ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

UNIVERSIDAD DE MÁLAGA

PROYECTO FIN DE CARRERA

Herramienta software para la corrección de disonancias en música polifónica

INGENIERÍA DE TELECOMUNICACIÓN

MÁLAGA, 2012 EMILIO MOLINA MARTÍNEZ

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

UNIVERSIDAD DE MÁLAGA

Titulación: Ingeniería de Telecomunicación

Reunido el tribunal examinador en el día de la fecha, constituido por:

D./Dª. _____

D./Dª. _____

D./Dª. _____

para juzgar el Proyecto Fin de Carrera titulado:

HERRAMIENTA SOFTWARE PARA LA CORRECCIÓN DE DISONANCIAS EN MÚSICA POLIFÓNICA

Del alumno: D. Emilio Molina Martínez

Dirigido por: Dª. Ana María Barbancho Pérez

ACORDÓ POR _____ OTORGAR LA

CALIFICACIÓN DE _____

Y, para que conste, se extiende firmada por los componentes del tribunal la presente diligencia.

Málaga a ____ de _____ del ____

El/La Presidente: El/La Vocal: El/La Secretario/a:

Fdo.: _____ Fdo.: _____ Fdo.: _____

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

UNIVERSIDAD DE MÁLAGA

HERRAMIENTA SOFTWARE PARA LA CORRECCIÓN DE DISONANCIAS EN MÚSICA POLIFÓNICA

Realizado por:
Emilio Molina Martínez

Dirigido por:
Ana María Barbancho Pérez

DEPARTAMENTO DE: Ingeniería de Comunicaciones

TITULACIÓN: Ingeniería de Telecomunicación

PALABRAS CLAVE: Música, Disonancia, Polifonía, Análisis, Procesado, Síntesis.

RESUMEN:

En este Proyecto Fin de Carrera, se ha desarrollado un sistema capaz de procesar material polifónico disonante, analizar la causa de dichas disonancias y resolverlas para conseguir una versión consonante del mismo sonido. Para ello se ha descompuesto la señal musical en una componente sinusoidal más una residual, se han aplicado transformaciones a la estructura armónica original y se ha resintetizado. Por último, se ha evaluado la herramienta combinando un análisis subjetivo y objetivo de los resultados.

Málaga, Febrero de 2012

Agradecimientos

Agradezco a mi familia, y en especial a mis padres, su apoyo incondicional y la confianza que han depositado en mí durante todos estos años. Sin ellos, nada de esto habría sido posible.

Además, agradezco los momentos compartidos a mis compañeros: Alex, Javi, Alhambra, Zoraida, Alfonso... la lista es muy larga, y seguro que estoy siendo injusto al no nombrar a muchas otras personas. Gracias a ellos por todo lo que me han aportado durante mis años en la Universidad.

En lo que respecta a la elaboración de este Proyecto, algunas personas han ofrecido su ayuda desinteresada y me gustaría mencionarlas. En especial, agradezco la colaboración de Manuel y Olga por haberme permitido evaluar la herramienta con alumnos del Conservatorio Manuel Carra de Málaga. También doy las gracias a Juan Luis por sus ideas y aportaciones para este Proyecto.

Por último, me gustaría agradecer a Ana María su excelente labor como tutora. Sus revisiones precisas y su confianza en mí han sido una importante motivación para conseguir los mejores resultados posibles dentro de este Proyecto.

Índice general

Capítulo 1

Introducción

La afinación en música ha sido objeto de interés y estudio a lo largo de toda la historia. Existen evidencias de que ya en la antigua Babilonia, alrededor del 1500 a.C, la afinación de ciertos instrumentos estaba estandarizada según un metodo basado en sucesivas quintas perfectas[1] [15]. En el siglo V a.C, Pitágoras formalizó el procedimiento que daba lugar a dicha escala musical, y por ello hoy día ésta es conocida como escala *Pitagórica* [4]. Los sonidos de esta escala resultaban armoniosos, al igual que lo era la matemática y la física sobre la que se sustentaba. Esta relación entre la armonía musical y las matemáticas sigue siendo aún hoy motivo de reflexión y debate. Numerosos teóricos y científicos posteriores han aportado interesantes conclusiones acerca del misterioso fenómeno de la armonía sonora [53, 18]. Sin embargo, los grandes maestros de la afinación han sido y son los músicos, que conscientes de su importancia la manejan con enorme intuición a la hora de componer e interpretar.

Al grabar música en un estudio de grabación, la afinación es un punto clave que ha de ser muy cuidado [8]. Dependiendo del estilo, se puede llegar a invertir mucho tiempo en lograr una toma en la que no existan desafinaciones indeseadas. No obstante, en ocasiones esto se hace muy difícil por falta de recursos o sencillamente por limitaciones técnicas del intérprete. Ante este problema, se pueden encontrar numerosas herramientas en el mercado que permiten mejorar "artificialmente" la afinación de melodías monofónicas (como es el caso de la voz). Las dos herramientas más famosas para este propósito son *Melodyne Studio* (lanzado en 2001 por la empresa alemana Celemony), y *Auto-Tune* (1997) de la empresa americana Antares Technology.

Sin embargo, para el caso de instrumentos polifónicos, o varios instrumentos monofónicos sonando de forma simultánea, no existen herramientas capaces de mejorar eficazmente la afinación del conjunto de sonidos. En los últimos años son varios los productos comerciales que se han aproximado de alguna forma a la solución de este problema. El más revolucionario ha sido *Melodyne Editor*, lanzado en Noviembre de 2009 por la empresa Celemony (ver figura 1.1).

[1] Dos sonidos están a intervalo de quinta perfecta si la razón entre sus frecuencias es $\frac{2}{3}$.

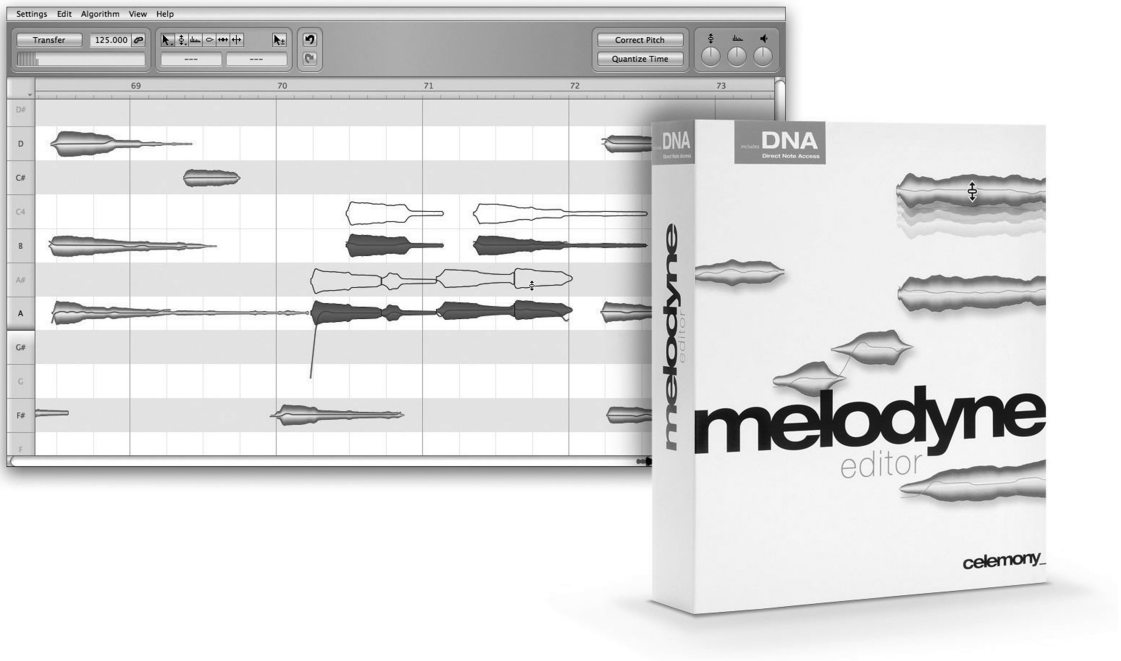


Figura 1.1: Imagen promocional de *Melodyne Editor* de la empresa alemana Celemony

En el caso de *Melodyne Editor*, la herramienta permite cierta edición independiente de los sonidos que forman una polifonía. Sin embargo, una nota no puede modificarse si no ha sido detectada como una entidad independiente, por lo que no permite procesar la polifonía como "un conjunto". Otras empresas como Antares o Apple, ya han anunciado el lanzamiento futuro de otros productos para competir con Celemony, cuyos resultados parece ser prometedores [1, 19]. Como se puede observar, el procesado de música polifónica es un terreno poco explorado que está actualmente en auge en el mercado del audio profesional.

En este Proyecto Fin de Carrera se va a desarrollar una herramienta que no ha sido implementada antes según la bibliografía consultada. Se trata de un "afinador" que procesa el audio polifónico como un todo, y permite que la relación entre armónicos de todas las notas sea consonante en su conjunto. Su implementación requerirá estudiar detalladamente el fenómeno de la disonancia sonora, así como las técnicas de análisis y síntesis más comunes para la señal musical.

1.1. Objetivos

1.1.1. Objetivo principal

El objetivo principal del Proyecto es:

Diseñar e implementar un sistema software capaz de procesar un fichero de audio que contenga material sonoro polifónico, analizar las desafinaciones indeseadas existentes en él y atenuar su efecto perceptual hasta un nivel aceptable.

Verificación de este objetivo

Resulta difícil verificar el cumplimiento de este objetivo debido a los numerosos factores subjetivos que intervienen en él y al excesivo número de casos a evaluar. Por ello la verificación del objetivo se realizará de la siguiente forma:

1. Se cuantificará el éxito obtenido atendiendo tanto a criterios subjetivos como a criterios objetivos. Para la valoración subjetiva se pondrá a prueba el sistema con personas de amplia experiencia musical. Para la valoración objetiva, se recurrirá a alguno de los numerosos algoritmos para la cuantificación de la disonancia percibida.

2. Esta evaluación se realizará con sonidos de complejidad creciente, observando el comportamiento del sistema en cada caso.

1.1.2. Objetivos secundarios

Además del objetivo principal, se han planteado una serie de objetivos secundarios para este Proyecto:

- Estudiar qué modelo de señal es el más adecuado para lograr una manipulación versátil del material musical acorde con las necesidades del Proyecto.

- Estudiar la bibliografía relacionada con el concepto de *disonancia* para adaptarlo computacionalmente a las necesidades del sistema desarrollado.

- Estudiar de qué manera se puede manipular la estructura armónica para reducir la disonancia percibida. Para ello se realizará un estudio de las diferentes escalas musicales y las relaciones armónicas existentes en ellas.

- Estudiar las técnicas de síntesis más importantes y aplicarlas a las necesidades concretas del sistema.

- Implementar un prototipo del sistema que permita estudiar con detalle su comportamiento.
- Diseñar los experimentos necesarios para realizar una correcta evaluación del sistema. Para ello se generará un banco de señales de prueba, combinando tanto señales reales como señales sintéticas.

1.2. Arquitectura general del sistema

El Sistema desarrollado tiene como entrada una señal de audio mono $x[n]$, y ofrece como salida la versión procesada $y[n]$. El diseño está basado en un esquema de *análisis-resíntesis*. En este esquema, cada parcial[2] de la señal $x[n]$ se parametriza en una primera fase de análisis, y posteriormente estos parámetros son manipulados y utilizados para resintetizar una nueva versión de la señal. Durante la fase de análisis también se extrae una componente residual que contiene todo aquello que no se haya parametrizado. Esta componente es necesaria para mantener la naturalidad del sonido original. En el apartado 3.5 se explican detalles sobre esta operación. En la figura 1.2 se muestra un diagrama de bloques general del sistema, y se explican los aspectos más importantes de su funcionamiento.

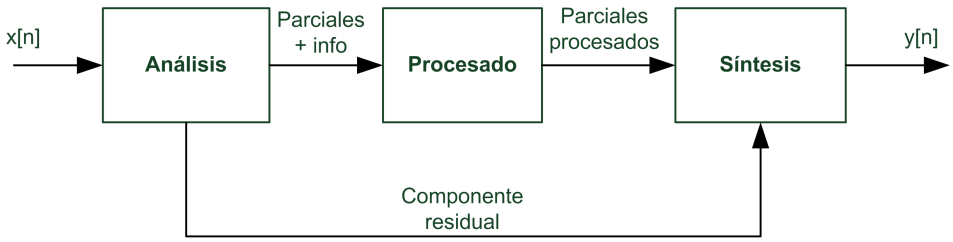


Figura 1.2: Diagrama de bloques general del sistema

En el diagrama se pueden observar tres bloques, que representan tres subsistemas funcionales diferentes:

1. **Subsistema de Análisis:** Es el encargado de extraer toda la información útil de la señal que va a servir para el procesado y la síntesis del sonido final. Además en este bloque se realiza la separación de la parte sinusoidal y la parte residual de la señal.

[2]Un parcial es cada una de las componentes sinusoidales que forman una señal musical, sea ésta armónica o inharmónica.

2. **Subsistema de Procesado:** En este bloque es donde se introducen modificaciones sobre el modelo de componente sinusoidal creado por el bloque de análisis. Lo que se hace es mover los parciales en frecuencia para lograr el efecto deseado. La calidad del procesado está muy influida la calidad del análisis.

3. **Subsistema de Síntesis:** En este bloque se sintetiza la nueva componente sinusoidal y se combina con la componente residual resultante del bloque de análisis. Esta mezcla da lugar al sonido final.

1.3. Organización de la memoria

En este subapartado se explicará cómo está organizado este documento y se resumirán brevemente las ideas más importantes de cada apartado.

La memoria está dividida en 7 capítulos:

1. **Introducción:** Se presenta y contextualiza el proyecto, se marcan claramente los objetivos y se expone un diagrama de bloques de alto nivel del sistema.

2. **Percepción de la disonancia:** Se exponen las teorías más importante que explican la percepción de la disonancia sonora. Este estudio ha sido necesario para definir adecuadamente el problema a resolver.

3. **Subsistema de Análisis:** En este apartado se explica con detalle el procedimiento de análisis utilizado para obtener una representación útil de la señal original.

4. **Subsistemas de Procesado:** En este apartado se expone de qué forma se ha utilizado y modificado la información resultante del análisis para reducir realmente las disonancias de la señal original.

5. **Subsistema de Síntesis:** Este apartado está dedicado a explicar cómo se realiza la síntesis de la señal y qué fundamentos matemáticos existen detrás de este procedimiento.

6. **Resultados** En este apartado se realiza una evaluación del sistema usando el banco de pruebas previamente generado. Se valorarán los resultados desde un punto de vista subjetivo (a través de encuestas) y desde un punto de vista objetivo (utilizando el modelo perceptual implementado en [49]).

7. **Conclusiones:** En este apartado se resumen las conclusiones de interés que se han extraído de la elaboración de este Proyecto.

Capítulo 2

Percepción de la disonancia

En este capítulo se presentan los conceptos de psicoacústica y fisiología del oído en los que ha sido necesario profundizar para llevar a cabo este Proyecto. En el apartado 2.1 se aclararán dos conceptos que pueden dar lugar a confusión: *desafinación* y *disonancia*, tratándose el primero de un fenómeno más relacionado con lo musical, y el segundo con lo perceptual. Posteriormente, en el apartado 2.2 se explicará qué efecto tiene la duración de tonos estables sobre la percepción de los mismos, y cómo estos resultados han sido aplicados al desarrollo del Sistema. Por último, el apartado 2.3 es el más extenso, donde se exponen las teorías más significativas que intentan explicar y modelar el fenómeno perceptual de la disonancia.

2.1. Desafinación y disonancia

El objetivo principal de este Proyecto es implementar un sistema capaz de atenuados los efectos indeseados de la desafinación. Pero la definición de "desafinación" y su efecto perceptual están lejos de ser evidentes, y más aún si se desea determinar qué se considera como efecto "indeseado". Por ello resulta imprescindible reflexionar sobre este concepto y aclarar los términos que se utilizarán a lo largo del Proyecto.

2.1.1. Desafinación

La desafinación se produce cuando la altura de los sonidos se desvía ligeramente de la afinación ideal [2]. Este concepto tiene gran dependencia del tipo de música y del instrumento al que se haga referencia. No obstante, existen numerosas convenciones comunes de la música occidental, como es el uso extendido de la escala temperada[1]. Este tipo de consideraciones, aplicables a distintos estilos y contextos, son las que han sido tenidos en cuenta para el diseño. Además, para cierto tipo de armonías, la afinación justa[2] está considerada la afinación óptima, algo que también ha sido contemplado en este Proyecto.

[1]La escala temperada se basa en la división de la octava en 12 intervalos iguales, llamados semitonos.

[2]La afinación justa se basa en relaciones de frecuencias numéricamente simples, imitando de alguna forma la serie armónica natural.

2.1.2. Disonancia

La disonancia es un concepto que puede ser interpretado de forma diferente en función del contexto. La *disonancia musical* se define como el intervalo que, según las reglas de la armonía clásica, resulta "desagrable" al oído [39]. Por otro lado, la *disonancia sensorial* se define en términos puramente perceptuales como la "aspereza" de un sonido dado, y se aplica a sonidos tanto musicales como no musicales [40]. Sin embargo, ninguna de estas definiciones explica totalmente el criterio por el que un oyente cualquiera va a considerar algo como "disonante" o "consonante" [16]. Existen numerosas consideraciones de tipo cognitivo (musicales, culturales, etc.) que han de tenerse en cuenta para que el Sistema realmente sea útil desde el criterio de un usuario estándar. En el apartado 2.3 se comentan las teorías sobre la percepción de la disonancia más interesantes.

En este Proyecto se ha partido de la hipótesis de que la disonancia, como concepto global, es una combinación de disonancia musical, disonancia sensorial, más una serie de consideraciones subjetivas dependientes del estilo musical y de la experiencia del oyente. De esta forma, la evaluación del sistema se ha realizado con material musical de uso común (acordes típicos, sonoridades frecuentes en música, etc). En la evaluación, se cuantifica la disonancia sensorial según el modelo perceptual definido en [49], y también se estudia la valoración subjetiva media de una serie de potenciales usuarios del Sistema. Para esta valoración subjetiva, los oyentes han tenido que valorar el sonido del 1 al 10 con cada uno de los siguientes adjetivos: disonante, desafinado, feo, antimusical, tenso, extraño, desagradable, antinatural y aburrido. Se ha considerado que el oyente percibe algo como disonante si la puntuación en la mayoría de estas características negativas es alta en término medio.

En conclusión, la disonancia puede considerarse como un posible efecto perceptual de la desafinación. Como ya se ha visto, la disonancia es difícil de cuantificar, ya que existen numerosos aspectos subjetivos implicados. La desafinación es más fácil de analizar si se conoce la escala a la que deberían estar ajustados una serie de sonidos. No obstante, en ocasiones esto puede llegar a ser complejo si las intenciones musicales de un sonido dado son ambiguas.

2.2. Efectos perceptuales de la duración

Algunos autores han investigado qué influencia tiene la duración de un sonido sobre la percepción de su tono o altura. La publicación más interesante que se ha encontrado al respecto es [33], en la cual Moore determina que es necesario una duración de al menos 200ms en el sonido para que la percepción de su altura sea precisa.

En este Proyecto se plantea la hipótesis de que la consonancia de dos sonidos no es posible si no existe una percepción clara de la altura de los mismos. Por tanto se considera irrelevante trabajar con fragmentos que no sean estables durante al menos 200ms. Esto es tenido en cuenta para evitar modificaciones sobre contenido espectral que no contribuye a la sensación de disonancia.

No se ha encontrado ninguna publicación que relacione la duración de un sonido con la sensación de disonancia que produce. No obstante los resultados por ensayo-error han determinado que 200ms es un valor razonable, apoyado en parte además por el experimento anteriormente mencionado.

2.3. Teorías sobre la percepción de la disonancia

Distintos autores han encontrado diferentes teorías para el fenómeno perceptual de la disonancia. Éstas se pueden considerar complementarias, ya que abordan el mismo fenómeno desde puntos de vistas diferentes. En [6], Cazden realiza un análisis de las teorías más importantes. Se podría establecer la siguiente clasificación en base a la bibliografía existente:

- **Teorías acústicas:** Intentan calificar un sonido de disonante o consonante en función de sus propiedades acústicas. Por ejemplo, la relación entre las frecuencias que lo forman.
- **Teorías psicofísicas:** Estas teorías tienen en cuenta los aspectos psicofisiológicos del sistema auditivo, como por ejemplo el comportamiento de la cóclea ante cierto tipo de sonidos.
- **Teorías cognitivas:** Estas teorías se basan en aspectos cognitivos de más alto nivel. Por ejemplo, el aprendizaje musical previo y su influencia en la percepción de la disonancia.
- **Teorías culturales:** Estas teorías defienden que la percepción de disonancia o consonancia está basada en aspectos sociales y normas aprendidas inconscientemente por el oyente.

Realmente, el fenómeno se adecúa a teorías diferentes en función del experimento. Por ello no existe un modelo perceptual consensuado para la percepción de la disonancia. En este apartado se detallan las siguientes teorías:

- *Teoría de las relaciones numéricas sencillas:* En la clasificación anterior, estaría dentro de las teoría acústicas, ya que considera que la consonancia o disonancia de un sonido depende de las características acústicas del mismo. En concreto, esta teoría estudia las relaciones númericas entre las frecuencias de sus parciales.

- *Teoría de los batidos:* También es una teoría acústica, pero en este caso la disonancia es analizada a partir de los batidos que aparecen entre parciales. Las conclusiones son muy parecidas a la teoría anterior.

- *Teoría de la disonancia tonotópica o sensorial:* Esta teoría tiene en cuenta las características fisiológicas del oído humano, así que estaría englobada dentro de las teorías psicofísicas. En ella se estudia el comportamiento mecánico del oído interno y se relaciona con el criterio subjetivo de un oyente medio.

- *Teoría de la disonancia por percepción de tono ambigua:* Esta teoría tiene gran cantidad de consideraciones psicoacústicas, aunque también considera aspectos cognitivos de más alto nivel. En ella se defiende que la disonancia aparece cuando el oyente es incapaz de distinguir las entidades sonoras de forma independiente.

- *Teoría de la disonancia por categorización de intervalos:* Esta teoría es de tipo cognitivo/cultural, y defiende que la exposición activa o pasiva a una serie de sonoridades a lo largo de la vida del oyente determinan en gran medida su juicio sobre la consonancia de un sonido dado.

Al final del apartado se exponen algunas conclusiones y se explica por qué estas teorías en concreto han sido consideradas interesantes para el desarrollo de este Proyecto.

2.3.1. Teoría de las relaciones numéricas sencillas

Esta teoría defiende que dos o más sonidos son consonantes entre sí cuando la relación numérica entre sus frecuencias es sencilla. Cuanto más sencilla es dicha relación, más consonante se considera dicho conjunto de sonidos. Esta teoría es consistente con las escalas occidentales comúnmente usadas a lo largo de la historia, así como con las reglas de la armonía clásica.

