modulación de 2.

Estos parámetros se incluyen para que las señales generadas, aún siendo artificiales, tengan modulaciones tanto en amplitud como en frecuencia que aportan riqueza al sonido -vibrato y trémolo-. Además, se han construido de forma que no tengan más de 20 armónicos además de la frecuencia fundamental.

Nota Los índices de las modulaciones se mantienen constantes aunque se modifique el *pitch* de la señal.

Como se ha explicado en la sección 2.1, las frecuencias de estas señales están definidas, así como las de todos sus armónicos. Para estudiar el comportamiento de los algoritmos, las mezclas están compuestas por dos pistas: una fija que corresponde a la nota *C*4 (261.6 Hz), y 5 sonidos distintos que corresponden a las notas *D*4, *E*4, *F*4, *G*4 y *A*4. La relación que existe

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

entre sus frecuencias fundamentales va a producir que un porcentaje variable de sus armónicos estén solapados en frecuencia. La tabla 4.1 muestra el porcentaje de armónicos que solapan y la energía respecto del total que representan.

Tabla 4.1: *Porcentaje de armónicos que solapan y de la energía que representan.*

Nota musical	*D*4	*E*4	*F*4	*G*4	*A*4
Solapamiento (%)	9.52	4.76	23.81	33.33	4.76
Energía (%)	1	4	10	23	7

La aparente incongruencia que se da en algunos valores (por ejemplo, la nota *E*4 tiene un 5 % de solapamiento que aporta un 4 % del total de la señal, mientras que *D*4 tiene un 1 % con un 9.5 % de solapamiento) se debe a que algunos de los armónicos que solapan tienen órdenes muy altos, por lo que apenas aportan energéticamente.

Los parámetros que van a determinar el comportamiento de los algoritmos son:

- **Signal to Residual Ratio**

  A partir de las pista estimada $\hat{x}[n]$ y de la original $x[n]$, se compara **la energía de las tramas** de $x[n]$ respecto de la diferencia con su estimación:

$$SRR_{\mathrm{dB}} = 10\log_{10}\left[\frac{\sum |x_1[n]|^2}{\sum |x_1[n] - \hat{x}_1[n]|^2}\right] \tag{4.1}$$

- **Frequency Domain Signal to Residual Ratio**

  A partir de los espectros $\hat{X}[k]$ y $X[k]$, éstos se comparan en las regiones donde se ha producido solapamiento, a partir de **la energía de las muestras implicadas**:

$$FDSRR_{\mathrm{dB}} = 10\log_{10}\left[\frac{\sum_{k_1}^{k_2} |X(k)|^2}{\sum_{k_1}^{k_2} |X(k) - \hat{X}(k)|^2}\right] \tag{4.2}$$

  $FDSRR_{\mathrm{dB}}$ permite poner a prueba la capacidad de cada método evitando que los verdaderos resultados sean enmascarados por los armónicos que no solapan, como si ocurre para la SRR_{dB}.

Los resultados se muestran en dos figuras que muestran la calidad de las estimaciones conforme la presencia del solapamiento crece en la mezcla.

Resultados

Los resultados de las simulaciones se muestran en las tablas 4.1 y 4.2. El eje vertical corresponde a los valores obtenidos en dB para SRR y $FDSRR$, mientras que el eje horizontal está

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

ordenado de menor a mayor porcentaje de energía implicada en el solapamiento; este orden no coincide si se hiciese de menor a mayor valor del *pitch*. Sin embargo, se ha considerado más relevante establecerlo según el nivel de solapamiento, pues en eso se centra este proyecto.