Aplicando rigurosamente el principio de simplicidad de relaciones armónicas, se deduce la escala musical más perfectamente consonante desde el punto de vista matemático. Esta escala es conocida como la escala de Aristógenes (350 a.C.), escala de afinación justa o escala de Zarlino (teórico de la música del siglo XVI). Consiste en elegir los sonidos buscando que las relaciones de frecuencias con el primer sonido de la escala (o Tónica), sean lo más sencillas posibles. Algunas de estas relaciones se pueden observar en la tabla 2.3.1:

Intervalo	**Relación entre frecuencias**	**Nombre del intervalo**
DO3-RE3	9/8	Segunda mayor
DO3-MI3	5/4	Tercera mayor
DO3-FA3	4/3	Cuarta justa
DO3-SOL3	3/2	Quinta justa
DO3-LA3	5/3	Sexta mayor
DO3-SI3	15/8	Séptima mayor
DO3-DO4	2/1	Octava

Tabla 2.1: Intervalos en la escala de afinación justa

El problema de esta escala es que sólo proporciona la afinación perfecta con respecto a la tónica. Si se desea variar la tonalidad, las frecuencias asociadas a cada nota varían. Actualmente se ha estandarizado el uso de la escala temperada, que es una versión aproximada de la escala de Zarlino que permite ser traslada a instrumentos de teclado. Este sistema se construye dividiendo la octava en 12 semitonos de igual distancia. La razón entre frecuencias para un intervalo de semitono por tanto resulta ser $2^{\frac{1}{12}}$. En esta escala, el único intervalo perfectamente afinado es la octava. Por ejemplo, en el caso de la quinta, $2^{\frac{7}{12}} = 1{,}498$ no llega a ser exactamente $\frac{3}{2}$.

Se podría decir que la escala comúnmente usada hoy día esta matemáticamente "desafinada", y de hecho en música clásica es habitual ajustar ciertos intervalos a la afinación perfecta si el instrumento lo permite para conseguir mayor consonancia. No obstante, debido al uso tan frecuencia de la afinación temperada, actualmente puede ser percibido como más consonante un acorde temperado dependiendo del contexto [16]. Esto evidencia la dificultad del problema que se pretende resolver, ya que el concepto de "afinado" o "desafinado" es muy subjetivo y dependiente de las circunstancias. En cualquier caso, se puede considerar que las relaciones sencillas entre frecuencias es la base de la afinación occidental, por lo que siempre existirá esa componente objetiva en el fenómeno de la consonancia.

2.3.2. Teoría de los batidos

Esta teoría fue presentada por el físico Helmholzt en 1877 [18], y aseguraba que la causa de la disonancia entre dos sonidos es la presencia de batidos entre sus armónicos adyacentes. Los batidos suceden cuando dos sinusoides tienen frecuencias cercanas, y como resultado se percibe una modulación en amplitud de baja frecuencia. Helmholzt llegó incluso a proponer que la máxima disonancia se percibe cuando los batidos tienen una modulación en amplitud de alrededor de 35 ciclos por segundo [6]. No obstante, esto ha sido ampliamente rebatido por investigadores posteriores, ya que esta teoría no explica el resultado de experimentos sobre la disonancia posteriormente realizados.

En cualquier caso, esta teoría es consistente con la teoría de las razones sencillas de frecuencia, ya que el mínimo número de batidos sucede cuando las relaciones efectivamente son simples. En la figura 2.1 se muestra un ejemplo gráfico que relaciona ambas teorías.

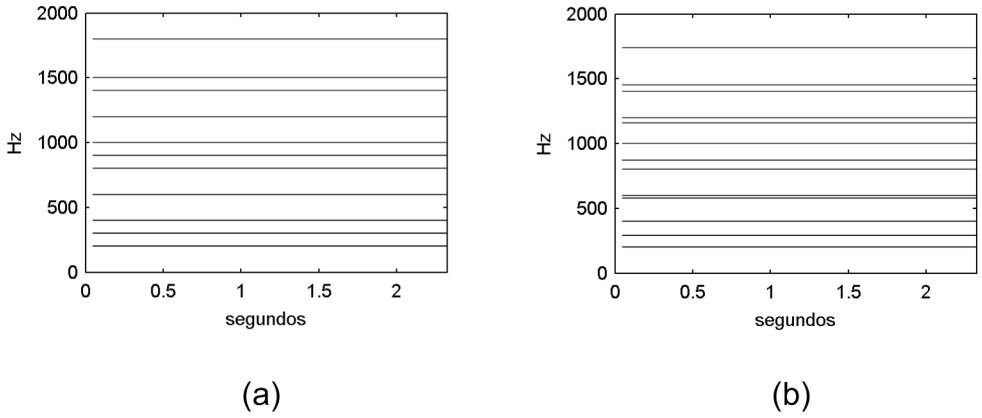


Figura 2.1: (a) Parciales de dos tonos complejos simultáneos de 200Hz y 300Hz (afinación exacta) (b) Parciales de dos tonos complejos de 200Hz y 290Hz (desafinación)

Como se observa, cuando las f_0 de las notas que forman el acorde siguen relaciones numéricas sencillas, existen multitud de armónicos coincidentes. Por ejemplo, en este caso el tercer armónico de 200Hz coincide con el segundo armónico de 300Hz (ambos están en 600Hz). Sucede lo mismo con muchos otros armónicos.

Sin embargo, cuando se trata de sonidos de baja frecuencia la teoría de Helmholzt no se cumple del todo, y es por ello que otros investigadores siguieron indagando en el fenómeno. La siguiente teoría arroja cierta luz sobre la base del fenómeno.

2.3.3. Teoría de la disonancia tonotópica o sensorial

La disonancia tonotópica es la que producen dos tonos puros cuando están próximos en frecuencia, y su cuantificación se basa exclusivamente en la sensación agradable o desagradable que producen. De esta forma, los juicios de tipo musical o basados en conocimiento previo no son contemplados por esta teoría.

La publicación más interesante al respecto es [40] (Plomp & Leveltz, 1965). En ella se relaciona la fisiología del oído y la percepción de la disonancia, argumentando que dos tonos son percibidos como disonantes cuando caen dentro de una misma "banda crítica". El término de banda crítica fue introducido por Harvey Fletcher en los años 40, y está relacionado con las propiedades mecánicas de la membrana basilar dentro de la cóclea. En la figura 2.2 se muestra un esquema de la cóclea para visualizar mejor el funcionamiento de la misma.

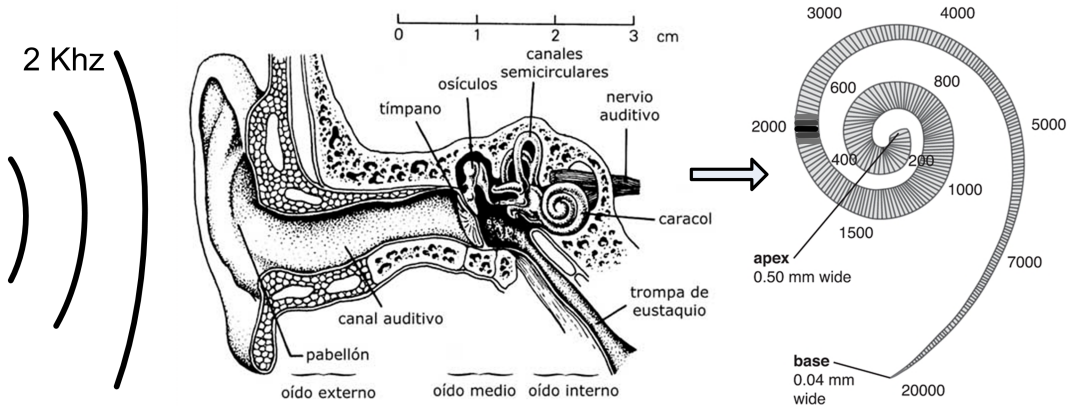


Figura 2.2: Esquema del sistema auditivo y respuesta de la cóclea a un tono de 2KHz. Como se observa, el tono excita una zona amplia de la cóclea, correspondiendo al ancho de banda crítico para dicha frecuencia. Imágenes extraída de [47, 30]

Cuando una frecuencia excita un punto de la membrana basilar, ese punto y sus alrededores entran en vibración. La banda crítica se puede considerar el rango de frecuencias que abarca la zona excitada de la membrana basilar debido a un tono puro. Si dos tonos caen dentro de la misma banda crítica, la interacción mecánica entre ellos va a impedir la correcta percepción de ambos, y es entonces cuando se percibe el fenómeno denominado disonancia sensorial. La curva de la figura 2.3 muestra el tamaño de la banda crítica para cada frecuencia. Se puede observar cómo la banda crítica es proporcionalmente mayor cuanto más baja es la frecuencia.

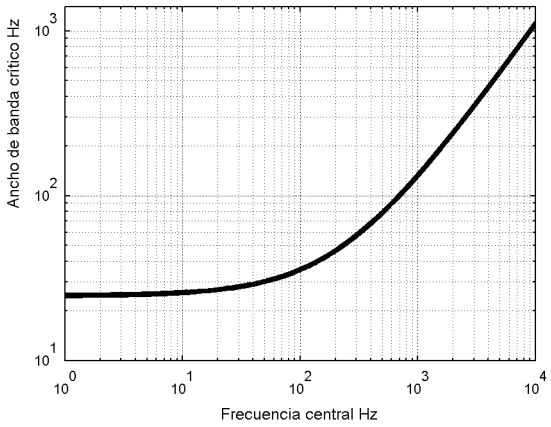


Figura 2.3: Ancho de banda crítico con respecto a la frecuencia central. La fórmula utilizada es la dada en [45], ERB = 0.108F + 24.7. Se trata del ancho del filtro rectangular equivalente

Plomp, en [40], estableció que la máxima disonancia percibida sucede cuando dos tonos están separados un 25 % de la banda crítica, tal y como se muestra en la figura 2.4. Debido a las variaciones de tamaño de la banda crítica a lo largo de la cóclea, la separación de frecuencias que produce la máxima disonancia también es variable.

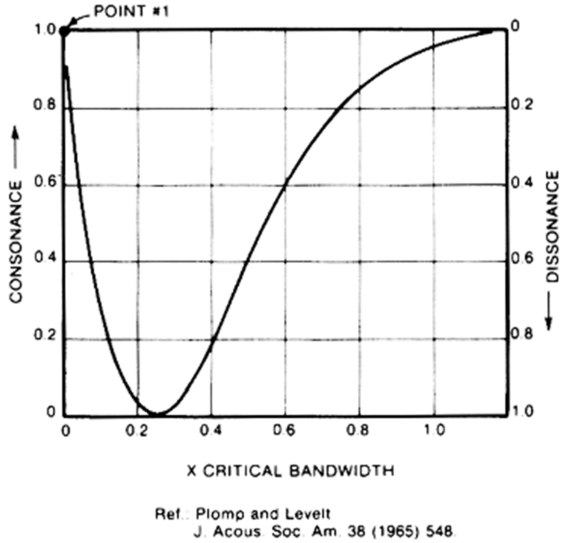


Figura 2.4: Curva que relaciona la consonancia percibida y la separación entre dos tonos en función de la banda crítica (extraído de Plomp&Leveltz 1965)

Estos resultados amplían la teoría de los batidos de Helmholzt, dando una explicación más genérica basada en factores objetivos, como es la fisiología del sistema auditivo. En su publicación, Plomp además analiza la interacción de tonos complejos, tal y como se muestra en la figura 2.5.

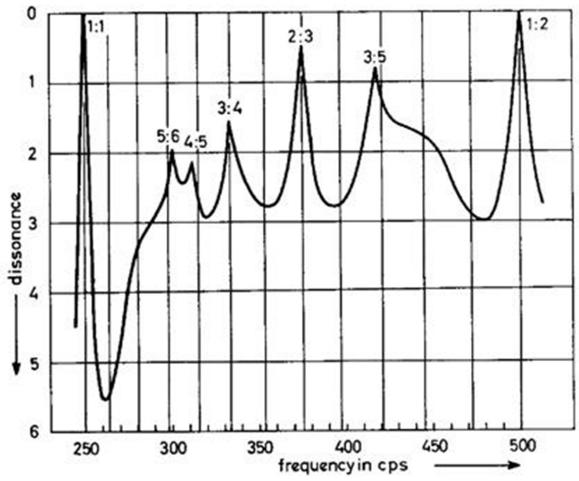


Figura 2.5: Consonancia percibida en relación con la separación entre dos tonos complejos

Como se observa, los resultados muestran coherencia con la teoría de las razones sencillas de frecuencia. Realmente, esta teoría y las anteriores son diferentes puntos de vista del mismo fenómeno, y es lógico por tanto llegar a conclusiones similares.

Existen otra serie de interesantes publicaciones [7, 20, 21, 23, 22, 24, 36, 38, 41, 48, 49, 51] que abordan el mismo tema de la disonancia tonotópica, intentando en algunos casos formular un modelo perceptual que permita cuantificar la disonancia de un sonido de forma algorítmica. Las conclusiones en estos casos son similares a [40], con modificaciones menores para adecuarlas a nuevas observaciones. En el caso de [49], el modelo perceptual ha sido implementado en una aplicación online, y ha sido el método objetivo usado en este Proyecto para el análisis de los resultados.

2.3.4. Teoría de la disonancia por percepción de tono ambigua

Varios autores (como bien se explica en [6]) han observado que un conjunto de sonidos puede ser percibido como disonante si sus componentes no se funden para formar una entidad única. Esto sucede cuando existe una inharmonicidad antinatural en el timbre de un sonido, que impide resolver la altura de las notas que suenan. En estos casos, se percibe un sonido "acampanado", ambiguo y que distorsiona la percepción del oyente.

Esta teoría complementa las anteriores, ya que estudia un tipo de disonancia que puede aparecer incluso en ausencia de batidos entre parciales. Es un tipo de disonancia más difícil de solucionar, y ha sido uno de los grandes problemas encontrados en el desarrollo de esta herramienta.

2.3.5. Teoría de la disonancia por categorización de intervalos

Esta teoría ha sido defendida en [6] y [16]. Arguye que la percepción de la disonancia tiene una gran componente cultural, basada en el aprendizaje previo y variante según la tradición musical. La idea base de esta teoría es que los intervalos musicales son interiorizados por exposición, estableciendo así una identificación categórica de los mismos. Si los sonidos se salen de esas categorías conocidas, serán percibidos como "raros", y en último extremo como disonantes.

2.3.6. Conclusión y aplicación de las teorías anteriores a este Proyecto

Una vez que se conocen las teorías que subyacen bajo la percepción de la disonancia, es necesario plantearse de qué forma este conocimiento puede ser útil en este Proyecto. Se han identificado fundamentalmente dos tipos de disonancia en el caso de sonidos desafinados:

- Disonancia tonotópica debido a la cercanía de los parciales de ambos sonidos
- Disonancia debido a que el intervalo resultante no llega a estar totalmente dentro de una categoría conocida.

Se deduce por tanto que para solventar un problema de afinación es necesario desplazar los parciales a las posiciones que deberían tener en el caso de un intervalo afinado. De esta forma, el intervalo entra dentro de una categoría conocida, y se reduce la disonancia tonotópica gracias a la coincidencia de parciales de ambos sonidos.

Capítulo 3

Subsistema de Análisis

En este apartado se detalla el diseño y la implementación del Subsistema de Análisis (ver diagrama del apartado 1.2).

En el apartado 3.1 se presenta un diagrama funcional del subsistema. En él se refleja el esquema clásico de un modelo sinusoidal-residual, basado en [43]. Los fundamentos de este modelo han sido detallados en el apartado 3.2. En los apartados posteriores de este capítulo se explica el funcionamiento de los diferentes bloques funcionales.

En el apartado 3.3 se explica cómo se ha implementado el bloque que realiza la Transformada Corta de Fourier (*Short-term Fourier Transform*, o STFT). En él se detallan las consideraciones tenidas en cuenta para que el espectrograma resultante se adapte a las necesidades del Sistema. Seguidamente, en el apartado 3.4, se detalla el procedimiento para estimar de la componente sinusoidal de la señal a partir del espectrograma. En este apartado no se contempla el seguimiento temporal de los parciales, sino que se estima la componente sinusoidal ventana a ventana. A partir de la estimación sinusoidal realizada ventana a ventana, se extrae la componente residual tal y como se explica en el apartado 3.5.

Posteriormente, se estima la continuación temporal de cada sinusoide según el procedimiento explicado en [42]. El resultado es un array de parciales que parametriza la señal de forma muy versátil para aplicar procesados interesantes. Los detalles se pueden encontrar en el apartado 3.6.

Por último, en el apartado 3.7 se explica el procedimiento que se ha utilizado para estimar las frecuencias fundamentales más importanes que componente la señal de entrada.

Observando el diagrama de bloques del apartado 3.1, se observa que falta por detallar el bloque *Síntesis de componente sinusoidal*. Éste ha sido desarrollado en el apartado 5.3 (dentro del Subsistema de Síntesis), ya que este bloque se utiliza tambien en dicho subsistema.

3.1. Diagrama de bloques

Para tener una idea global del funcionamiento de este subsistema, en la figura 3.1 se muestra un esquema del mismo.

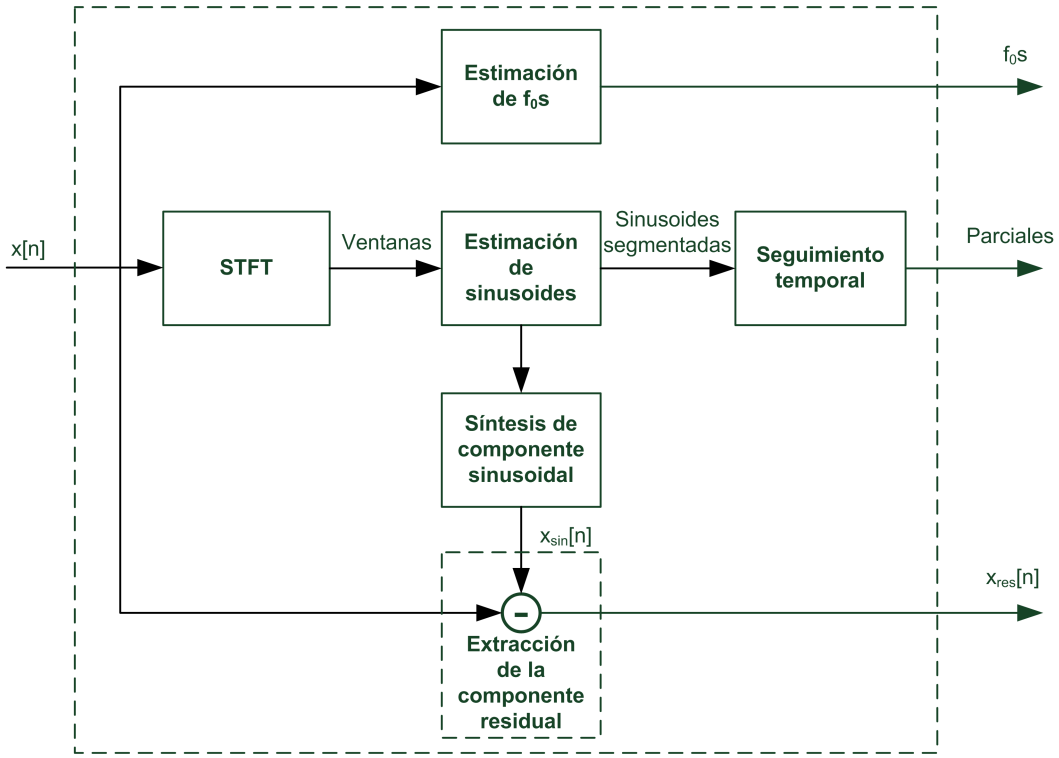


Figura 3.1: Diagrama de bloques del Subsistema de Análisis

Este diagrama de bloques corresponde es el estándar para un sistema basado en modelado sinusoidal [43]. En los siguientes apartados se explicará en qué consiste este modelado, y se detallarán aspectos concretos del funcionamiento del sistema.

3.2. Modelo sinusoidal + residual

Cuando se trabaja con sonidos musicales, es importante tener un buen modelo cuyos parámetros proporcionen una fuente rica de transformaciones útiles. Existen básicamente tres tipos de modelos: modelos físicos de la fuente sonora, modelos basados en el espectro y modelos abstractos (como la síntesis FM).

En este trabajo se ha optado por la segunda categoría. La ventaja de este grupo de técnicas es que existen procedimientos de análisis capaces de extraer los parámetros más relevantes de los sonidos reales. El enfoque utilizado ha sido modelar los sonidos como la suma de sinusoides estables (parciales) y una componente residual.

Para comprender mejor en qué consiste la descomposición sinusoidal + residual, en la Figura 3.2 se muestra la forma de onda y el espectograma de una señal musical, y sus dos componentes tras dicha descomposición.

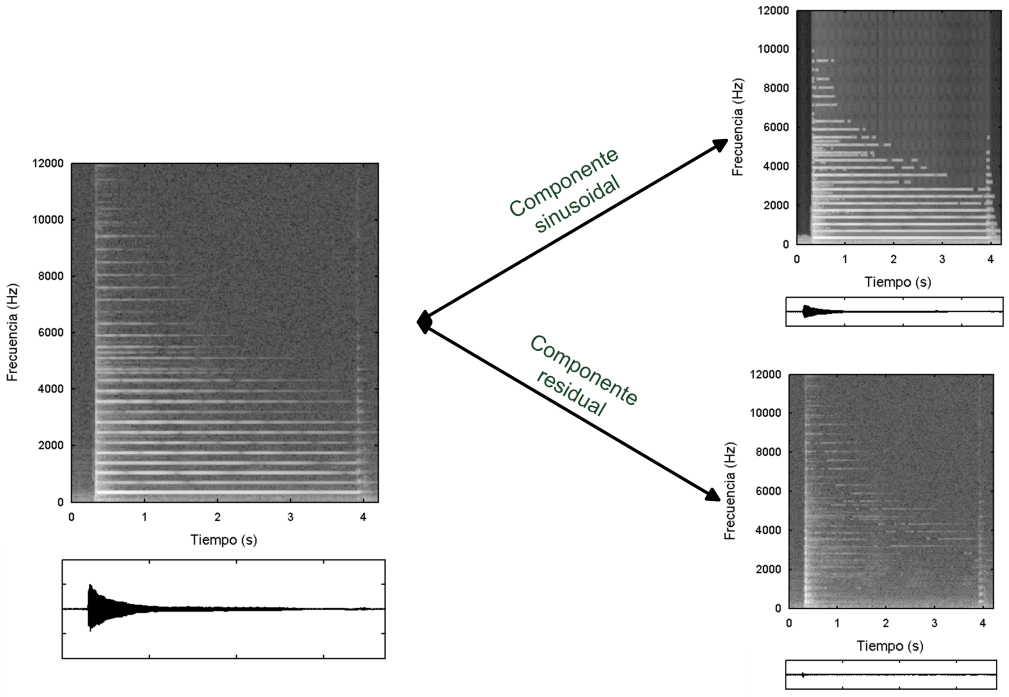


Figura 3.2: Ejemplo gráfico de descomposición sinusoidal + residual de una nota de piano F4

No siempre es posible ajustar un cierto sonido a dicho modelo. Por ejemplo, en el caso de sonidos muy cambiantes, o con mucha reverberación, las componentes sinusoidal y residual suelen estar mezcladas. Sin embargo, en grabaciones limpias de instrumentos musicales o sonidos vocales cada componente suele aparecer bien diferenciada.

3.2.1. Descripción analítica del modelo

El modelo escogido asume que la señal puede expresarse como la suma de una componente sinusoidal y una componente residual. La siguiente expresión resume el modelo para tiempo contínuo:

$$x(t) = x_{sin}(t) + x_{res}(t) = \sum_{r=1}^{R} A_r(t) cos[\theta_r(t)] + x_{res}(t) \tag{3.1}$$

Donde:

$$\theta_r(t) = \int_0^t \omega_r(\tau) + \phi_r \tag{3.2}$$

Siendo:

$$\begin{aligned} A_r(t) &= \text{Amplitud instantánea de la sinusoide } r \\ \omega_r(t) &= \text{Frecuencia instantánea de la sinusoide } r \\ \phi_r &= \text{Fase inicial de la sinusoide } r \\ R &= \text{Número de sinusoides que forman la señal} \end{aligned}$$

Trasladado a tiempo discreto, el modelo quedaría de la siguiente forma:

$$x[n] = x_{sin}[n] + x_{res}[n] = \sum_{r=1}^{R} A_r[n]\ cos(\theta_r[n]) + x_{res}[n] \tag{3.3}$$

$$\theta_r[n] = \sum_{k=0}^{n} \omega_r[k] + \phi_r \tag{3.4}$$

La componente residual $x_{res}[n]$ no se parametriza de ninguna forma, sino que se almacenan sus valores de amplitud en el dominio del tiempo sin alteraciones.

La obtención de los valores de amplitud y frecuencia instantáneos requeriría una resolución temporal y frecuencial perfecta durante la fase de análisis. Sin embargo, esto no es posible utilizando la STFT como representación tiempo-frecuencia. En la práctica, de cada ventana $l \in [1, L]$ se obtendrán una serie de triadas $(\hat{A}_r^l, \hat{\omega}_r^l, \hat{\phi}_r^l)$ correspondientes a cada uno de los parciales $r \in [1, R]$. La síntesis de la componente sinusoidal una determinada ventana l se realiza con la siguiente expresión:

$$x_{sin}^l[m] = \sum_{r=1}^{R^l} \hat{A}_r^l cos(\hat{\omega}_r^l \cdot m + \hat{\phi}_r^l) \qquad m = 1 \dots N_s \tag{3.5}$$

Donde r representa el índice del parcial, l representa el índice de ventana, y N_s representa el tamaño de la ventana de síntesis. Se observa que cada una de estas ventanas

está formada por sinusoides estables, no contemplando variaciones de la amplitud o frecuencia instantáneas. Durante la síntesis, cada una de estas ventanas se sintetiza y se solapa con la anterior, dando lugar a una aproximación de la componente sinusoidal. Para que esta aproximación sea realmente útil, el tamaño de ventana ha de adaptarse al tipo de señal con que se esté trabajando.

3.3. Transformada Corta de Fourier: STFT

La STFT es una potente herramienta de propósito general para procesado de señales de audio. Se trata de un tipo de distribución tiempo-frecuencia especialmente útil. Esta transformada se basa en el cálculo del espectro de pequeñas ventanas temporales a lo largo de toda la señal (figura 3.3). De esta forma se obtiene una amplitud compleja para cada valor de tiempo y frecuencia de la señal.

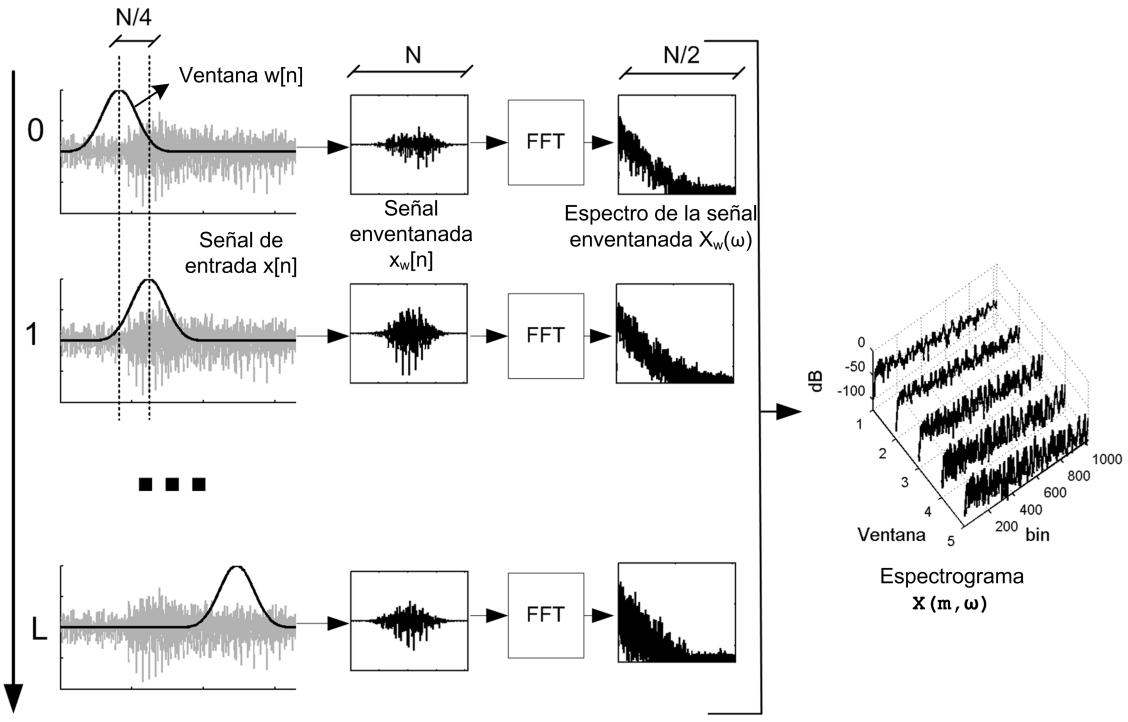


Figura 3.3: Procedimiento para el cálculo de la Transformada Corta de Fourier. En este caso se ha utilizado una ventana de tipo Blackman-Harris 92dB, con un factor de superposición del 75 %.

El tipo de ventana va a influir sobre el resultado de la STFT, como se explica con detalle en el apartado 3.3.3.

3.3.1. Definición de la STFT

En la expresión 3.6 se muestra analíticamente la definición matemática más común de la STFT para señales de longitud infinita (la cual aparece en [44], una excelente referencia).

$$X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]\ w[n-mH]\ e^{-j\omega n} \tag{3.6}$$

Donde

$$\begin{aligned} x(n) &= \text{Señal de entrada} \\ w(n) &= \text{Ventana de tamaño M centrada en el origen} \\ H &= \text{Salto entre ventanas, o } \textit{hop-size}\text{, en muestras} \\ m &= \text{Índice de ventana} \end{aligned}$$

En el caso de este Proyecto, se ha deseado que la posición absoluta de la ventana no influya en el resultado de la STFT. Esto se consigue normalizando el término $e^{-j\omega n}$ para que esté siempre centrado alrededor de la ventana:

$$\begin{aligned} X(m,\omega) &= \sum_{n=-\infty}^{\infty} x[n]\ w[n-mH]\ e^{-j\omega(n-mH)} \\ &= \sum_{n=-\infty}^{\infty} x[n+mH]\ w[n]\ e^{-j\omega n} \end{aligned} \tag{3.7}$$

Sin embargo, la implementación computacional de la STFT no permite hacer cálculos con infinitas muestras. Una posible definición de la STFT para señales finitas sería:

$$\widetilde{X}[m,k] = \sum_{n=-N/2}^{N/2-1} \widetilde{x}[n+mR]\ \widetilde{w}[n]\ e^{-j\omega_k n} \tag{3.8}$$

Donde la tilde $\widetilde{x}[n]$ y $\widetilde{w}[n]$ representa la extensión periódica de periodo N de la señal truncada:

$$\widetilde{x}[n] = x[((n))_N] \tag{3.9}$$

$$\widetilde{w}[n] = w[((n))_N] \tag{3.10}$$

$$\omega = \omega_k = \frac{2\pi k}{N} \tag{3.11}$$

3.3.2. Cálculo de la STFT sin distorsión de fase

En Matlab, al generar una ventana $w[n]$ de longitud M (definida en $[1, M]$), el punto central de la ventana suele estar en $M/2$ (donde se sitúa el eje de simetría). Sin embargo, esto produce distorsión sobre el espectro de fases de la STFT. En este subapartado se explicará de qué forma se puede evitar esta distorsión, analizando el efecto de la simetría de $w[n]$ sobre la fase de la DFT.

Para simplificar el desarrollo matemático, se realizará primero un análisis en tiempo discreto con señales de duración infinita (DTFT, *Discrete-Time Fourier Transform*). Posteriormente, el mismo análisis se trasladará a señales discretas de duración finita (DFT, *Discrete Fourier Transform*).

DTFT: Señales de duración infinita

La DTFT se define como:

$$X(\omega) = \sum_{n=-\infty}^{\infty} x[n]\ e^{-j\omega n} \tag{3.12}$$

Supóngase $x[n]$ una sinusoide enventanada con $w[n]$, una ventana genérica de duración impar M:

$$x_w[n] = x[n] \cdot w[n] = \begin{cases} 0 & si \quad n < 0 \\ w[n] \cdot A \cdot cos(\omega_0 n + \phi) & si \quad 0 \geq n \geq M-1 \\ 0 & si \quad n \geq M \end{cases} \tag{3.13}$$

La DTFT de esta señal es:

$$\begin{aligned} X_w(\omega) &= X(\omega) * W(\omega) \\ &= \frac{A}{2}\Big(e^{j\phi}\delta(\omega-\omega_0) + e^{-j\phi}\delta(\omega+\omega_0)\Big) * |W(\omega)|e^{-j\omega\frac{(M-1)}{2}} \\ &= \frac{A}{2}\Big(e^{-j[(\omega-\omega_0)\frac{(M-1)}{2}-\phi]}|W(\omega-\omega_0)| + e^{-j[(\omega+\omega_0)\frac{(M-1)}{2}+\phi]}|W(\omega+\omega_0)|\Big) \end{aligned} \tag{3.14}$$

Es habitual extraer la fase de una sinusoide a partir del semieje positivo del espectro de fases. En este caso:

$$\angle X_w(\omega > 0) = (-\omega + \omega_0)\frac{(M-1)}{2} + \phi \tag{3.15}$$

Como se observa, el espectro de fases no es constante con ω, sino que crece de forma lineal. El objetivo del análisis es obtener un único valor de fase, por lo tanto esta variabilidad de la fase conlleva varios problemas:

- Interpretar el espectro de fase es más complicado, puesto que los lóbulos tienen la fase en pendiente y no constante.
- En el proceso de síntesis, es necesario computar las pendientes combinando ϕ, M y ω_0.

Para simplificar el análisis, se ha realizado un desplazamiento de $x_w[n]$ que da lugar a un espectro de fases constante. Supóngase la señal $x_w[n + (M-1)/2]$, es decir, centrada en el origen, con $(M-1)/2$ valores en índices positivos y $(M-1)/2$ valores en índices negativos. Aplicando la propiedad de desplazamiento:

$$x_w[n + (M-1)/2] \stackrel{DTFT}{\Longrightarrow} X_w(\omega) \cdot e^{j\omega(M-1)/2} \tag{3.16}$$

Resultando:

$$\angle X_w(\omega > 0) = \omega_0 \frac{(M-1)}{2} + \phi \tag{3.17}$$

En este caso se observa que la fase es independiente de ω. A partir del análisis, el único valor que se extrae es éste, y contiene toda la información necesaria para proceder posteriormente a la síntesis. A este nuevo valor de fase se le ha llamado ϕ':

$$\phi' = \phi + \omega_0 \frac{(M-1)}{2} \tag{3.18}$$

Por tanto, la terna de análisis para una sinusoide r en una ventana l es: $(\hat{A}_r^l, \hat{\omega}_r^l, \hat{\phi'}_r^l)$.

Una vez comprendido este procedimiento en señales de longitud infinita, se puede aplicar el mismo razonamiento a señales de longitud finita.

DFT: Señales de duración finita

La definición de DFT es la siguiente:

$$\widetilde{X}[k] \stackrel{def}{=} \sum_{n=0}^{N-1} \widetilde{x}[n]\ e^{-j\frac{2\pi k}{N}n} \tag{3.19}$$

Donde:

$$\widetilde{x}[n] = x[((n))_N] \tag{3.20}$$

Los resultados obtenidos con la DTFT pueden ser aplicados a señales de duración finita a través de un elaborado desarrollo matemático. Sin embargo, se ha considerado más didáctico y práctico mostrar ejemplos reales antes y después del desplazamiento. En la Figura 3.4 se muestra el resultado real de la DFT de un tono puro de duración finita antes de desplazarla, mientras que en la figura 3.5 se muestra el resultado tras el desplazamiento:

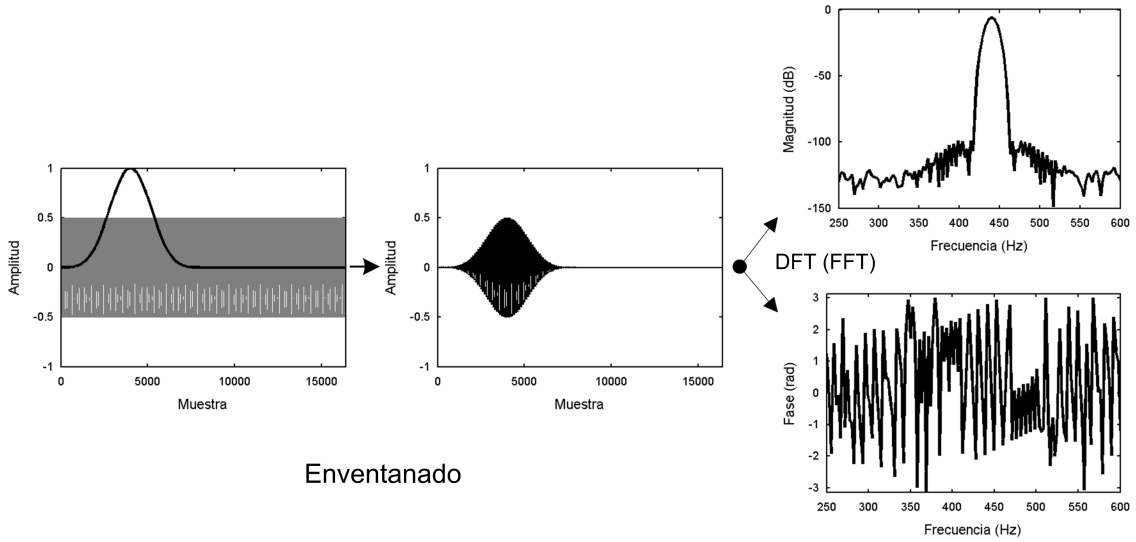


Figura 3.4: DFT en módulo y fase de un tono puro enventanado. Tamaño de ventana (M): 8001 muestras, Tamaño de FFT (N): 16384, Tipo de ventana: Blackman-Harris 92dB.

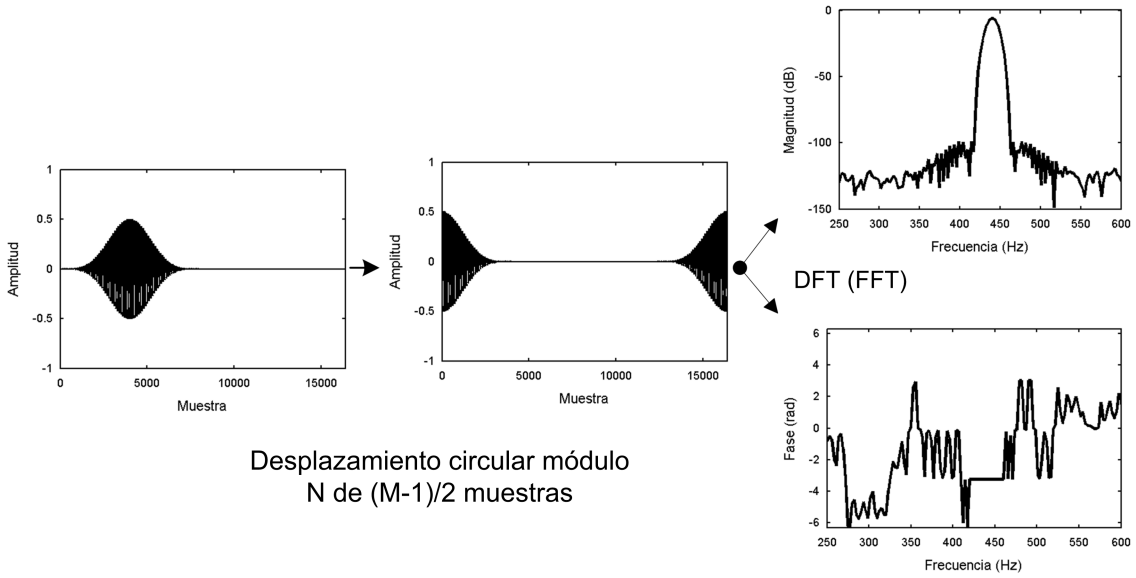


Figura 3.5: Desplazamiento circular de (8001-1)/2 muestras para evitar la distorsión de fase

En la Figura 3.4, se observa que la fase está en pendiente, como bien se deducía de las expresiones en el caso de la DTFT. Esto impide una interpretación y parame-

trización directa del espectro de fase, lo cual resulta incómodo para el análisis. Para conseguir valores constantes en la fase es necesario centrar en el origen la señal enventanada. En el caso de la DTFT, el desplazamiento necesario era $x_w[n + (M-1)/2]$. En el caso de la DFT, este desplazamiento es en realidad:

$$\widetilde{x_w}[n + (M-1)/2] = \widetilde{x_w}[((n + (M-1)/2))_N] \tag{3.21}$$

Esto significa que la primera mitad de la ventana se sitúa al final, la segunda mitad al principio y en medio se queda un relleno de $N - M$ ceros. Al igual que sucedía en el caso de la DTFT, este desplazamiento produce resultados de la DFT con fase constante a lo largo de los lóbulos principales (figura 3.5).

Como se observa, el espectro de fase es ahora constante dentro del lóbulo principal, pudiéndose parametrizar en el análisis con un único valor. No obstante, este valor de fase corresponde a la señal desplazada, por ello es importante tenerlo en cuenta para revertir la operación posteriormente en el proceso de síntesis.

3.3.3. Elección del tipo de ventana

La STFT se basa en el cálculo de la FFT de ventanas temporales a lo largo de la señal. Esta ventana temporal se obtiene multiplicando en el dominio del tiempo la señal original $x[n]$ por una señal ventana $w[n]$:

$$x[n] \cdot w[n] \overset{DTFT}{\Longrightarrow} X(\omega) * W(\omega) \tag{3.22}$$

El tipo de ventana que se escoja va a influir notablemente sobre varios aspectos: resolución frecuencial, atenuación de lóbulos secundarios, factor de superposición durante la síntesis, etc. Para tomar decisiones acerca del tipo de ventana a utilizar, se han consultado las referencias [37, 42, 17, 35]. Especialmente relevante ha sido la referencia [44], que aporta un análisis brillante del problema. En este apartado se resumirán brevemente los aspectos más intesantes encontrados en dicha referencia.

En la figura 3.6 se muestran tres resultados diferentes tras la FFT de un tono puro de 440Hz, correspondientes a una ventana rectangular, una ventana de Hamming y una ventana Blackman-Harris de 92dB.

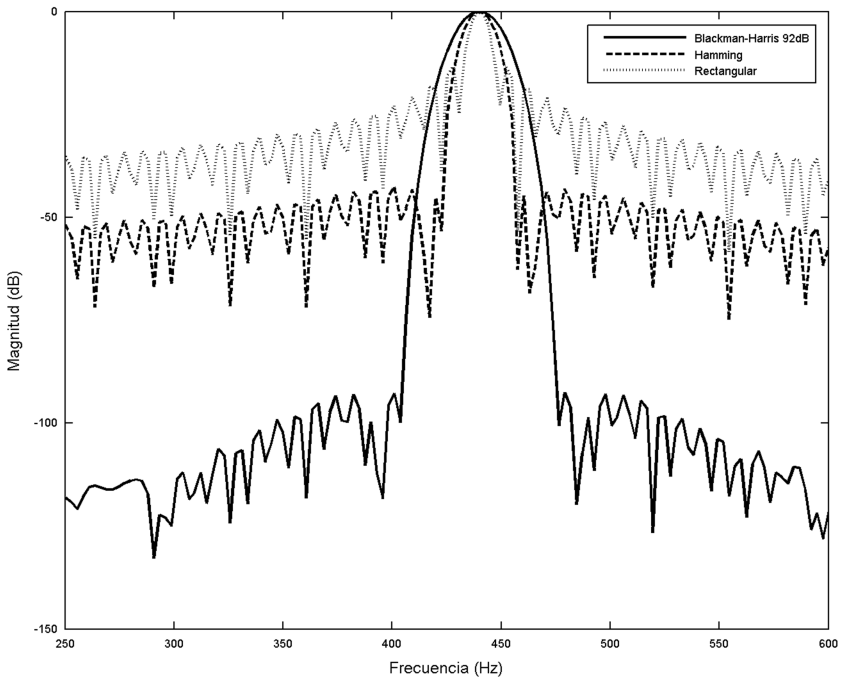


Figura 3.6: Resultado de la transformación DFT usando tres tipos de ventanas diferentes. Los dos parámetros más importantes son la anchura del lóbulo principal y la atenuación de los lóbulos secundarios, existiendo un compromiso entre ellos. La señal analizada es un tono puro de 440Hz de amplitud unidad. Tamaño de ventana (M): 5001 muestras. Tamaño de FFT (N): 16384.

En [17] se puede encontrar un análisis detallado de cada tipo de ventana que puede servir de referencia para tomar decisiones al respecto. En la siguiente tabla se resumen algunas características de cada tipo de ventana.

Tipo de ventana	**Atenuación de lóbulos secundarios (dB)**	**Ancho del lóbulo principal (K) (bins)**
Rectangular	13	2
Hamming	43	4
Blackman-Harris 92dB	92	8

La anchura del lóbulo principal puede estar dada con diferentes definiciones. La que se ha aplicado en este caso es la distancia entre los dos mínimos a los lados del lóbulo (situados a magnitud $-\infty$ dB, o la equivalente amplitud compleja nula en unidades lineales). Otras definiciones pueden ser encontradas en [17], como la definición de ancho de lóbulo a -6 dB.

Resolución frecuencial

Teniendo en cuenta la anchura del lóbulo principal, se puede conocer la separación mínima necesaria entre dos tonos puros para que ambos puedan ser identificados separadamente. El análisis necesario para comprender con detalle esto no es obvio, aunque puede ser encontrado en [44]. La conclusión es que para que ello sea posible, la separación ha de ser mayor que la anchura del lóbulo principal K.