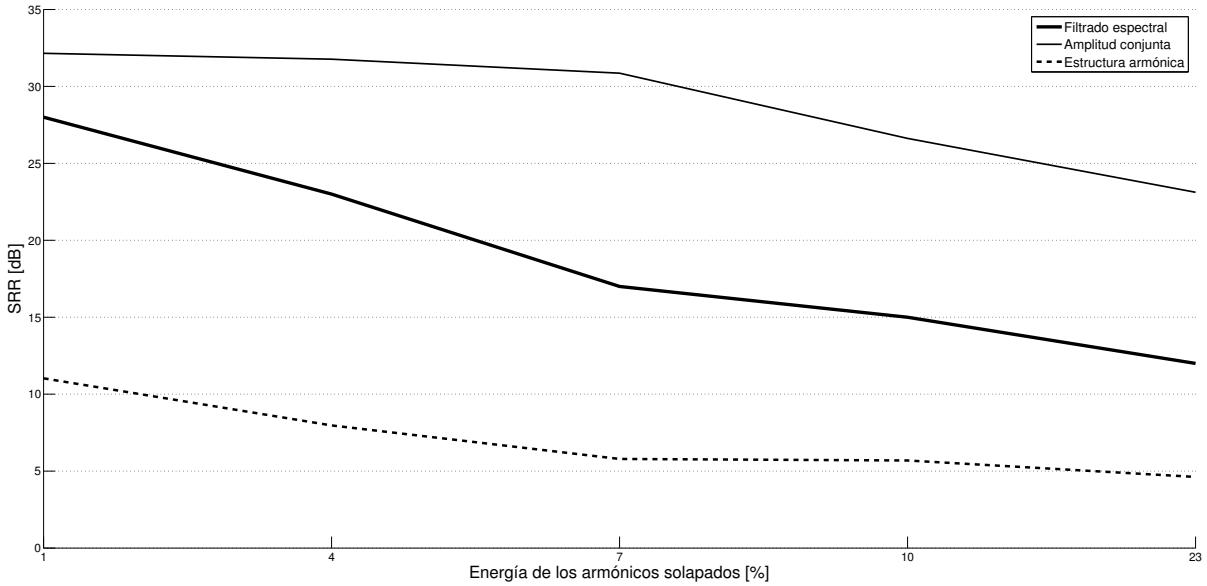


Figura 4.1: *SRR obtenido con los distintos métodos para distintos valores de solapamiento.*