Dado un tipo de ventana con anchura de lóbulo pricipal K, una separación de frecuencias mínima Δf, y una frecuencia de muestreo f_s, se puede calcular el mínimo tamaño de ventana M necesario de la siguiente forma:

$$M > K\frac{f_s}{\Delta f} \tag{3.23}$$

Usando un tamaño M tal y como se ha expresado, podemos asegurar que dos tonos separados por una distancia Δf van a identificarse correctamente, ya que estos estarán correctamente situados y aparecerá un mínimo detectable entre ambos lóbulos.

En este Proyecto, la ventana escogida para todos los casos es **Blackman-Harris de 92dB**. Esta ventana tiene un lóbulo principal muy ancho ($K = 8$), por lo tanto una baja resolución espectral. Sin embargo, tiene una gran ventaja sobre las demás, y es que un tono puro se traduce prácticamente en un único lóbulo principal, puesto que los lóbulos secundarios están muy atenuados.

A la hora de detectar sinusoides, se observó que los lóbulos secundarios daban lugar a picos ficticios que confundían al sistema. La atenuación de los lóbulos secundarios primó sobre la resolución frecuencial. Además, el problema de su baja resolución es compensado gracias a la posibilidad de usar ventanas relativamente largas, ya que las señales a manipular suelen ser estables en el tiempo. De hecho, se han obtenidos buenos resultados con ventanas de 4096 muestras y de 8192 muestras (90ms y 180ms respectivamente a 44100Hz, permitiendo resoluciones de $\Delta f \approx 40Hz$).

Por otro lado, esta ventana tiene la desventaja de necesitar un factor de superposición del 25 %. Esto aumenta el número de ventanas y por tanto el coste computacional del sistema final. Sin embargo, dado que las ventanas suelen ser bastante largas, este factor de superposición pequeño permite que la síntesis se realice con más suavidad y naturalidad, y por ello esto no se ha considerado negativo.

3.4. Estimación de sinusoides

Una vez que se ha calculado el espectrograma de la señal utilizando parámetros adecuados (tamaño de ventana M, tipo de ventana, tamaño de FFT N, tamaño de salto H, etc.), el siguiente paso en la fase de análisis es la estimación sinusoidal. En esta fase se estimarán las sinusoides predominantes, obteniendo así una parametrización útil del contenido armónico de la señal.

Esta estimación se realiza a través de la detección de máximos locales en el espectro de magnitudes, tal y como se explica en el subapartado 3.4.1. Estos máximos locales, idealmente, deberían representar el vértice de cada uno de los lóbulos principales de las sinusoides predominantes. Sin embargo, debido a la discretización del eje de frecuencias, el máximo local detectado no siempre coincide con el máximo ideal del lóbulo. Por ello se realiza una interpolación parabólica de los lóbulos, que permita estimar la posición del vértice del lóbulo, tal y como se explica en el subapartado 3.4.2. En el subapartado 3.4.3 se muestra la estructura de datos interna que ha sido utilizada para los parámetros de la señal en esta fase del análisis. Se observa en ella la necesidad de introducir un seguimiento temporal de las sinusoides para aplicar procesados que de mantengan sentido musical del sonido.

Por último, para comprender mejor las limitaciones de la estimación sinusoidal a través de la STFT, en el subapartado 3.4.4 se analiza una situación que se dará con frecuencia en el análisis de señales disonantes. Este es el caso de parciales con frecuencias próximas entre sí que no son resueltos adecuadamente como sinusoides independientes.

3.4.1. Detección de picos espectrales

Como ya se ha explicado en el subapartado anterior, el espectrograma (resultado de la STFT) ha sido calculado con una ventana Blackman-Harris de 92dB. Gracias al uso de esta ventana, las sinusoides aparecen con un único lóbulo principal en frecuencia, y el problema se reduce exclusivamente a calcular la cima de estos picos. Para detectar estos picos se ha usado un sencillo algoritmo encargado de detectar los máximos locales. Para ello simplemente busca los valores de magnitud que sean mayores que su anterior y su posterior al mismo tiempo. Es decir, existe un máximo en $k = k_r$ si cumple la siguiente condición:

$$|X_w[k_r]| \geq |X_w[k_r - 1]| \quad \wedge \quad |X_w[k_r]| \geq |X_w[k_r + 1]| \quad \wedge \quad |X_w[k_r]| \geq t \quad (3.24)$$

Estos máximos locales son tenidos en cuenta sólo si superan un cierto valor um-

bral t. De esta forma se evita que se detecten picos cuando sólo existe el suelo de ruido. Además, se detectarán sólo los K picos de mayor magnitud, evitando de esta forma un excesivo número de sinusoides detectadas. En realidad, si se trata de un sonido estándar con una componente sinusoidal estable, el número de armónicos significativos no debe superar algunas decenas. En la figura 3.7 se muestra un ejemplo de detección de picos espectrales dentro de una ventana.

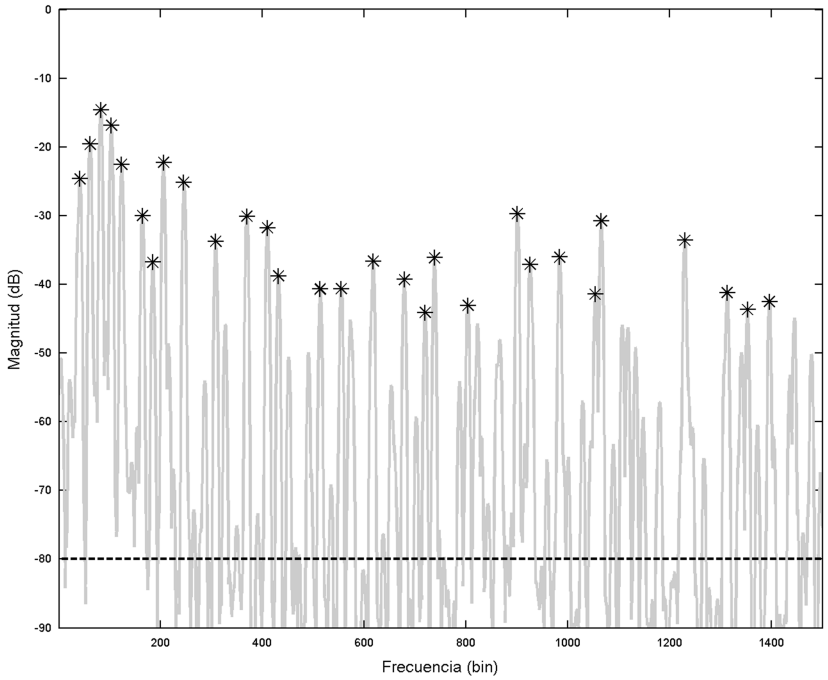


Figura 3.7: Detección de máximos locales en el espectro de un acorde La Mayor en guitarra acústica (sólo se muestra hasta 4KHz: 1500 bins en una FFT de 16384 bins). Umbral: -80 dB, Número máximo de picos detectados: 30, Tamaño de ventana (M): 8001, Tamaño de FFT (N): 16384, Tipo de ventana: Blackman-Harris 92dB

3.4.2. Interpolación parabólica de los picos

Debido a la discretización del eje de frecuencias, existe un problema asociado a la detección de picos que ha de tenerse en cuenta. Para entender de qué se trata, en la figura 3.8 se observa un zoom sobre uno de los picos detectados.

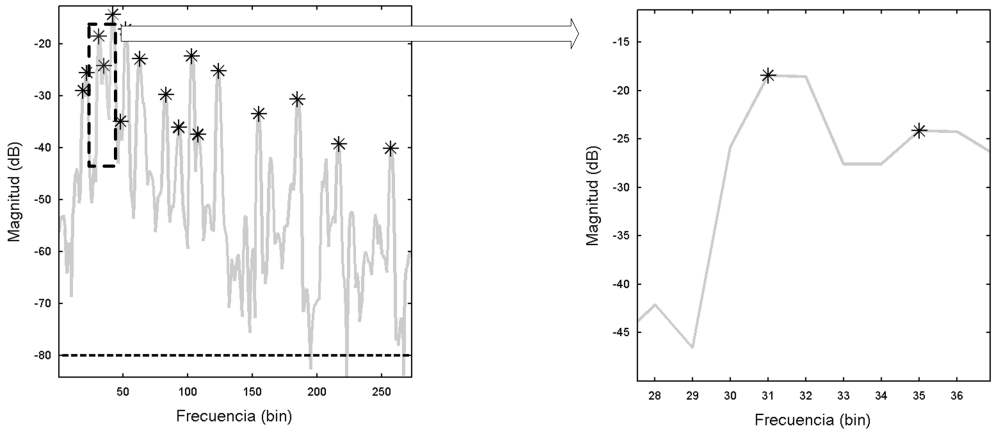


Figura 3.8: Zoom sobre dos de los picos detectados

En este zoom se observa que el valor detectado no está donde idealmente debería estar situado el máximo del lóbulo. La razón es que el eje de frecuencias es discreto, y el máximo del lóbulo no siempre coincide en un valor entero de k. Para evitar este error y estimar el valor de frecuencia correcto, se ha realizado una interpolación parabólica del logaritmo de las magnitudes del lóbulo. En [3] se demuestra que la parábola puede aproximar adecuadamente el lóbulo principal de varios tipos de ventanas.

Esta interpolación cuadrática se realiza tomando el máximo detectado junto con los dos valores colindantes, y se realiza una interpolación polinómica. Esta aproximación se realiza a partir de los siguientes valores:

$$y(-1) = \alpha = 20log(|X_w[k_p - 1]|) \quad (3.25)$$
$$y(0) = \beta = 20log(|X_w[k_p]|) \quad (3.26)$$
$$y(1) = \gamma = 20log(|X_w[k_p + 1]|) \quad (3.27)$$
$$(3.28)$$

Siendo k_p el bin correspondiente al máximo erróneo detectado. Esta información se puede utilizar para encontrar la posición exacta del máximo a partir de una aproximación parabólica:

$$\text{Frecuencia precisa del máximo (bin):} \qquad k_p^* = k_p + \frac{\alpha - \gamma}{2(\alpha - 2\beta + \gamma)} \quad (3.29)$$

$$\text{Magnitud precisa del máximo (dB):} \qquad a_p^* = \beta - \frac{\alpha - \gamma}{4}(k_p^* - k_p) \quad (3.30)$$

En la figura 3.9 se muestra la interpolación parabólica de la magnitud del espectro.

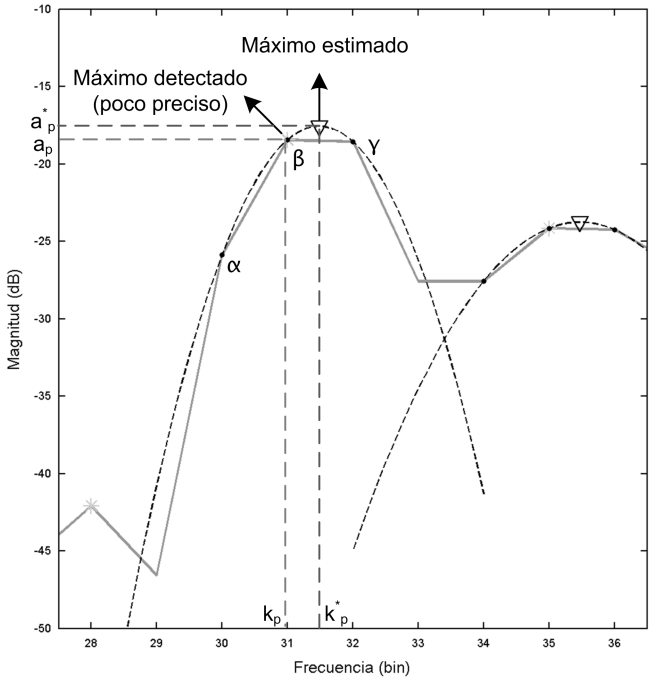


Figura 3.9: Interpolación de los lóbulos principales mediante parábolas. Los vértices de dichas parábolas han sido marcados con triángulos.

De esta forma, los valores de frecuencia y magnitud estimados para las sinusoides son suficientemente precisos. En cuanto a la fase, se aproxima en la posición k^* mediante interpolación lineal.

3.4.3. Estructura de datos utilizada

En este subapartado se expone la estructura de datos utilizada para almacenar la información sinusoidal en esta fase del análisis (figura 3.10). Esto permite comprender mejor qué información exactamente se ha utilizado en la parametrización. No obstante, para conseguir una representación más versátil desde el punto de vista musical es necesario un seguimiento temporal de sinusoides. Tras dicho seguimiento temporal, la unidad de trabajo no es el pico, sino el parcial (ver apartado 3.6.2).

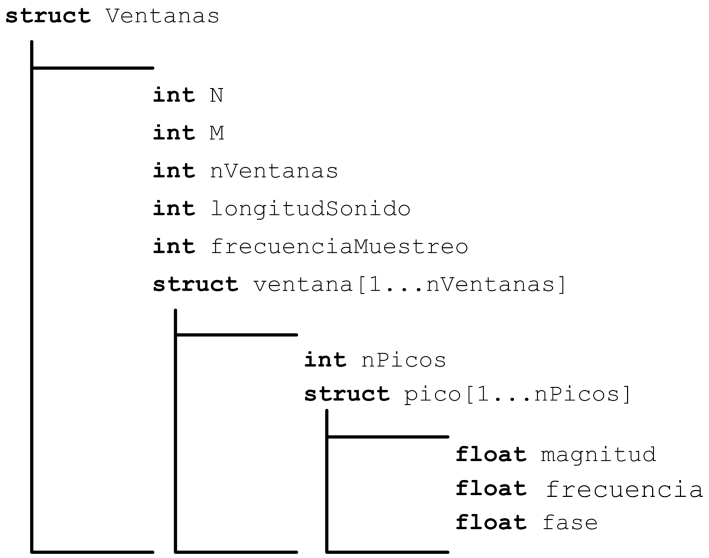


Figura 3.10: Estructura de datos utilizada para almacenar toda la información sinusoidal ventana a ventana. En Matlab no ha sido necesario especificar los tipos, pero se ha considerado conveniente incluirlos en el esquema.

3.4.4. Análisis de parciales cercanos en frecuencia

En este subapartado se analizará el resultado de la STFT ante parciales cercanos en frecuencia que no llegan a resolverse como sinusoides independientes. Esto es especialmente notable en el caso de sonidos desafinados, donde las sinusoides suelen "chocar" a menudo para producir batidos.

En la STFT siempre existe un compromiso entre la resolución temporal y la resolución frecuencial. Esto da lugar a una serie de consecuencias:

- Un tamaño de ventana demasiado pequeño ofrece buena resolución temporal, necesaria para un seguimiento fiel de los cambios de la componente sinusoidal. Sin embargo, no ofrece suficiente resolución frecuencial para distinguir parciales demasiado cercanos, por lo que apareran fusionados en una única sinusoide modulada en amplitud y/o frecuencia. Estas oscilaciones pueden empeorar el seguimiento temporal y puede dar lugar a un análisis pobre.

- Un tamaño de ventana demasiado grande permite una buena resolución frecuencial. Sin embargo, sólo se obtendrán buenos resultados con señales muy estables en el tiempo, debido a la mala resolución temporal.

- El compromiso óptimo entre ambos tipos de resolución depende de la señal de

entrada y del tipo de análisis que se vaya a realizar, por lo tanto no se puede definir a priori un correcto tamaño de ventana.

- En el caso de disonancias causadas por batidos de dos parciales, la separación entre ellos puede ser de unos pocos Hz. En estos casos, es necesario ventanas gigantescas (de varios segundos) para ser capaz de resolver correctamente ambos parciales. Sin embargo, el uso de ventanas tan grandes resulta inútil debido a la pésima resolución temporal que ofrecen, y por tanto surge la necesidad de abordar el problema de forma alternativa (apartado 4.2).

La cercanía entre parciales, y los problemas de resolución derivados son por tanto uno de los grandes problemas que han surgido en el desarrollo del Proyecto. En vista de esto, resulta conveniente comprender con detalle qué resultados ofrece la STFT en estos casos, además de disponer de algún apoyo matemático para comprender el por qué de dichos resultados.

Parciales cercanos con la misma amplitud

Supónganse dos parciales que tienen frecuencias próximas entre sí y la misma amplitud:

$$x(t) = Acos((\omega_0 - \Delta\omega)t) + Acos((\omega_0 + \Delta\omega)t) \tag{3.31}$$

Al realizar la STFT, dependiendo del tipo y del tamaño de la ventana se conseguirá mayor o menor resolución frecuencial. Si la ventana es demasiado pequeña, la señal será interpretada como un único tono modulado en amplitud. Si la ventana es suficientemente grande, el resultado mostrará dos picos distinguidos (correcta resolución de parciales). La siguiente expresión determina el tamaño adecuado para la ventana:

$$M > K\frac{2\pi f_s}{\Delta\omega} \tag{3.32}$$

Generalmente, las disonancias a baja frecuencias pueden ser apreciables con separaciones de unos cuantos Hz. Supóngase, por ejemplo, una $\Delta f = 5Hz$, el cual es un valor que puede aparecer fácilmente en parciales cercanos de señales reales. El tamaño adecuado de ventana, usando $f_s = 44100Hz$ y $K = 8$ (ventana Blackman-Harris 92dB) sería:

$$M > K\frac{2\pi f_s}{\Delta\omega} \rightarrow M > 70560 \tag{3.33}$$

Sin embargo este tamaño de ventana (1.6 segundos) resulta inútil para señales reales, ya que ofrecería una insuficiente resolución temporal. Por ello, conviene entender qué sucede cuando la ventana no ofrece resolución frecuencial suficiente.

En la figura 3.11 se muestra gráficamente la STFT de dos sinusoides cercanas para dos tamaños de ventana diferentes.

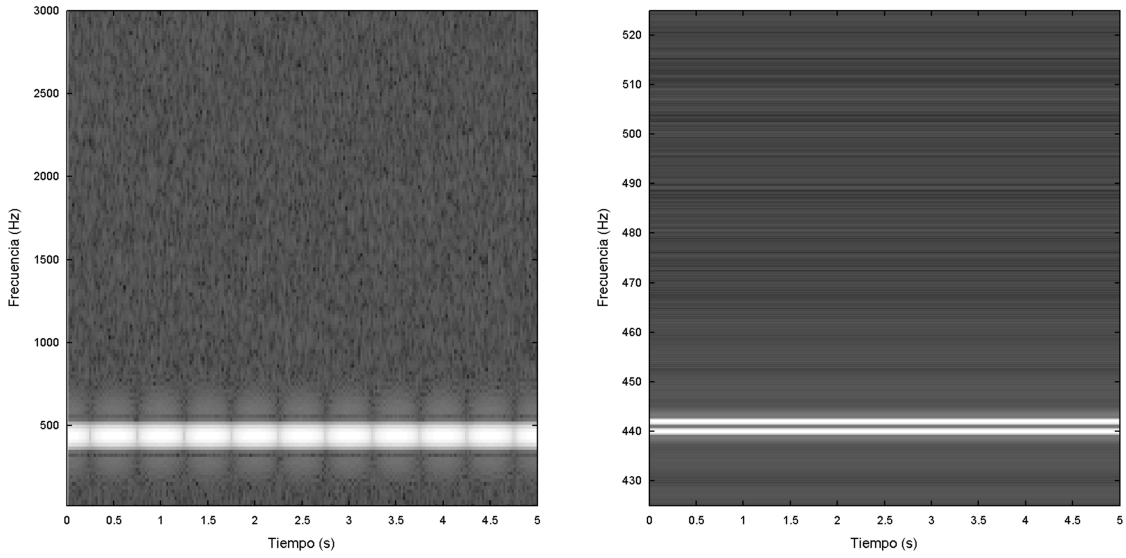


Figura 3.11: Resultado de la STFT para la suma de dos tonos cercanos en frecuencia (440Hz y 442Hz simultáneos, en este caso). El tamaño mínimo de ventana para la resolución de ambos parciales es M muestras. **(a)** Ventana de 2048 muestras (parciales no resueltos) **(b)** Ventana de 262144 muestras (parciales resueltos)

Ambas representación tiempo-frecuencia son válidas, ya que la suma de dos tonos puede ser expresada como un único tono con una modulación en amplitud de baja frecuencia:

$$x(t) = Acos((\omega_0 + \Delta\omega)t) + Acos((\omega_0 - \Delta\omega)t) = 2A \cdot cos(\Delta\omega_0 t) \cdot cos(\omega_0 t) \quad (3.34)$$

Dependiendo del tamaño de ventana, la estimación de la componente sinusoidal se acercará a una suma o a una modulación (existiendo ambiguas situaciones intermedias). En la figura 3.12 se muestra la representación tiempo-frecuencia de los máximos espectrales detectados para distintos tamaños de ventana:

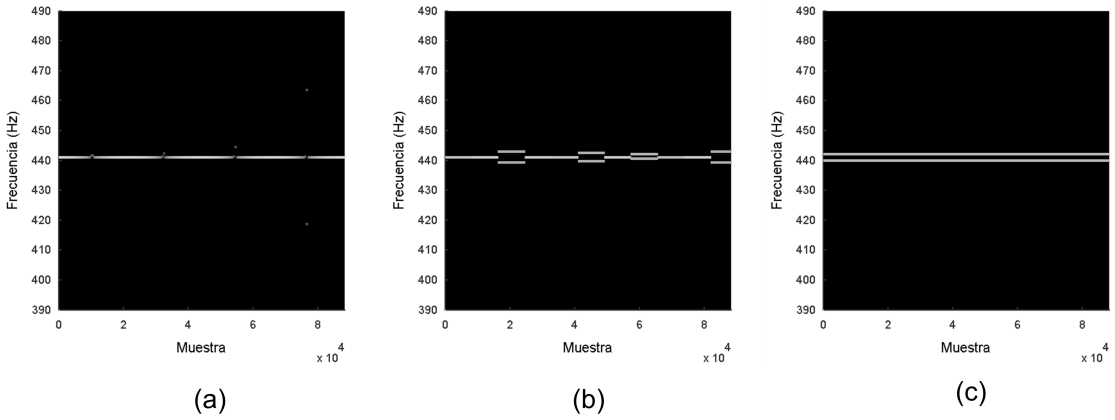


Figura 3.12: Estimación sinusoidal de dos tonos de 440Hz y 442Hz, con 0.5 de amplitud cada uno. **(a)** M = 2001, N = 2048 **(b)** M = 32001, N = 32768 **(c)** M = 128001, N = 131072

Parciales cercanos de distinta amplitud

Cuando dos parciales cercanos en frecuencia tienen amplitudes diferentes, se producirá una modulación en amplitud y una modulación en fase. Esta modulación en fase producirá variaciones de la frecuencia instantánea que dificultarán más aún la interpretación de la componente sinusoidal analizada. A continuación se realiza un análisis matemático para comprender mejor este fenómeno.

Supónganse dos parciales que tienen frecuencias próximas entre sí y amplitudes diferentes:

$$x(t) = A_1 cos((\omega_0 + \Delta\omega)t) + A_2 cos((\omega_0 - \Delta\omega)t) \tag{3.35}$$

En este caso, no existe una relación directa entre suma y producto de sinusoides. Sin embargo se puede llegar a algo parecido a través de un adecuado desarrollo matemático. Dicho desarrollo es complejo, y puede ser encontrado en [31], por lo que se aquí sólo se muestra el resultado final:

$$x(t) = \sqrt{A_1^2 + A_2^2 + 2A_1A_2cos(2\Delta\omega\ t)} \cdot cos(\omega_0\ t + \arctan\Big(\Delta\omega\ t \cdot \frac{A_1 - A_2}{A_1 + A_2}\Big)) \tag{3.36}$$

De esta expresión se puede extraer la frecuencia instantánea como la derivada de la fase instantánea:

$$\hat{\omega}(t) = \omega_a + \frac{(A_1^2 - A_2^2)\Delta\omega}{A_1^2 + A_2^2 + 2A_1A_2cos(2\Delta\omega t)} \tag{3.37}$$

Como se observa, la frecuencia instantánea es estable sólo en el caso de que las amplitudes de ambas sinusoides sean iguales. Cuando las amplitudes son diferentes, además de la modulación en amplitud existirá una modulación de la frecuencia. Para ventanas pequeñas, esta oscilación suele estar presente y es necesario tenerla en cuenta como un artificio indeseable propio del análisis, no de la señal (ver figura 3.13).

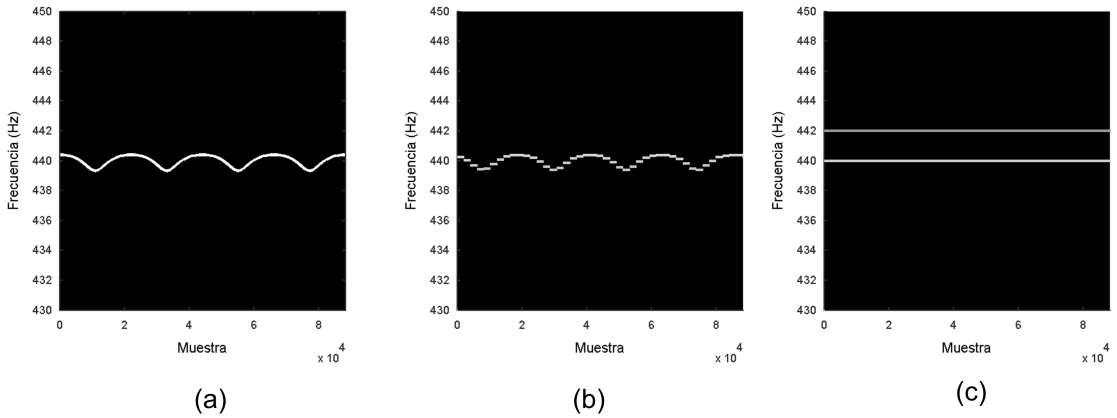


Figura 3.13: Análisis de dos tonos simultáneos de 440Hz y 442Hz, con amplitudes 0.8 y 0.2 respectivamente. **(a)** Frecuencia instantánea calculada según la expresion 3.37 **(b)** Resultado del análisis sinusoidal usando una ventana de M = 8001, N = 8192 **(c)** Resultado del análisis sinusoidal usando una ventana de M = 128001, N = 131072

En señales reales, cuando existen disonancias es común encontrar este tipo de patrones en los parciales que la forman. En la figura 3.14 se muestra un acorde de guitarra desafinado (Re Mayor), analizado con 3 tamaños de ventana diferentes.

Es evidente que los batidos entre parciales complican el análisis por temas de resolución frecuencial y temporal. Esto sucede habitualmente con sonidos disonantes, que desafortunadamente es el material de trabajo del sistema. Esta complicación añadida en la etapa de análisis se ha enfocado por métodos alternativos (4.2) con resultados interesantes.

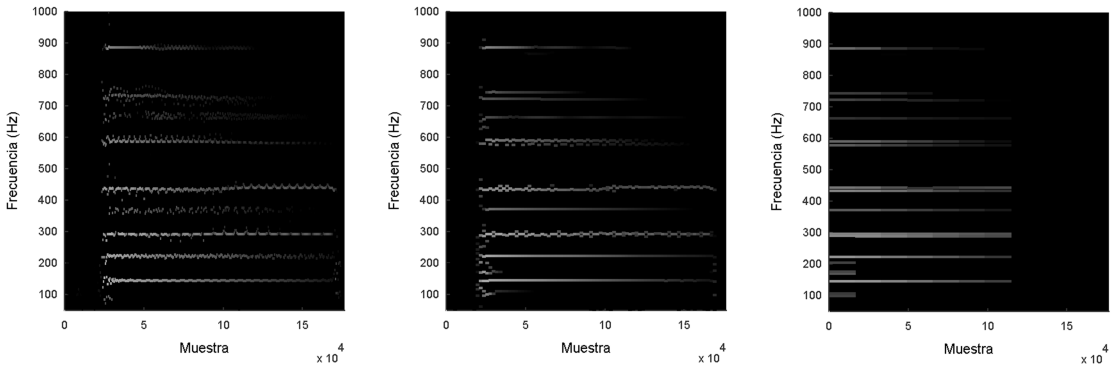


Figura 3.14: Acorde Re Mayor desafinado de guitarra acústica analizado con tres tamaños de ventana diferentes **(a)** M = 2001, N = 2048 **(b)** M = 8001, N = 8192 **(c)** M = 64001, N = 65536

3.5. Extracción de la componente residual

Tal y como se explicó en el apartado 3.2, el modelo que se está utilizando parametriza la componente sinusoidal y almacena aparte la componente residual. El procedimiento habitual es estimar la componente sinusoidal, sintetizarla y sustraérsela a la señal original. La componente residual suele almacenar los ataques y todos los elementos ruidosos que no dan lugar a picos pronunciados en el espectro.

La síntesis sinusoidal se expone con detalle en el apartado 5.3, ya que es un bloque se utiliza también en el Subsistema de síntesis. La información necesaria para la síntesis es la parametrización ventana a ventana de la componente sinusoidal, es decir, la estructura de datos mencionada en 3.4.3.

Una vez que se dispone de la componente sinusoidal $x_{sin}[n]$, teniendo en cuenta el modelo:

$$x[n] = x_{sin}[n] + x_{res}[n] \tag{3.38}$$

se puede obtener la componente residual por simple substracción en el dominio temporal, muestra a muestra. Esta substracción generalmente ofrece buenos resultados, y garantiza que la señal original puede ser regenerada sin pérdidas. La operación realizada es:

$$x[n] - x_{sin}[n] = x_{res}[n] \tag{3.39}$$

La "limpieza" de la señal residual va a depender en gran medida de la calidad del análisis sinusoidal. En la figura 3.15 se observa cual es el resultado tras la extracción de la componente residual en un tono ruidoso.

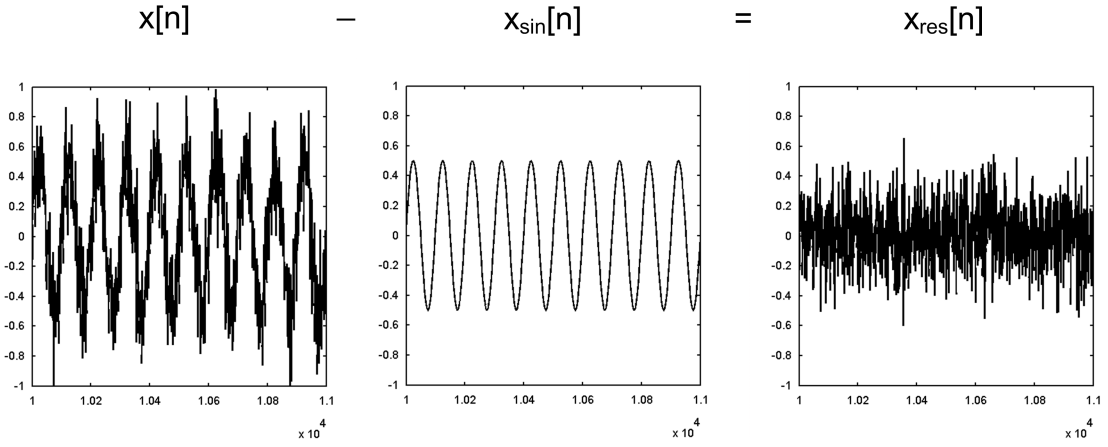


Figura 3.15: Descomposición de un tono de 440Hz de amplitud 0.5 sumado con una componente de ruido gaussiana de desviación estándar de 0.2. Correctamente ajustado, el sistema de análisis consigue detectar la componente sinusoidal con suficiente precisión.

En este caso el análisis sinusoidal es muy fácil de realizar, ya que se trata de una única sinusoide estable a lo largo de toda la señal. Gracias a la eficacia del análisis sinusoidal en este caso, la componente residual puede ser limpiamente extraída de la señal original.

En [42] se propone un método alternativo para extraer la componente residual. Este método no se basa en la substracción en el dominio del tiempo, sino en el espectro de magnitudes. En el siguiente subapartado se realizan algunos comentarios acerca de esta técnica "alternativa".

Descomposición basada en el espectro de magnitudes

El método alternativo para extraer la componente residual, propuesto por Serra en [42], consiste en "cortar" del espectro de magnitudes los lóbulos correspondientes a sinusoides y obtener así la componente sinusoidal. De esta forma se dispondría del espectro de la componente sinusoidal por un lado, y de la componente residual por el otro. En este trabajo se ha comprobado que este sistema no ofrece resultados satisfactorios por varias razones:

- Es habitual encontrar lóbulos solapados, correspondientes a sinusoides cercanas entre sí. Estos lóbulos tienen formas anómalas, y al recortarlos no se corresponden con ninguna ventana en concreto. Para que un lóbulo represente una sinusoide enventanada, la forma de dicho lóbulo debe corresponder fielmente al de alguna ventana.

- Dado que la componente sinusoidal no puede extraerse correctamente con este sistema, la componente residual tampoco ofrece resultados satisfactorios.

El único caso en el que la extracción de la componente sinusoidal a partir del espectro de magnitudes puede realizarse con éxito es cuando las sinusoides están separadas y sus lóbulos están perfectamente definidos sobre el resto de componentes espectrales. Esto, no obstante, no suele darse al trabajar con señales musicales.

3.6. Seguimiento temporal de parciales

En el apartado anterior se ha explicado el procedimiento para detectar las sinusoides encontradas en cada ventana. Sin embargo, en dicho procedimiento no se contempla la continuación temporal de un mismo parcial. En este apartado se explicará de qué forma se ha realizado el seguimiento temporal de parciales.

Como ya se explicó en el capítulo 2, el procedimiento escogido para el seguimiento temporal es el descrito en [42]. En él se establece una máscara dentro de la cual debe caer la sinusoide para ser detectada como continuación de la anterior. Esta máscara se define en magnitud, frecuencia, y tiempo. Por tanto, dos picos se han considerado pertenecientes al mismo parcial si se cumplen simultáneamente las siguientes condiciones:

- **Magnitud:** Si la diferencia de magnitud entre esos dos picos es menor de 20dB.
- **Frecuencia:** Si la distancia en frecuencia es menor de 0.2 semitonos.
- **Tiempo:** Si la separación temporal entre los dos picos es menor de 700ms.

Esto garantiza que sinusoides próximas de características similares son tomadas como parte del mismo parcial. En la figura 3.16 se muestra un ejemplo que lo ilustra gráficamente:

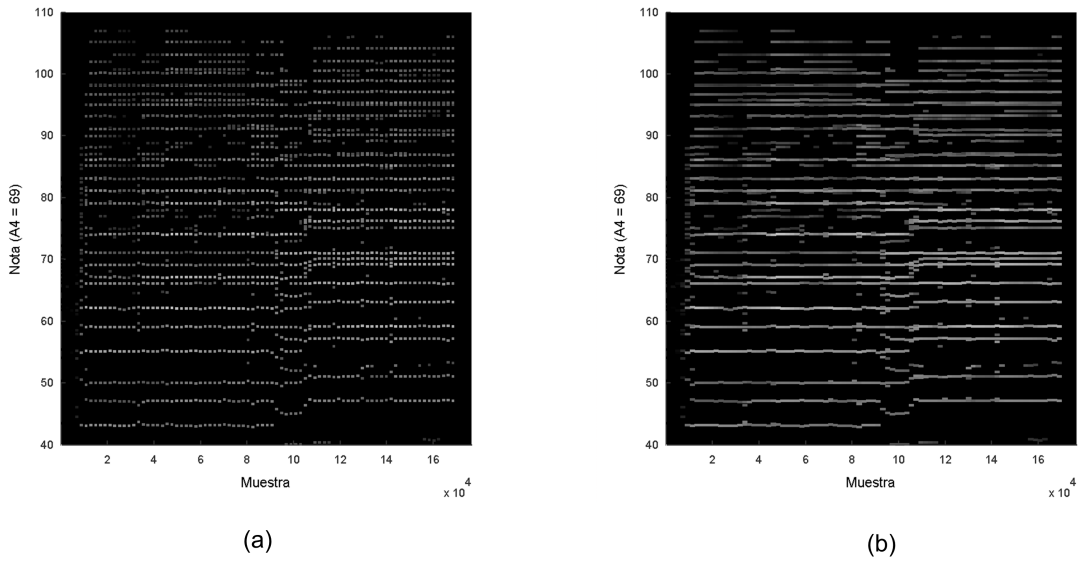


Figura 3.16: Sonido analizado: Piano desafinado tocando acordes estables en registro medio/grave. **(a)** Conjunto de sinusoides detectadas en cada ventana a lo largo de toda la señal. **(b)** Resultado del seguimiento temporal. Se pueden observar casos en los que ha sido necesario enlazar un parcial "quebrado" mediante interpolación.

En el caso de que dos sinusoides no sean contiguas, el enlace de ambas se realiza mediante interpolación. La interpolación entre dos picos no contiguos es lineal en frecuencia, magnitud y fase. Como se observa en la figura 3.16, al realizar el seguimiento temporal aparecen nuevos picos, fruto de la interpolación mencionada.

Esta interpolación es necesaria porque en ocasiones un mismo parcial es detectado de forma intermitente por fluctuaciones del mismo, dejando pequeños huecos que lo fragmentan en varios falsos parciales. Estos falsos parciales no son representativos para conocer la duración del parcial real, ni tampoco para conocer el momento de comienzo y fin del mismo. La interpolación, por tanto, ayuda a mantener la coherencia y evita la fragmentación innecesaria de los parciales.

El seguimiento temporal de parciales es una etapa crítica del sistema, tanto a nivel de coste computacional como a nivel funcional. La forma en la que está planteado el seguimiento temporal requiere un barrido exhaustivo de todos los picos, lo cual ralentiza notablemente el funcionamiento global del sistema. Pero además, el método propuesto comete errores importantes en el caso de parciales inestables. Esta dificultad para identificar picos pertenecientes al mismo parcial se puede considerar un cuello de botella del sistema a nivel computacional, y a nivel de comportamiento global.

3.6.1. Eliminación de parciales de menos de 200ms

Dentro del bloque de seguimiento temporal, también se tiene en cuenta lo comentado en el apartado 2.2. Una vez que los parciales ha sido delimitados, se comprueba su duración y de descartan todos aquellos que no superen los 200ms. De esta forma se evita trabajar con contenido sinusoidal que debería pertenecer a la componente residual. El valor de 200ms puede ser incrementado en algunos casos para mejorar la representación de la componente sinusoidal. En la figura 3.17 se muestra la estructura de parciales antes y después de realizar el filtrado por duración.

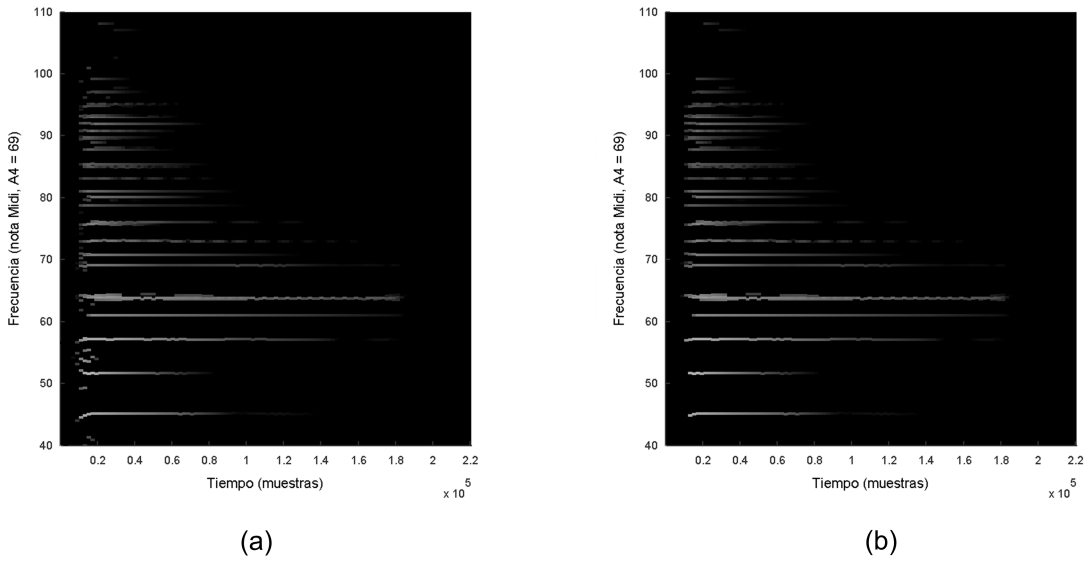


Figura 3.17: Sonido analizado: Acorde La Mayor tocado en guitarra desafinada. **(a)** Estructura de parciales sin filtrado por longitud. **(b)** Estructura de parciales tras eliminar los parciales demasiado cortos.

3.6.2. Estructura de datos utilizada

La información resultante del seguimiento temporal es almacenada en una estructura de datos diferente. Esta estructura pretende ser más útil para aplicar transformaciones, y consta de los siguientes elementos:

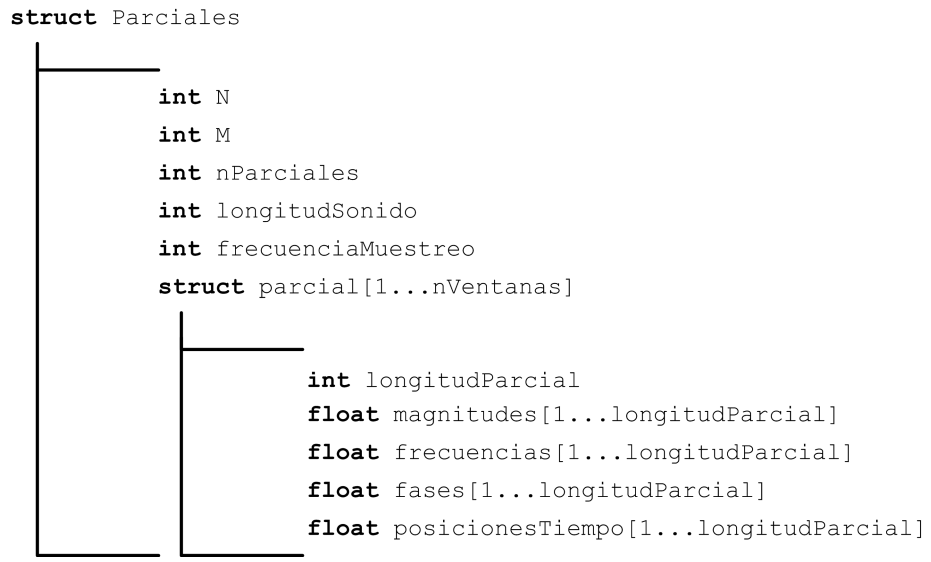


Figura 3.18: Estructura de datos de los parciales tras el seguimiento temporal.

Este nuevo nivel de abstracción en la estructura de datos hace de la información algo fácilmente manipulable para aplicar transformaciones musicales.

3.7. Detección de las frecuencias fundamentales predominantes

El Sistema generalmente va a trabajar con señales compuestas de distintas notas, formando un acorde con sentido musical que será analizado y procesado. La bonanza del Sistema va a estar relativamente determinado por su capacidad para encontrar las notas predominantes del acorde. No es imprescindible la correcta detección de todas las notas que están sonando, pero al menos sí debe ser capaz de estimar aquellas de mayor importancia.

Este es un problema muy complejo, sobre el cual se investiga activamente hoy día y que no está totalmente resuelto. Una de las aportaciones más interesantes corresponde a Klapuri (véase [25]), quien desarrolla un método interesante para la detección de múltiples frecuencias fundamentales basado en conceptos perceptuales. En este Proyecto se planteó inicialmente la implementación de este algoritmo, pero debido a su complejidad, su inclusión se propone como línea de investigación futura.

La solución propuesta es mucho más simple, y también ofrece peores resultados, aunque puede llegar a ser suficiente para casos sencillos. Consiste en considerar los

n parciales de mayor amplitud y duración[1] como las n "frecuencias fundamentales", incluyendo la norma de no repetir la misma nota en diferentes octavas. Este método no ofrece realmente las frecuencias fundamentales, pero ofrece una serie de resultados que pueden llegar a ser aceptables por dos motivos:

- Generalmente, los armónicos de mayor peso en una polifonía corresponden a notas de gran importancia armónica en el acorde, o con una relación armónica cercana con las mismas. Esto es así porque las notas más importantes del acorde suelen ser las más potentes, y los primeros armónicos suelen ser los más presentes de cada nota, por lo que los armónicos con más peso dentro del acorde serán los primeros armónicos de las notas importantes.

- El sistema no requiere la perfecta detección de todas las notas del acorde, sino que sólamente necesita una serie de referencias que le permita determinar aproximadamente las posiciones del contenido armónico. Una desviación de octava, o la no detección de ciertas notas puede producir la pérdida de algunos armónicos, pero seguirá existiendo consonancia entre ellos.

A partir del procedimiento descrito se extrae un array de frecuencias fundamentales `f0[i]`. El número de frecuencias detectadas se considera un parámetro del Sistema, y por defecto ha sido fijado a 5. En general, para acordes de 3 notas, el resultado es relativamente satisfactorio con 5 f_0s. En la figura 3.7 se muestra el resultado aplicado a un acorde de guitarra acústica.

A pesar de que numerosas f_0s no han sido correctamente detectadas, el resultado del Sistema con el sonido original es relativamente satisfactorio. La razón es que el A2 (nota 45) detectado contiene prácticamente toda la información armónica del acorde completo. Tal es la robustez del diseño global a la mala detección de frecuencias fundamentales. En cualquier caso, si la detección de f_0s fuera correcta, el resultado global resultaría mejor.