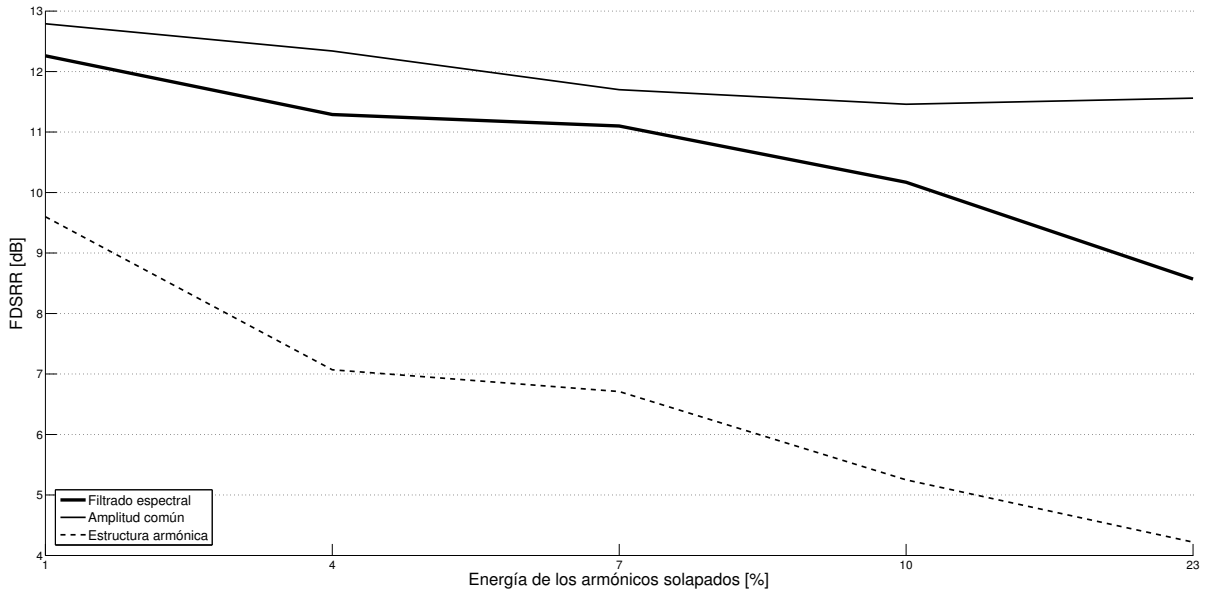


Figura 4.2: *FDSRR obtenido con los distintos métodos para distintos valores de solapamiento.*

Análisis de los resultados

Las imágenes de la sección anterior muestran una clara diferencia de resultados entre los tres métodos. El método del modelado armónico ofrece los peores resultados; la figura 4.1 muestra

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

que conforme aumenta el solapamiento, la calidad de las estimaciones empeora, siguiendo un patrón similar en la figura 4.2. De estos resultados se deduce que:

- La calidad con la que este algoritmo resuelve los armónicos solapados es baja: 11 dB en el caso más favorable frente a 5 dB en el caso peor.
- No existe diferencia de calidad entre los armónicos que solapan y los que no solapan: de 11.03 dB a 4.62 dB para la SRR y de 9.60 dB a 4.22 dB para la $FDSRR$.

La conclusión para este método es que ofrece los peores resultados globales. Una posible causa de este bajo rendimiento es que no ha sido diseñado para resolver con éxito el solapamiento, sino que se centra en estimar la distribución de energía de los picos "ignorando" este problema. Además, para su puesta en marcha es necesario definir un número elevado de parámetros, además de los mostrados en la tabla 3.4. Para conseguir la máxima calidad en la extracción, la primera parte del proceso descrito en el desarrollo del método ha sido "simulada": al programa se le han suministrado los modelos armónicos de las señales empleadas, que son bien conocidos al ser de origen artificial. Aún así, no se han conseguido mejores resultados que en los otros métodos.

Existe pues un problema de índole práctica, ya que por norma general no es posible suministrarle la información de los modelos -cualquier situación que implique trabajar con señales reales-. Entonces, es necesario conocer a fondo el funcionamiento del programa, si bien en el proceso de aprendizaje de los modelos los parámetros óptimos varían en cada caso, luego es necesario un estudio previo para determinar la mejor configuración. Esta situación es totalmente opuesta a la que se describe en la sección 1.2 de la página 6: ayudar a un usuario **inexperto** a elegir el mejor método para su escenario de trabajo. Por lo tanto, la conclusión es que este método **no sirve para los propósitos establecidos**.

Para el método de la Modulación en Amplitud Conjunta, los resultados muestran una disminución de la calidad con el aumento de la energía de los armónicos solapados. Es capaz de extraer con buena calidad -12.79 dB- cuando el solapamiento no es significativo, y en el peor caso mantiene una $FDSRR$ de 11.56 dB; estos resultados lo convierten en el mejor de los tres métodos en cuanto a la capacidad de resolución de los armónicos solapados. La SRR que consigue es la más alta obtenida aunque disminuye conforme aumenta el solapamiento, aunque el menor valor (23.12 dB) es superior al mejor conseguido con el modelado armónico, por ejemplo. La buena calidad de los armónicos solapados extraídos ratifica la validez de la Modulación en Amplitud Conjunta, que está demostrada a partir de la página 30.

Este método es adecuado para que lo utilice un usuario inexperto, pues **mantiene la calidad independientemente del nivel de solapamiento**. Esto, añadido a que el programa **no requiere de configuración adicional** salvo determinar tipo y tamaño de la ventana, tamaño de las tramas y umbral de solapamiento -y si se usan los valores por defecto mostrados en la tabla 3.3, no es necesario configurar nada- le dota de una gran utilidad práctica.

Por último, el método del filtrado espectral consigue resultados ligeramente inferiores al anterior y claramente superiores al primero, aunque no consigue mantener la calidad de los armónicos solapados cuando el nivel de solapamiento llega al $10\,\%$ y al $23\,\%$, situación en la que la calidad se reduce a la mitad: $\Delta FDSRR = 11.56 - 8.57 \approx 3$ dB. Estos resultados están unos 4 dB por encima de los del modelado armónico.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

La inferior calidad en la resolución del solapamiento afecta también a la SRR, comenzando sólo 4 dB por debajo del segundo método pero decrementando rápidamente. Estos valores también están por encima de los conseguidos con el modelado armónico. De forma similar a como funciona el método de la Modulación en Amplitud Conjunta, no es necesaria ninguna configuración inicial, salvo el tipo y tamaño de ventana, tamaño de la trama y umbral de solapamiento, lo que también dota a este método de utilidad para un uso práctico.

Conclusiones

Como se ha descrito en la sección 1.2, con los resultados aquí obtenidos se busca ayudar a los usuarios a elegir el método adecuado para realizar la separación de fuentes. Tras observar las figuras 4.1 y 4.2, se concluye que:

1. El método que peores resultados ofrece en cualquier escenario es el del modelado armónico. Ésto, sumado a la dificultad que acarrea la puesta en marcha del algoritmo, debido a los múltiples parámetros de configuración que requiere, hace que **no sea adecuada su utilización**.

2. El método de la Modulación en Amplitud Conjunta es el que extrae las estimaciones con más calidad y su funcionamiento es transparente para el usuario, lo que lo convierte en **el más adecuado** a emplear.

3. El método del filtrado espectral, si bien es igual de sencillo de utilizar, ofrece resultados peores. Sin embargo, existe una situación en la que es más apropiado emplearlo: siempre que **se anteponga la rapidez a la calidad de las estimaciones**, pues durante las simulaciones se ha comprobado que este método es capaz de extraer las estimaciones el doble de rápido: 6.18 s frente a los 12.41 s empleados por el segundo método.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira

Bibliografía

[1] Antonio López Martín: Apuntes de Teoría de la Comunicación.

[2] Mark R. Every and John E. Szymanski "Separation of Synchronous Pitched Notes by Spectral Filtering of Harmonics", IEEE Transactions on audio, speech and language processing", vol 14, no 5, September 2006.

[3] John Woodruff, Yipeng Li and DeLiang Wang, "Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation".

[4] Zhiyao Duan, Yungang Zhang, Changshui Zhang and Zhenwei Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling" IEEE Transactions on audio, speech and language processing, vol 16, no45, May 2008.

[5] Yipeng Li, John Woddruff and DeLiang Wang, "Monaural Musical Sound Separation Based on Pitch and Common Amplitude Modulation", IEEE Transactions on audio, speech and language processing, vol 17, no 7, September 2009.

[6] Tuomas Virtanen and Anssi Klapuri, "Separation of Harmonic Sound Sources Using Sinusoidal Modeling".

[7] Tuomas Virtanen and Anssi Klapuri, "Separation of Harmonic Sounds Using Linear Models for the Overtone Series".

[8] Francisco Jesús Cañadas Quesada, "Investigación y desarrollo de técnicas de estimación multi-pitch y su aplicación a la transcripción automática de señales musicales polifónicas".

[9] Alan V. Oppenheim, Ronald W. Schafer y John R. Buck, "Tratamiento de señales en tiempo discreto", 2a ed., Prentice Hall Iberia, Madrid, 2000.

[10] Víctor Domínguez Báguena y Ma Luisa Rapún Banzo, "Matlab en cinco lecciones de Numérico", Universidad Pública de Navarra, 2007.

upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa
Todos los derechos reservados
Eskubide guztiak erresalbatu dira