[1] Los parciales se ordenan según el producto (Magnitud máxima × Duración)

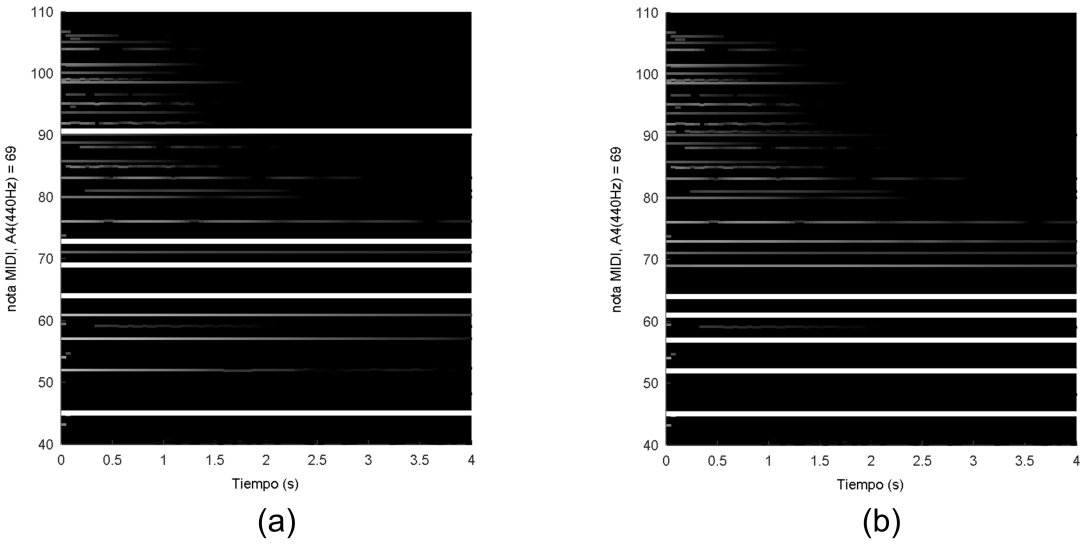


Figura 3.19: Acorde La mayor de guitarra acústica afinada (compuesto por A2, E3, A3, C#4, E4). En blanco intenso se muestran las f0s. **(a)** Estimación de las f0s según el método propuesto **(b)** Posición real de las f0s]

Además del método automático para la detección de frecuencias fundamentales, el Sistema ofrece la posibilidad de introducir manualmente estas f_0s. Esto puede ser práctico en determinadas ocasiones, ya que en numerosas ocasiones el usuario puede saber de antemano qué notas forman el acorde al que le quiere mejorar la afinación.

Capítulo 4

Subsistema de Procesado

En este apartado se detalla el diseño y la implementación del Subsistema de Procesado (ver diagrama del apartado 1.2).

En el apartado 4.1 se presenta un diagrama funcional del subsistema. Este Subsistema es la parte más original del Proyecto, puesto que las técnicas utilizadas no han sido extraídas de trabajos previos. Este diagrama consta de cinco bloques funcionales, los cuales están detallados en los apartados posteriores.

En el apartado 4.2, se explica de qué forma se han evitado las fluctuaciones indeseadas de parciales que son fruto de varios parciales interfiriendo entre sí. Esto sucede cuando la separación entre parciales está por debajo de la resolución del sistema (ver subapartado 3.4.4). Los parciales a partir de este bloque estarán caracterizados con un único valor de frecuencia, un único valor de fase y la envolvente de amplitud.

De forma paralela, las $f_0 s$ estimadas son procesadas para corregir las desviaciones de afinación de forma musicalmente coherente, tal y como se explica en el apartado 4.3. En este bloque funcional hay que introducir algunas presuposiciones de tipo musical acerca del material de entrada. En general dichas presuposiciones serán bastante acertadas para el tipo de material para el que está orientado el Sistema.

Una vez que se conocen las posiciones ideales de las $f_0 s$ que componente el acorde, se crea la estructura armónica a la que debería estar mayormente ajustado el sonido original para que sonara de forma consonante. Esto se consigue introduciendo una cantidad limitada de armónicos a las f_0s detectadas. Los detalles pueden encontrarse en el apartado 4.4.

En el apartado 4.5 se detalla el procedimiento con el cual se traslada la estructura armónica original (disonante) a la nueva estructura armónica (consonante). El funcionamiento es simple, ya que se basa en aproximar cada parcial a su posición más cercana de la "rejilla" ideal de armónicos. Los parciales cuya frecuencia sea muy alta no se procesan, ya que en numerosas ocasiones existe una gran inharmonicidad

en ellos, y no influyen significativamente en la percepción de la disonancia.

Por último, para evitar desplazamientos absolutos de frecuencia en el sonido original, se realiza un ajuste a la afinación estándar $La3 = 440Hz$ (o en la notación anglosajona, $A4 = 440Hz$) (véase apartado 4.6).

El resultado de este procesado ofrece como salida un nuevo array de parciales, que teóricamente representa una versión consonante de la señal de entrada.

4.1. Diagrama de bloques

En este apartado se muestra un diagrama de bloques de este subsistema.

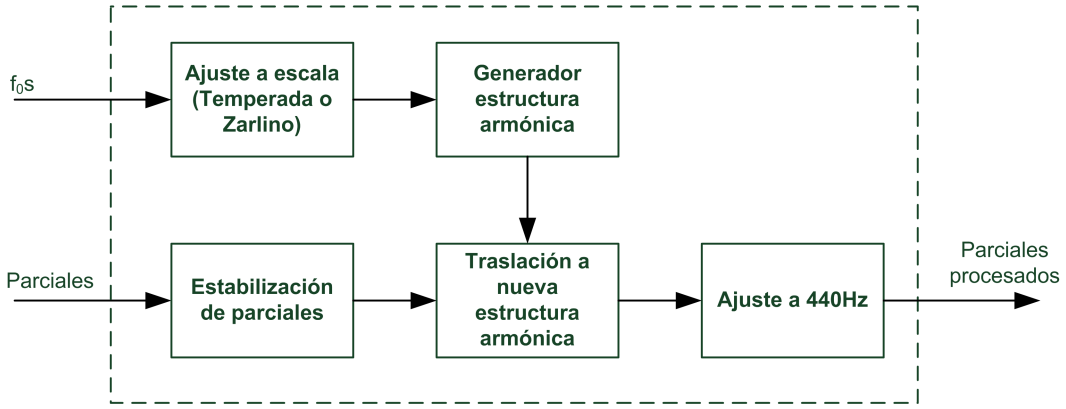


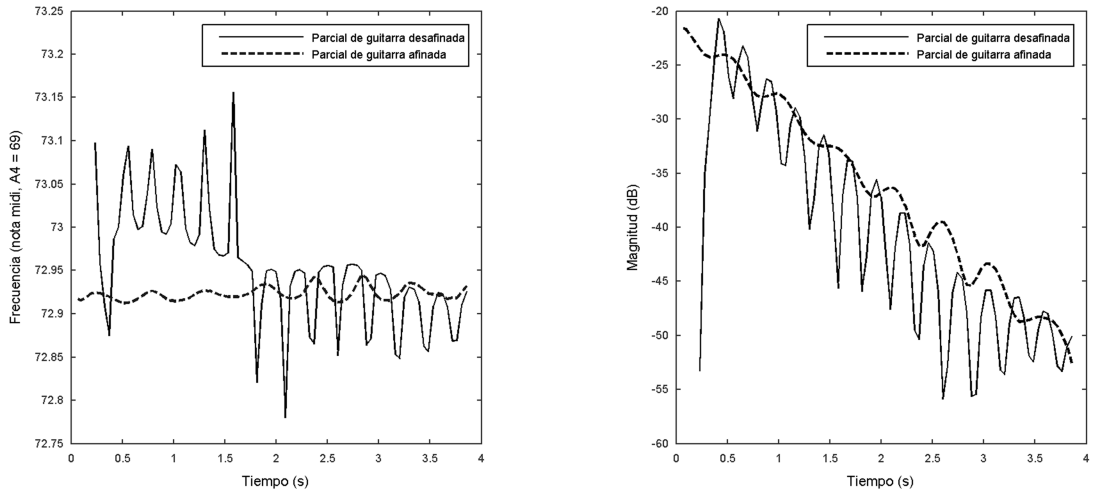
Figura 4.1: Diagrama de bloques del Subsistema de Procesado

En los siguientes apartados se explican los detalles de cada bloque funcional.

4.2. Estabilización de parciales

Una de las suposiciones que se hacen sobre el audio de entrada es que consiste en un único acorde estable en frecuencia. Esto significa que idealmente, cada parcial no debería fluctuar demasiado a lo largo de su evolución temporal. Esta suposición descarta todo tipo de vibratos, glissandos, u otras modulaciones en el sonido de entrada al Sistema.

Sin embargo, cuando se trabaja con desafinaciones es común encontrar parciales inestables en frecuencia, fruto del batido de dos parciales que no han sido resueltos correctamente. En la figura 4.2 se muestra una comparación entre dos parciales, extraídos de un mismo acorde de guitarra (La Mayor en su posición estándar). Un parcial corresponde a la guitarra afinada, y el otro corresponde a la misma guitarra tras desafinarla.



(a) (b)

Figura 4.2: **(a)** Evolución de la frecuencia respecto al tiempo. El parcial se situa alrededor de la nota 73, que es un Do#3 (o C#4). Este parcial corresponde a un acorde La Mayor dado con una guitarra acústica. **(b)** Evolución de la magnitud con respecto al tiempo. Al igual que sucede con la frecuencia, la magnitud también está modulada en el caso del parcial perteneciente a la guitarra desafinada.

Ante la imposibilidad de interpretar adecuadamente estos parciales con fluctuaciones, se plantearon inicialmente dos enfoques diferentes para tratar con este problema:

1. Aplicar algún tipo de técnica multiresolución para conseguir detalles sobre los parciales que están produciendo el batido. Se planteó calcular la FFT con un tamaño de ventana igual al parcial completo, y detectar con precisión la frecuencia de las sinusoides que lo forman. Esto combinaría la resolución frecuencial de una ventana larga con la resolución temporal de una ventana corta. Sin embargo, la implementación de este análisis es compleja, por lo que se ha recurrido a otro método más sencillo que ofrece resultados aceptables (con mucho menos coste computacional).

2. El enfoque finalmente llevado a cabo consiste en eliminar las modulaciones del parcial problemático aunque no se disponga de información detallada sobre las sinusoides que están produciendo el batido. Esta técnica tiene dos objetivos: resolver el problema de la disonancia tonotópica a bajas frecuencias (donde las distancias entre parciales son tan pequeñas que no se pueden resolver), y parametrizar los parciales con un único valor de frecuencia. El problema que se plantea es que no es posible conocer el valor de frecuencia correcto antes de valorar el sonido de forma global. Por tanto, esta técnica corrige la disonancia tonotópica, pero introduce estructuras armónicas antinaturales que son también percibidas como disonantes. No obstante, en etapas posteriores estos parciales se trasladan a posiciones que son percibidas como naturales, por lo que esta técnica resulta ser adecuada.

La observación que llevó a plantear este último enfoque es que cuando se trabaja con sonidos perfectamente afinados, dejan de existir parciales oscilantes y el contenido sinusoidal se vuelve más estable. En base a esto, para resolver la desafinación se empieza por estabilizar los parciales del sonido para acercarse al aspecto de un espectro consonante. En la figura 4.3 se muestra en qué consiste este proceso aplicado a un sonido real.

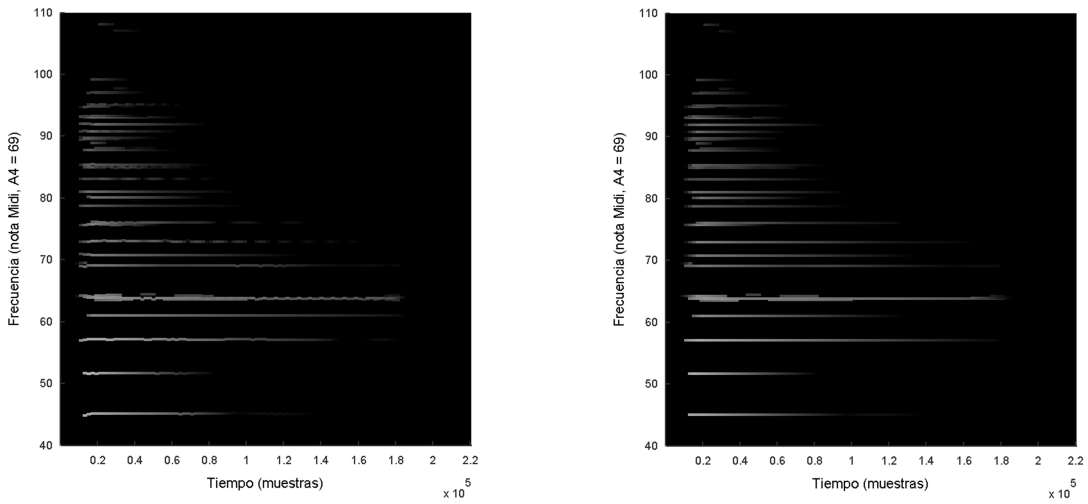


Figura 4.3: Acorde La Mayor desafinado de una guitarra acústica: **(a)** Parciales antes de la estabilización **(b)** Parciales tras la estabilización. Se observa que dejan de existir modulaciones, tanto en frecuencia como en amplitud.

A continuación se desarrollan los detalles sobre la estabilización de los distintos

parámetros del parcial.

4.2.1. Reconstrucción de la envolvente

Uno de los efectos del batido entre parciales es la modulación en amplitud. Para eliminarla se procesa la envolvente original del parcial, interpolando los máximos de cada oscilación. Sería algo parecido a la demodulación incoherente de AM. En la figura 4.4 se muestra en qué consiste este proceso.

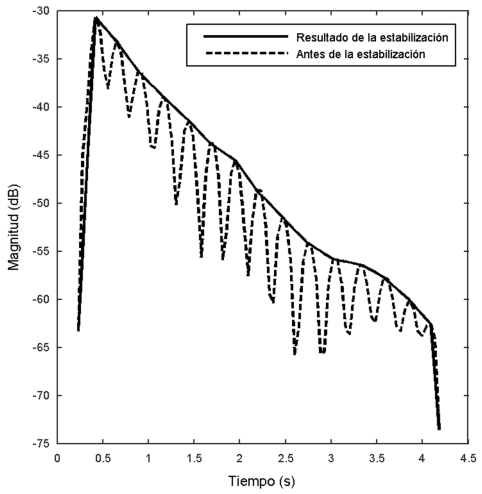


Figura 4.4: Evolución de la magnitud con respecto al tiempo para el mismo parcial, antes y después de reconstruir su envolvente.

La envolvente final intenta ser fiel al original en el ataque de la nota, y tan sólo se elimina la modulación en amplitud de la parte sostenida de la nota. Esta forma de reconstruir la envolvente tiene varios problemas:

- La potencia media del parcial aumenta, ya que se están interpolando los máximos producidos por la interferencia coherente de sinusoides. En ocasiones, esto puede influir en el sonido final, haciéndole perder naturalidad.

- Este tipo de interpolación es válida solamente para parciales con un solo ataque. Sería necesario implementar un método alternativo para contemplar casos más generales.

4.2.2. Estabilización de la frecuencia

Otro de los efectos de la desafinación sobre los parciales es que se producen modulaciones de la frecuencia (ver apartado 3.4.4). La técnica propuesta consiste en hacer completamente constante los valores de frecuencia a lo largo del parcial. El valor de frecuencia aplicado se calcula mediante la media aritmética:

$$\overline{f_k} = \frac{\sum_{l=1}^{\text{longitudParcial}} f_k[l]}{\text{longitudParcial}}$$

Donde

k	=	Índice de parcial
l	=	Índice de ventana temporal
longitudParcial	=	Longitud del parcial en ventanas temporales
$f_k[l]$	=	Valor de frecuencia del parcial k en la ventana temporal l
$\overline{f_k}$	=	Promedio temporal de la frecuencia del parcial

Posteriormente se aplica dicho valor a todo el parcial, estabilizando su frecuencia y evitando oscilaciones indeseadas. Es importante ser consciente de que el valor de frecuencia f_k no tiene por qué corresponder con el valor que idealmente eliminaría la disonancia. Esto se ejemplifica adecuadamente en la figura 4.5:

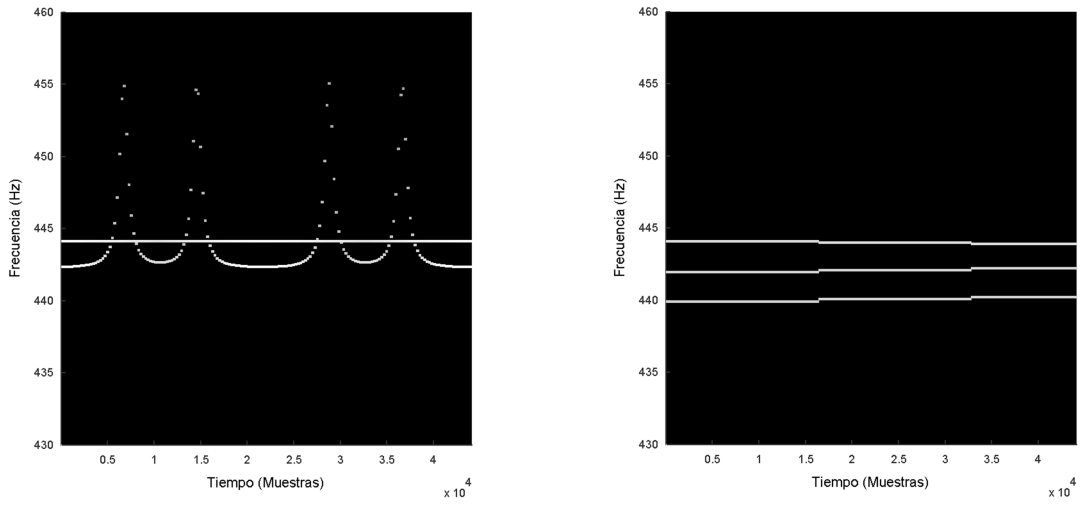


Figura 4.5: En este ejemplo se han usado 3 tonos, de frecuencias 440Hz, 442Hz y 444Hz, con amplitudes 0.3, 0.3 y 0.4 respectivamente. **(a)** Parcial oscilante resultado de una ventana pequeña, y resultado de la estabilización según se ha explicado anteriormente (línea horizontal), ambos sobre la misma gráfica. **(b)** Resolución precisa de los tres parciales usando una ventana larga.

Los valores de frecuencia resultantes tras la estabilización no tienen por qué ser los convenientes para la consonancia de conjunto, ya que no se está teniendo en cuenta el resto de parciales. Sin embargo, esto se corrige en etapas posteriores del Subsistema de Procesado.

4.2.3. Cálculo de la fase

Una vez que se han modificado los valores de frecuencia, la evolución de la fase ha de ser adaptada a este nuevo valor para garantizar la correcta continuidad de la sinusoide. Para ello se ha aplicado la siguiente fórmula que proporciona el valor de fase necesario para cada instante del parcial:

$$\phi_k^{l\,\prime} = \phi_k^{l-1\prime} + \frac{H \cdot 2\pi f_k}{f_s}$$

Donde:

$\phi_k^{l\,\prime}$ = Valor de fase del parcial k para la ventana l
H = Tamaño de salto. Normalmente $N/4$ para la ventana Blackman-Harris

f_k = Frecuencia tras la estabilización para el parcial k
f_s = Frecuencia de muestreo (4.1)

Un correcto valor para el vector de fase es imprescindible para conseguir que un parcial sea realmente una sinusoide correctamente continuada de ventana a ventana.

4.3. Ajuste de las $f_0[i]$s a una escala dada

Como ya se explicó en el apartado 3.7, una de las salidas del Subsistema de Análisis es una array de frecuencias fundamentales `f0[i]`, que idealmente deberían ser las de las notas que forman el acorde analizado. Generalmente, la desafinación indeseada va a suceder cuando estas $f_0[i]$ deberían formar parte de una escala determinada, pero en realidad no lo hacen del todo.

El procedimiento que se ha utilizado para "afinar" las $f_0[i]$ es estimar cuáles son las posiciones que idealmente deberían tener estas notas y desplazarlas adecuadamente. Para ello se toma como nota referente la más baja de todas las detectadas, y considerando ésta como primera nota, se contruye toda la escala. En concreto se ha trabajado con escalas de uso común en música (ver subapartados 4.3.1 y 4.3.2). Posteriormente, todas las $f_0[i]$s detectadas (supuestamente desafinadas) se trasladan a la nota de la escala más cercana (subapartado 4.3.3).

La notación utilizada en los siguientes subapartados es:

$\mathrm{f}_0[i]$ = Frecuencias fundamentales estimadas en el acorde de entrada
$\mathrm{f}_{escala}[i]$ = Frecuencias de la escala a la que pertenece el acorde
$\mathrm{f}_0 * [i]$ = Frecuencia de las notas ajustadas a la afinación correcta

En la figura 4.6 se representa gráficamente en qué consiste este proceso.

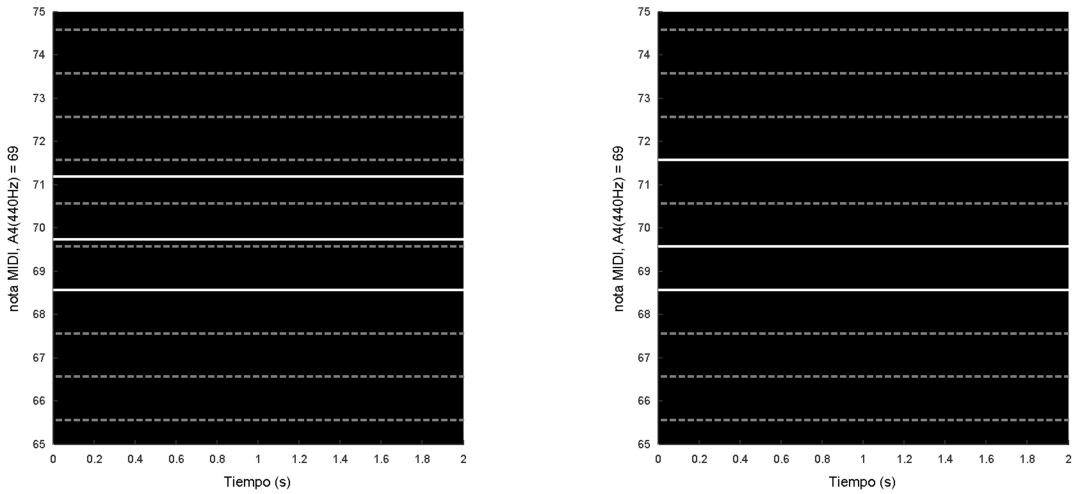


Figura 4.6: **(a)** Posición de las $f_0[i]$ estimadas (línea blanca contínua), y las posiciones de la $f_{escala}[i]$ (línea gris discontinua) **(b)** Ajuste de las $f_0[i]$ estimadas a la escala

A continuación se explicará cómo se construye $f_{escala}[i]$ a partir de $f_0[0]$ en el caso de utilizar la escala Temperada (subapartado 4.3.1) o la escala de Zarlino (subapartado 4.3.2).

4.3.1. Escala Temperada

En el caso de la escala temperada (escala más comúnmente usada hoy día), los intervalos pueden ser medidos en semitonos. Un semitono es idéntico a lo largo de toda la escala, y corresponde con una razón de frecuencias de $2^{\frac{1}{12}}$. Por ello, el factor $2^{\frac{i}{12}}$ permite generar toda la escala simplemente variando el parámetro i. De esta forma, partiendo de la f_0 más grave de todas las estimadas (es decir, `f0[0]`), las frecuencias de la escala han sido calculadas de la siguiente forma:

$$f_{escala}[i] = 2^{\frac{i}{12}} \cdot f_0s[0] \qquad i \in [-12, 48] \tag{4.2}$$

Es decir, la "rejilla" se sitúa en cada semitono de la escala temperada. El rango de posibles f_0s abarca desde una octava por debajo de la f_0 más grave, hasta 4 octavas por encima de la misma, ya que cada octava tiene 12 semitonos.

La expresión 4.2 es práctica cuando no se tiene absolutamente ninguna información del audio de entrada. Si se sabe que los acordes de entrada van a ser triadas[1]

[1] Acorde formado por tres notas

mayores o menores exclusivamente, se puede construir una escala que facilite el ajuste de cada f_0 a su posición correcta.

Triada mayor

Teniendo en cuenta que un acorde mayor puede estar dado en distintas inversiones, los intervalos posibles entre las notas que forman el acorde y la más grave son los siguientes:

- **Estado fundamental:** Tercera mayor y quinta justa. Ejemplo: DO - MI - SOL
- **Primera inversión:** Tercera menor y sexta menor. Ejemplo: MI - SOL - DO↑[2]
- **Segunda inversión:** Cuarta justa y sexta mayor. Ejemplo: SOL - DO↑ - MI↑

Teniendo en cuenta esto, se pueden excluir algunos intervalos de la escala y aun así se garantizaría el correcto procesado de las f_0s de un acorde mayor. Teniendo en cuenta la distancia en semitonos para cada intervalo:

Intervalo	Nº semitonos
Unísono	0
Tercera menor	3
Tercera mayor	4
Cuarta justa	5
Quinta justa	7
Sexta menor	8
Sexta mayor	9
Octava	12

Se puede reducir el número de notas necesarias usando el siguiente vector de escala:

```
E = [0 3 4 5 7 8 9]
```

Y aplicando la expresión 4.3 de la siguiente forma:

$$f_{escala}[i] = 2^{\frac{E[i]}{12}} \cdot f_0[0] \qquad i \in [-7, 28] \tag{4.3}$$

En este caso, la nueva escala tiene 7 notas (en lugar de 12), por lo que $i \in [-7, 28]$ garantiza que se están abarcando 5 octavas. Básicamente se trata de la misma escala temperada de la expresión 4.2 eliminando los intervalos de segunda menor, segunda mayor, cuarta aumentada, séptima menor y séptima mayor. La razón es que estos intervalos no aparecen en ninguna inversión de la triada mayor.

[2]La flecha ascendente indica octava alta

Triada menor

Aplicando el mismo razonamiento, se puede analizar qué intervalos pueden existir en este caso:

- **Estado fundamental:** Tercera menor y quinta justa. Ejemplo: DO - MI♭[3] - SOL
- **Primera inversión:** Tercera mayor y sexta mayor. Ejemplo: MI♭ - SOL - DO↑
- **Segunda inversión:** Cuarta justa y sexta menor. Ejemplo: SOL - DO↑ - MI♭ ↑

Si se ordenan los intervalos se llega exactamente a la misma escala que en el caso de acordes mayores:

```
E = [0 3 4 5 7 8 9]
```

Por lo que el mismo desarrollo realizado para acordes mayores puede ser aplicado a acordes menores.

4.3.2. Escala de Zarlino

En el caso de la afinación perfecta (escala de Zarlino), los intervalos no se pueden expresar en cantidad de semitonos, ya que no existe una única definición de semitono. Por ello, la forma más intuitiva de representar cada intervalo es como una razón entre frecuencias. A continuación se exponen estas razones para los intervalos más usados en la música occidental:

Intervalo	Razón de frecuencias
Segunda menor	15/16
Segunda mayor	9/8
Tercera menor	6/5
Tercera mayor	5/4
Cuarta justa	4/3
Cuarta aumentada	10/7
Quinta justa	3/2
Sexta menor	8/5
Sexta mayor	5/3
Séptima menor	7/4
Séptima mayor	15/8
Octava	2/1

[3] el ♭ baja un semitono a la nota

En este caso, la escala se puede construir definiendo el vector:

```
E = [1 15/16 9/8 6/5 5/4 4/3 10/7 3/2 8/5 5/3 7/4 15/8]
```

Aplicando el mismo método que en el caso de la escala temperada, se puede elaborar toda la escala. Para ello, se aplica la fórmula:

$$f_{escala}[i] = 2^{\text{octava}(i)} \cdot E[((i))_{12}] \cdot f_0[0] \qquad i \in [-12, 48] \tag{4.4}$$

Donde `octava(i)` devuelve valores entre -1 y 4, dependiendo de la octava a la que corresponda el valor i. La función en MATLAB sería:

```
octava(i) = floor(i/12)
```

Al igual que sucede en el caso de la escala temperada, el vector de escala `E[i]` puede adecuarse al tipo de material con el que se vaya a trabajar.

Triada mayor o menor

En el caso de trabajar con triadas mayores o menores, se puede aplicar el mismo razonamiento que para la escala temperada, pudiendo simplificar el vector de escala de la siguiente forma:

```
E = [1 6/5 5/4 4/3 3/2 8/5 5/3]
```

Y aplicado la fórmula:

$$f_{escala}[i] = 2^{\text{octava}(i)} \cdot E[((i))_{12}] \cdot f_0[0] \qquad i \in [-7, 28] \tag{4.5}$$

Se obtienen unos valores de $f_{escala}[i]$ que mejoran los resultados para acordes mayores y menores, ya que existe menos riesgo de desviar una cierta $f_0(i)$ a un valor equivocado.

El mismo razonamiento puede ser aplicado a otro tipo de acordes (de séptima[4], disminuidos[5], etc.), aunque no se han desarrollado en este Proyecto.

[4]Un acorde de séptima consta de 4 notas. Estas notas se encuentran a distancia de tercera (mayor o menor), quinta justa y séptima (mayor o menor) con respecto a la tónica.

[5]Un acorde disminudo consta de una tercera menor y una quinta disminuída con respecto a la tónica.

4.3.3. Traslación de las $f_0[i]$ a $f_{escala}[j]$

La etapa de análisis proporciona un array de frecuencias fundamentales $f_0[i]$. A partir de éste se estima la escala en la que estaría trabajando el sistema ($f_{escala}[j]$) según se explica en el apartado anterior. La versión "afinada" de $f_0[i]$ se ha nombrado $f_0^*[i]$.

El procedimiento para ajustar las $f_0[i]$ a las posiciones $f_{escala}[j]$ está basado en un criterio de cercanía en frecuencia (dada en un eje logarítmico). Es decir, si se quiere afinar un determinado $f_0[i]$, se busca la posición de $f_{escala}[j]$ más cercana (menor distancia en semitonos), y se realiza la asignación ${f_0}^*[i] \leftarrow f_{escala}[j]$. De esta forma ${f_0}^*[i]$ contiene el valor "afinado" de $f_0[i]$. Este procedimiento está basado en la hipótesis de que las desafinaciones son leves y normalmente las notas se encuentran relativamente cerca de la posición ideal.

En el algoritmo se mantienen intactas las frecuencias de $f_0[i]$ que estén por encima del valor más alto de $f_{escala}[j]$. Además, si una cierta $f_0[i]$ está por debajo de $f_{escala}[0]$ (valor más bajo de $f_{escala}[j]$), se realiza directamente la asignación ${f_0}^*[i] \leftarrow f_{escala}[0]$. De esta forma el sistema respeta la inharmonicidad que pueden encontrarse en parciales de muy alta frecuencia.

Para aclarar un poco más la idea se plantea el siguiente ejemplo. Supóngase el acorde de La Mayor formado por las notas:

$$[A3, E4, A4, C\#5]$$

Cuyas frecuencias según la afinación perfecta serían:

$$[220, 330, 440, 550]\ Hz$$

Supóngase una versión desafinada que corresponde con las notas:

$$[A3, E4 - 20 \text{ cents}^6, A4 + 35 \text{ cents}, C\#5 + 10 \text{ cents}]$$

Este acorde desafinado corresponde con el siguiente vector de frecuencias fundamentales:

$$f_0[i] = [220, 326{,}20, 448{,}98, 553{,}18]\ Hz$$

Si se sabe que el acorde sólo puede ser mayor o menor en cualquier inversión, y se decide utilizar la afinación perfecta, la $f_{escala}[j]$ estimada sería la siguiente (según se explicó en subapartados anteriores):

[6] Un *cent* es un término común en música, y representa la centésima parte de un semitono.

Nota	$f_{escala}[j]$ **(Hz)**
A3	220 = f0[0]
C4	264
C#4	275
D4	293.3
E4	330
F4	352
F#4	366.6
A4	440
C5	528
C#5	550
D5	586.6
E5	660
...	...

Al aproximar el vector $f_0[i]$ a las posiciones más cercanas de $f_{escala}[j]$, el resultado es:

$$f_0^*[i] = [220, 330, 440, 550]\ Hz$$

Que coincide con la versión afinada del acorde. En este punto, se conocen las notas que deben formar el acorde para que éste suene afinado. En base a esta información, se procede a procesar los diferentes parciales del sonido original para lograr que la sonoridad final sea consonante.

4.4. Cálculo de la nueva estructura armónica

Una vez que se conocen los valores "correctos" de las distintas ${f_0}^*[i]$, es necesario estimar toda la estructura armónica del acorde ideal a lo largo del espectro. Para ello simplemente se generará una especie de rejilla, que contendrá los n primeros armónicos de cada $f_0^*[i]$.

El bloque funcional que genera esta estructura armónica utiliza como parámetros de entrada el array ${f_0}^*[i]$ y un valor n correspondiente al número de armónicos usados en cada nota. Posteriormente se eliminan las frecuencias repetidas y se ordena de menor a mayor. El array que contiene las posiciones de la rejilla puede se ha llamado `rejilla[k]`.

A continuación se muestra un ejemplo para dos notas situadas a 200Hz y 300Hz (distancia de quinta perfecta):

$$f_0 \quad = \quad [200, 300]$$

$$n = 5$$

A partir de lo cual se puede estimar la estructura armónica de cada nota:

```
armonicos(1,:) = [200 400 600 800 1000]
armonicos(2,:) = [300 600 900 1200 1500]
```

Si se unifican todos estos valores en el mismo vector de forma ordenada, se consigue la estructura armónica de las dos notas [A3, E4] (como si de una rejilla se tratara):

```
rejilla = [200 300 400 600 800 900 1000 1200 1500]
```

En la figura 4.7 se muestra gráficamente la estructura armónica calculada para el ejemplo anterior.

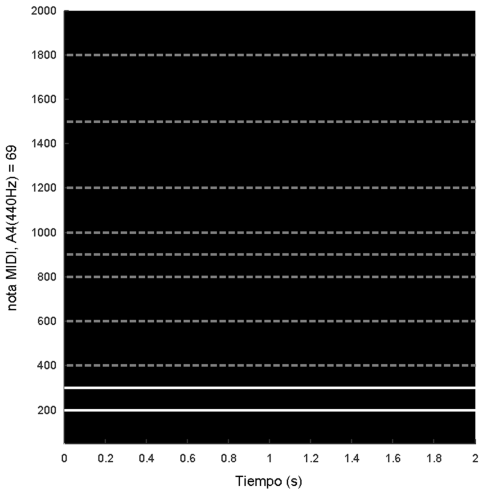


Figura 4.7: Posición del array f0[i] (blanco) y del array rejilla[k] (gris discontinua).

Una vez que se dispone de la estructura armónica del acorde, es necesario desplazar los parciales del sonido original a estas posiciones para conseguir que la sonoridad final sea "consonante".

4.5. Traslación de los parciales a la nueva estructura

Como ya se ha explicado anteriormente, la nueva estructura armónica funciona como una rejilla que limita las posiciones de los parciales. Para ajustar un determinado sonido a la nueva estructura, simplemente se aproxima la frecuencia de cada

parcial a su posición más cercana de la rejilla. En la figura 4.8 se muestra el funcionamiento de este sistema.

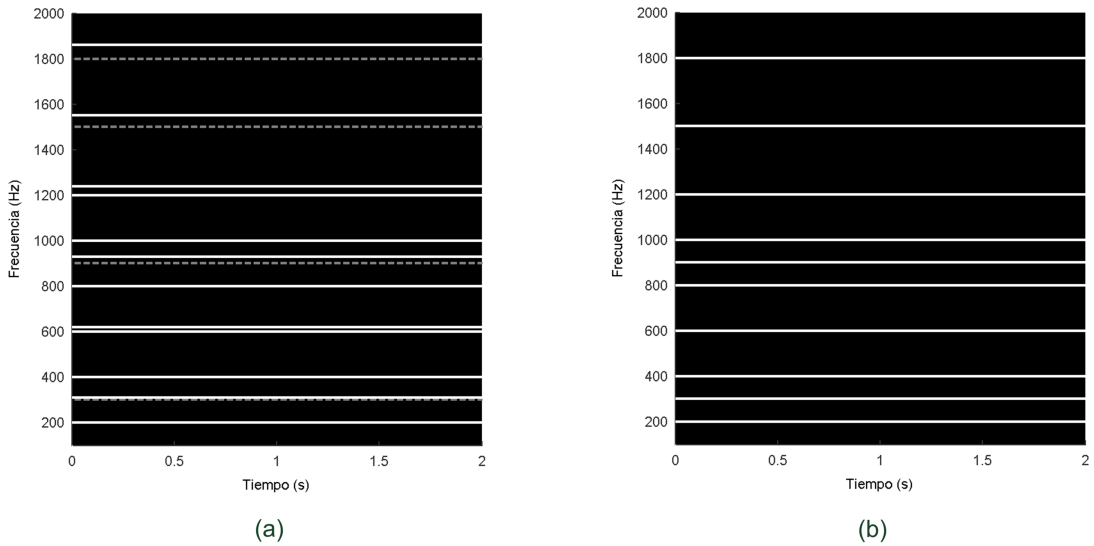


Figura 4.8: Sonido de entrada: 2 tonos de 200Hz y 310Hz. **(a)** Parciales del sonido de entrada (blanco) y posiciones del vector calculado rejilla[k] (gris discontínua) **(b)** Parciales después de la traslación

En el caso de parciales que superen el valor máximo de `rejilla[i]`, no sufrirán ninguna modificación. La razón es que estos armónicos de alta frecuencia no contribuyen fuertemente a la percepción de la disonancia, ya que suelen tener poca potencia y aportan más bien características tímbricas. Además, en sonidos metálicos suele existir una gran inharmonicidad en los armónicos agudos. Si se procesaran se forzaría un timbre excesivamente armónico que puede restar naturalidad al sonido original.

4.6. Ajuste a 440Hz

Una vez que se ha procesado el sonido original, es posible que la afinación resultante de todo el acorde no esté referenciada a 440Hz. Esto puede deberse a dos factores:

- La afinación inicial del acorde no estaba referenciada a 440Hz

- La nota más grave de las f_0s estimadas estaba desafinada con respecto al 440Hz, aunque el acorde completo sí lo estuviera.

En cualquier caso, el sistema siempre ajustará a 440Hz el resultado final, aunque el acorde inicial estuviese referenciado a otra afinación. En ocasiones puede no ser deseable, pero se considera que en la mayoría de las ocasiones sí lo será.

Para conseguir esto se aplica un procedimiento simple pero ingenioso, el cual ha sido tomado de [12]. La hipótesis es que los parciales de un acorde afinado a 440Hz deberían caer en posiciones de la escala temperada tomando el A4 como 440Hz. Esto no es totalmente cierto en la realidad, ya que las relaciones armónicas siguen intervalos perfectos, no temperados. Sin embargo, en promedio esta afinación se puede considerar muy aproximada.

El procedimiento consiste en calcular la desviación media de todos los parciales con respecto a la escala temperada ($A4 = 440Hz$). Esta desviación ha de ser medida de forma logarítmica (en semitonos), considerando así el rango de desviación $[-0{,}5, 0{,}5]$ semitonos. Esto puede expresarse en Matlab de la siguiente forma:

```
desviaciones_stonos=log2(fparciales/440)*12-round(log2(fparciales/440)*12)
```

Posteriormente al vector `desviaciones_stonos` se usa para calcular la desviación media de los parciales.

En este Proyecto se ha introducido además una mejora para evitar problemas derivados del análisis. A la hora de calcular la media, se pondera la desviación de cada parcial con el producto $magnitud \cdot longitud$, evitando así que parciales espúreos afecten de la misma forma que parciales estables de larga duración. Con esta media ponderada, el resultado en general ofrece muy buenos resultados.

Capítulo 5

Subsistema de Síntesis

Una vez que los parciales han sido procesados, éstos se usan para sintetizar una nueva versión de la señal. El conjunto de funciones que permiten la síntesis están englobadas dentro del Subsistema de Síntesis, del cual se muestra un diagrama de bloques en el apartado 5.1.

El procedimiento de síntesis se lleva a cabo sobre ventanas temporales sucesivas dentro de la señal, las cuales se superpondrán posteriormente para formar la señal completa. Para ello, cada parcial ha de segmentarse antes de proceder al proceso de síntesis. Esta segmentación se explica en el apartado 5.2).

A partir de esta información, se sintetiza la componente sinusoidal tal y como se explica en el apartado 5.3. Para ello se sigue el esquema superposición-suma (*overlap-adding*). Este procedimiento se expone más detalladamente en el subapartado 5.3.3.

El último paso de la síntesis es añadir la componente residual que fue extraída en la fase de análisis (subapartado 5.4).

5.1. Diagrama de bloques

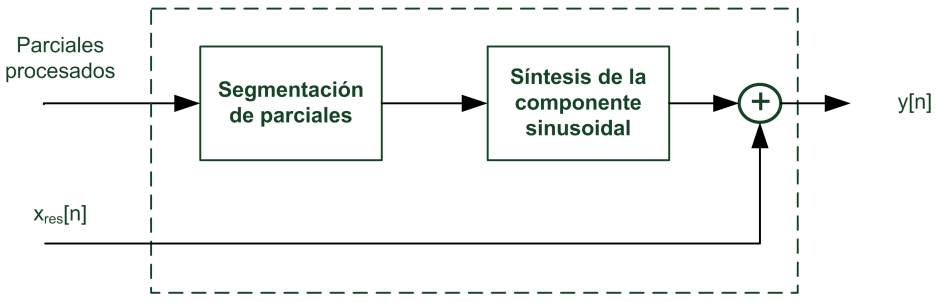


Figura 5.1: Diagrama de bloques del Subsistema de Síntesis

El Subsistema de Síntesis es mucho más sencillo que el resto de subsistemas. Sin embargo, funcional y conceptualmente es muy importante, ya que es la interfaz de salida del Sistema.

5.2. Segmentación de parciales

Este bloque funcional es sencillo, ya que sólamente representa la misma información sinusoidal de forma diferente. El objetivo aquí es eliminar el seguimiento temporal de los parciales, obteniendo la información ventana a ventana como se explicaba en el subapartado 3.4.3. Para ello, cada parcial se segmenta en L partes, y posteriormente se agrupan todos los picos espectrales pertenecientes a la misma ventana. En la figura 5.2 se muestra gráficamente este proceso.

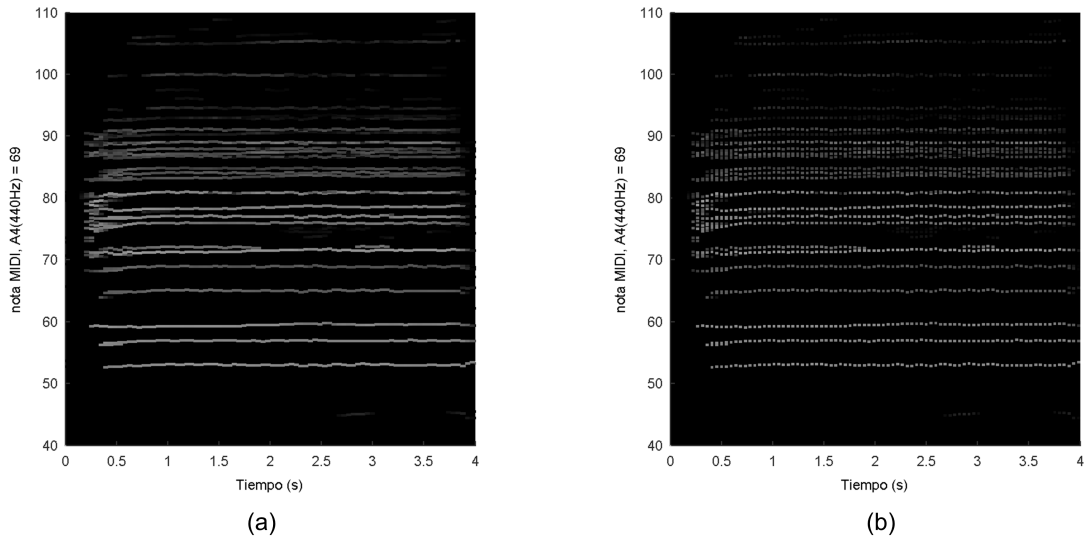


Figura 5.2: **Parciales correspondientes a un cuarteto vocal. (a)** Parciales antes de la segmentación **(b)** Parciales tras la segmentación que produce la funcion `Parciales2Ventanas()`|

5.3. Síntesis de la componente sinusoidal

El bloque más complejo de este subsistema es la síntesis de la nueva componente sinusoidal. Este es un proceso que se realiza en dos ocasiones a lo largo de todo el Sistema: en el análisis (para extraer la componente residual) y en la síntesis (para

sintetizar la señal procesada).

La información de entrada a esta función es un array de ventanas, contiendo cada una de ellas un array de picos (estructura explicada en el subapartado 3.4.3). El procedimiento consiste en sintetizar cada ventana partiendo sólo de la información sinusoidal para finalmente sumarlas con un factor de superposición del 75 %.

Esta estrategia se podría haber abordado directamente en el dominio del tiempo, realizando una síntesis aditiva de sinusoides con los parámetros extraídos en el análisis. Sin embargo la complejidad computacional de este proceso sería muy alta, y por ello se ha optado por trabajar en el dominio de la frecuencia.

La síntesis de una sinusoide en el dominio de la frecuencia resulta trivial si se trabaja con enventanado Blackman-Harris 92dB, ya que los lóbulos secundarios son despreciables. Por tanto, el procedimiento para generar la componente sinusoidal de una ventana es generar un único lóbulo Blackman-Harris para cada sinusoide (con su magnitud, frecuencia y fase correspondiente).

5.3.1. Generación del lóbulo principal en frecuencia de la ventana Blackman-Harris 92dB

La expresión en el dominio del tiempo de la ventana Blackman-Harris de 92dB es la siguiente:

$$w(n) = a_0 + a_1 cos(\frac{2\pi n}{N}) + a_2 cos(\frac{4\pi n}{N}) + a_3 cos(\frac{6\pi n}{N}) \tag{5.1}$$

Donde $a_0 = 0{,}35875$, $a_1 = -0{,}48829$, $a_2 = 0{,}14128$ y $a_3 = -0{,}01168$. La transformada de Fourier es por tanto:

$$W_{BH}(\omega) = \sum_{k=-3}^{3} a_{|k|} W_R(\omega + k\Omega_M) \tag{5.2}$$

Donde $W_R(\omega) = \frac{sin(M\omega/2)}{sin(\omega/2)}$ (transformada de una ventana rectangular centrada en el origen de longitud M muestras) y $\Omega_M = \frac{2\pi}{M}$.

Para la síntesis se cada ventana temporal, se utilizan ventanas de tamaño N_s (tamaño de la ventana de síntesis). Al contrario de lo que sucedía en el análisis, aquí no se utilizan valores diferentes para M y N, sino que se supone que $M = N_s$ (no hay *zero-padding*). Bajo esta circunstancia, el lóbulo principal de una ventana Blackman-Harris se puede aproximar bastante bien con 9 muestras significativas (recuérdese la anchura del lóbulo principal $K = 8$), calculadas alrededor del valor de frecuencia de

la sinusoide.

5.3.2. Síntesis de una ventana temporal en el dominio frecuencial

Una vez que se dispone del espectro deseado, se realiza la transformada de Fourier inversa (IFFT) de tamaño N_s. El resultado de esta operación es una suma de sinusoides dentro de una ventana de tamaño N_s multiplicadas por una ventana Blackman-Harris, también de tamaño N_s. Este procedimiento se ejemplifica gráficamente en la figura 5.3.

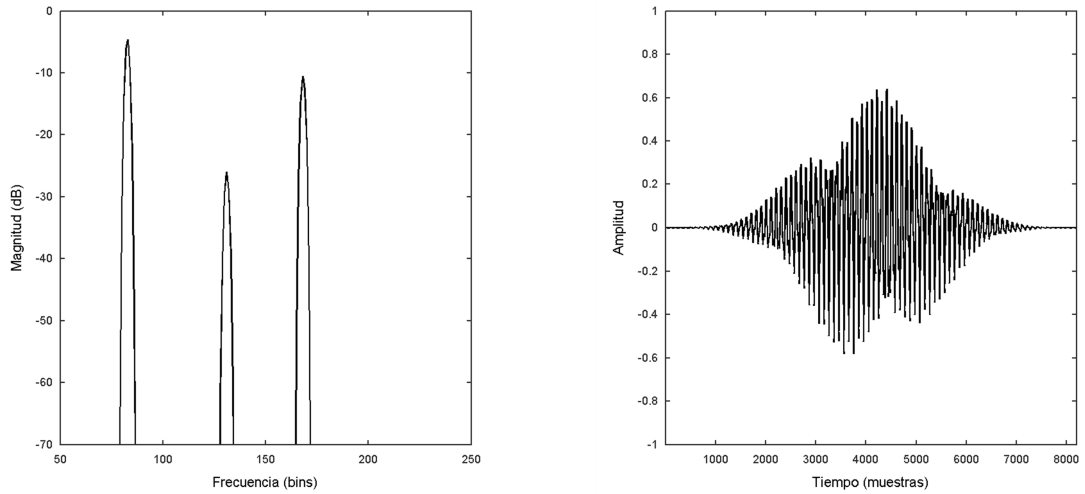


Figura 5.3: **(a)** Espectro generado a partir de la superposición de varios lóbulos principales Blackman-Harris 92dB. **(b)** Resultado de la Transformada Inversa de Fourier (IFFT). Se observa que el enventanado en el dominio del tiempo corresponde efectivamente a Blackman-Harris 92dB

Es importante tener en cuenta lo explicado en el apartado 3.3.2 sobre el efecto de la fase. Si a los lóbulos se les aplica el valor de fase directamente extraído del análisis, la ventana resultante de la IFFT estará desplazada $(M-1)/2 = (N_s-1)/2$. Es necesario aplicar posteriormente otro deplazamiento de $(N_s-1)/2$ para obtener la ventana tal de la figura 5.3.

5.3.3. Proceso Superposición-Suma

En este subapartado se explicará el procedimiento Superposición-Suma (del inglés *overlap-adding* u *OLA*). A través de este proceso se consigue "invertir" la transformación STFT, sintetizando la señal temporal completa a partir de su espectrograma. Existen ciertos detalles a tener en cuenta en este proceso, los cuales han sido consultados en las referencias [44] y [37]. Este proceso se esquematiza gráficamente en la figura 5.4.

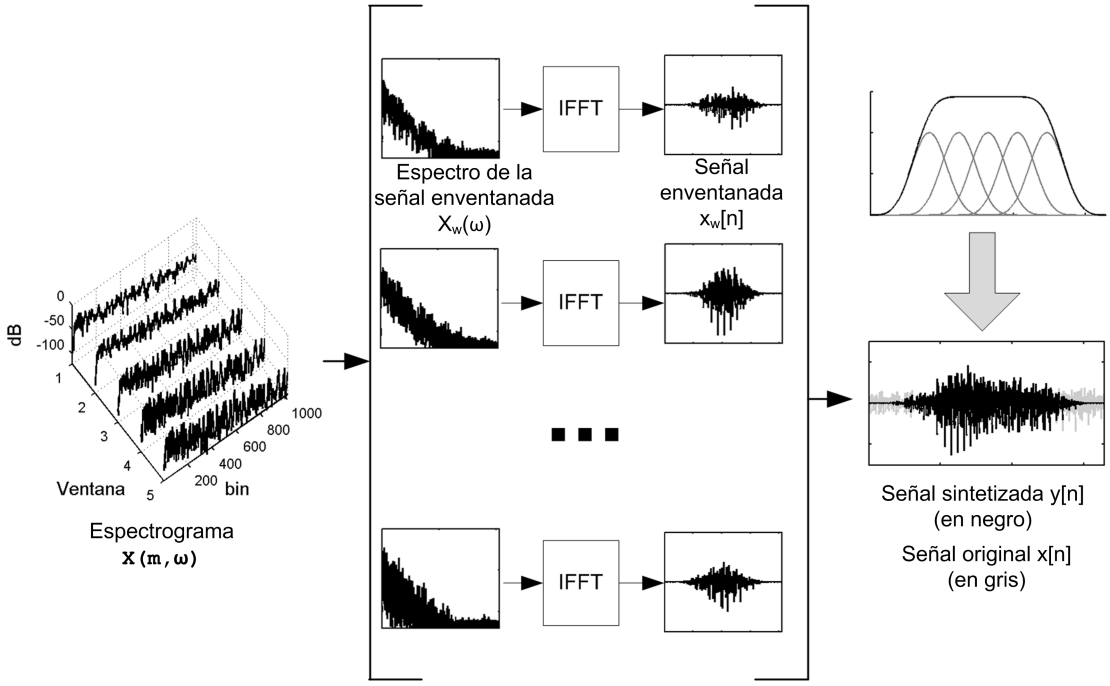


Figura 5.4: Proceso de superposición-suma utilizando una ventana Blackman-Harris de 92dB y un factor de superposición del 75 %. En la gráfica final, se ha superpuesto la reconstrucción en negro sobre el audio original en gris. Se observa que existe cierta amplificación (factor 1.435), y un suavizado en los bordes debido a las primeras y las últimas ventanas.

Es posible conseguir una reconstrucción exacta de la componente sinusoidal si se utilizan ventanas cuya superposición de lugar a un valor constante. Para cada tipo de ventana es necesario escoger el factor de superposición adecuado. En general los factores de superposición más habituales son 50 % o 75 % (tamaño de salto $H = N_s/2$ o $H = N_s/4$). No obstante, para cierto tipo de ventanas no es posible usar un factor de superposición del 50 %, como por ejemplo sucede con la Blackman-Harris 92dB. En la figura 5.5 se muestra este proceso para varios tipos de ventanas.

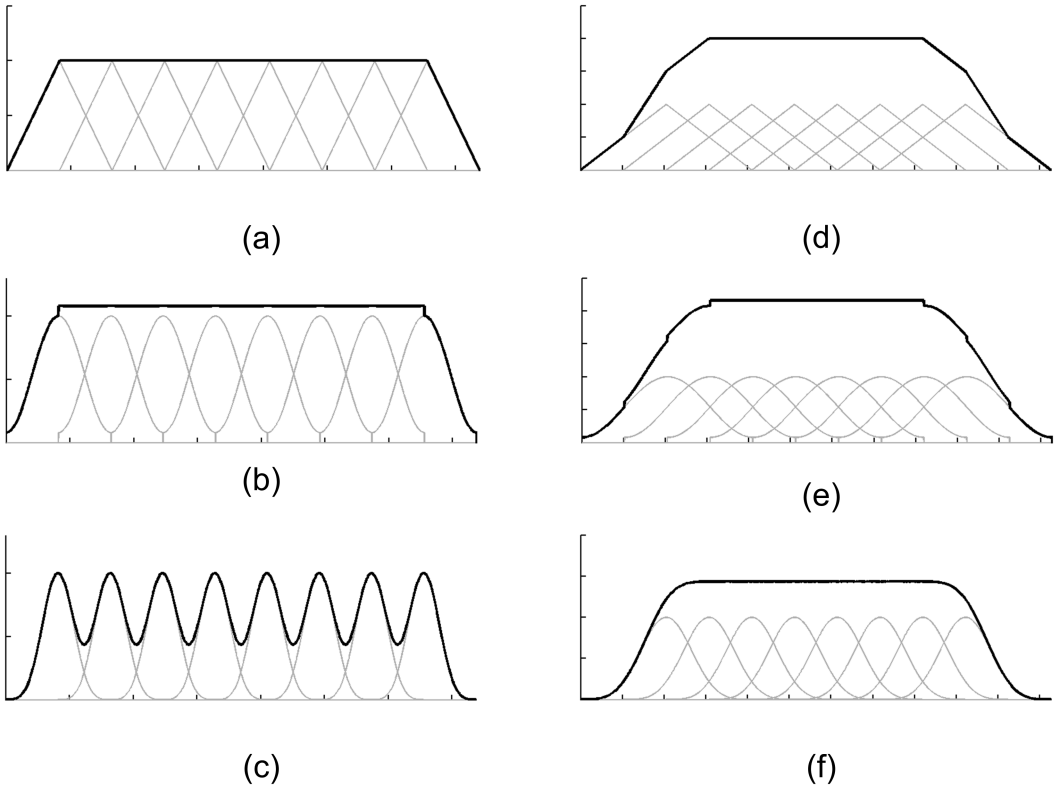


Figura 5.5: (a) Ventana triangular con factor de superposición del 50 % (b) Ventana Hamming / 50 % (c) Ventana Blackman-Harris 92dB / 50 % (factor de superposición incorrecto) (d) Ventana triangular / 25 % (e) Ventana Hamming / 25 % (f) Ventana Blackman-Harris / 25 %

Analíticamente, una ventana tiene la propiedad de una superposición-suma constante (comúnmente llamada *COLA*, de *Constant Overlap-Adding*) si para un determinado valor de R se cumple que:

$$\sum_{m=-\infty}^{\infty} w(n - mR) = \lambda$$

Siendo λ un valor constante independiente de n.

En el Subsistema de Síntesis se ha utilizado una ventana Blackman-Harris 92dB de tamaño N_s con un tamaño de salto $H = N_s/4$ (factor de superposición 75 %). En la figura 5.6 se muestra gráficamente cómo se lleva a cabo este proceso en el caso de una señal real.

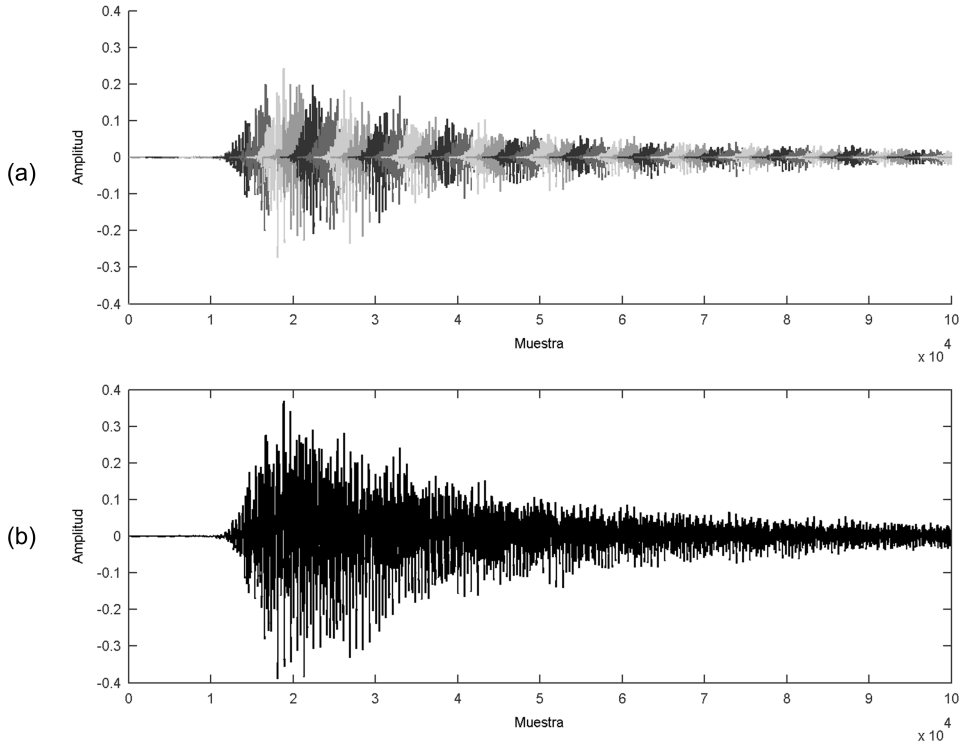


Figura 5.6: Sonido: Acorde La Mayor de guitarra acústica **(a)** Secuencia de ventanas solapándose con un factor de superposición del 75 % **(b)** Resultado final de la suma de todas las ventanas

5.4. Adición de la componente residual

Por último, a la componente sinusoidal sintetizada se le añade la componente residual extraída en en análisis. Esta componente residual suele contener las componentes ruidosas y los ataques, y no se le aplica ningún procesado a lo largo del Sistema. Analíticamente se puede expresar esta operación de la siguiente forma:

$$x[n] = x_{sin}[n] + x_{res}[n] \tag{5.3}$$
$$y[n] = y_{sin}[n] + x_{res}[n] \tag{5.4}$$

En la figura 5.4 se muestra gráficamente qué efecto tiene sobre una señal real la adición de la componente residual.

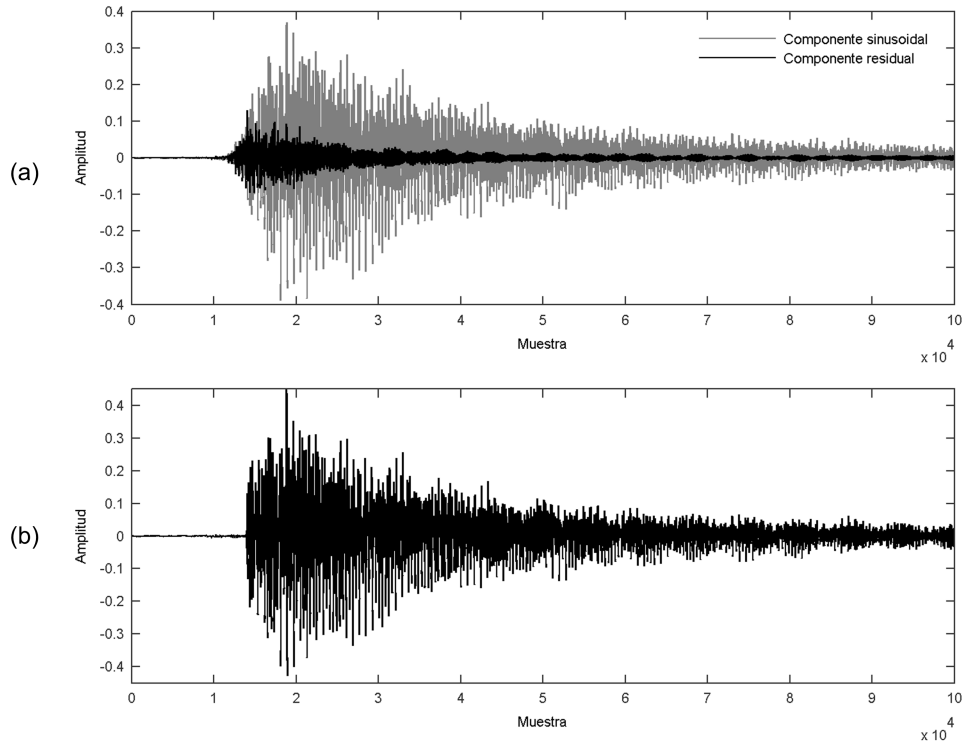


Figura 5.7: Sonido: Acorde La Mayor de guitarra acústica. **(a)** Representación separada de la componente residual y la componente sinusoidal **(b)** Suma de ambas componentes para generar la señal final

Capítulo 6

Resultados

En este capítulo se exponen los resultados obtenidos tras la evaluación del sistema desarrollado. Para ello, en primer lugar se exponen las consideraciones generales que ayudan a entender el procedimiento de evaluación (parámetros utilizados y metodología), y posteriormente se exponen los resultados de los experimentos llevados a cabo para la evaluación. Estos experimentos son:

1. Experimento 1: Señales sintéticas
2. Experimento 2: Guitarra acústica
3. Experimento 3: Cuartetos de cuerda frotada e instrumentos de viento

En cada experimento se han tomado varias muestras de acordes desafinados, se han procesado con el sistema y posteriormente se han evaluado las diferencias entre la entrada y la salida. Además, se ha procesado cada muestra con el software comercial más relevante: *Melodyne Editor*. En total se han tomado 8 acordes diferentes, de los cuales se han evaluado 3 variantes (original, procesado con el Sistema, procesado con Melodyne), resultando por tanto 24 variantes en total. Esta evaluación se ha realizado por triangulación, combinando un análisis subjetivo y un análisis objetivo de los resultados.

El análisis subjetivo se ha realizado mediante un cuestionario debidamente diseñado para conocer la opinión general del usuario. Este procedimiento está inspirado en [27].

El análisis objetivo está basado en la herramienta online *SRA: Spectral and Roughness Analysis of sound signals*, basada en el algoritmo de Vassilakis [49]. Esta herramienta cuantifica la "aspereza" (roughness) del sonido que se analiza, utilizando las ideas expuestas por Plomp en 1965 [40]. Este algoritmo no contempla aspectos musicales como el tipo de acorde o el gusto personal, ya que se centra únicamente en aspectos perceptuales. En cualquier caso, a pesar de la ausencia de aspectos musicales de esta evaluación, se ha considerado conveniente este análisis ya que aporta una perspectiva interesante que complementa la evaluación subjetiva.

6.1. Consideraciones generales

En este apartado se especifican los parámetros del sistema utilizados en los experimentos, además de detallarse los procedimientos de evaluación utilizados.

6.1.1. Parámetros utilizados en los experimentos

En el sistema desarrollado existen numerosos parámetros que pueden ser modificados para adecuarlo a un tipo concreto de señal. Sin embargo, se ha conseguido un conjunto de parámetros que ofrece resultados aceptables para una gama amplia de señales. Sin duda este conjunto podría ser refinado para obtener mejores resultados en casos concretos, pero requeriría una mayor intervención del usuario y es algo que se ha intentado evitar.

El conjunto de parámetros utilizados es el siguiente:

1. **STFT:** Tamaño de ventana (M) = 8001, Tamaño de FFT (N) = 8192, Tipo de ventana = Blackman-Harris 92dB, Factor de superposición = 75 %
2. **Estimación de sinusoides:** Número máximo de sinusoides detectadas = 30 (50 en el caso del cuarteto de cuerda frotada), Umbral de detección = -80dB
3. **Seguimiento temporal de sinusoides:** Se consideran el mismo parcial si las sinusoides varían en menos de 0.2 semitonos, en menos de 20dB y no están más distantes de 700ms.
4. **Eliminación de parciales demasiado cortos:** Se eliminan los parciales por debajo de 500ms
5. **Máximo número de $f_0 s$ estimadas:** Cuando la introducción de $f_0 s$ es manual, el número de $f_0 s$ estimadas puede variar en función del interés del usuario. Cuando se trabaja con el sistema automático, utilizan como máximo 5 $f_0 s$ diferentes.
6. **Escala utilizada** En general se ha utilizado la escala cromática Temperada, aunque en el caso de triadas mayores con sonidos sintéticos o instrumentos de cuerda frotada se ha utilizado la escala de Zarlino.
7. **Posiciones permitidas en la nueva estructura armónica:** 30 armónicos de cada una de las f_0 estimadas.

6.1.2. Análisis subjetivo y objetivo de los resultados

Grupo de sujetos

Los sujetos encuestados son alumnos de grado profesional del Conservatorio Profesional Manuel Carra de Málaga. Se pasó el cuestionario en 3 aulas diferentes, abarcando un total de 31 alumnos. No hubo ningún tipo de preferencia de género o edad, al considerarse que no son parámetros que puedan sesgar notablemente los resultados. El elemento común que comparten todos los sujetos es más de 7 años de formación musical en conservatorio, y por tanto experiencia suficiente para poder hacer un juicio crítico sobre los resultados.

En la tabla 6.1 se muestra la cantidad de sujetos de cada género.

Género	Número de sujetos
Masculino	16
Femenino	15

Tabla 6.1: Cantidad de sujetos de cada género.

En la figura 6.1 se muestra un histograma de las edades de los sujetos encuestados. La mayoría de los usuarios ellos bastante jóvenes (20 años o menos), aunque algunos sujetos tenían mayor edad.

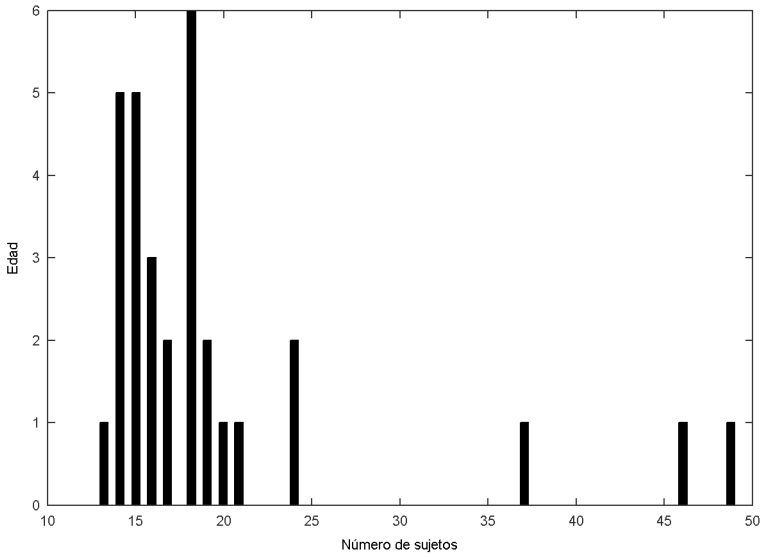


Figura 6.1: Histograma de edades de los sujetos.

En la figura 6.2 se muestra un histograma de años de formación musical de los sujetos. Todos ellos tienen 7 años de formación o más.

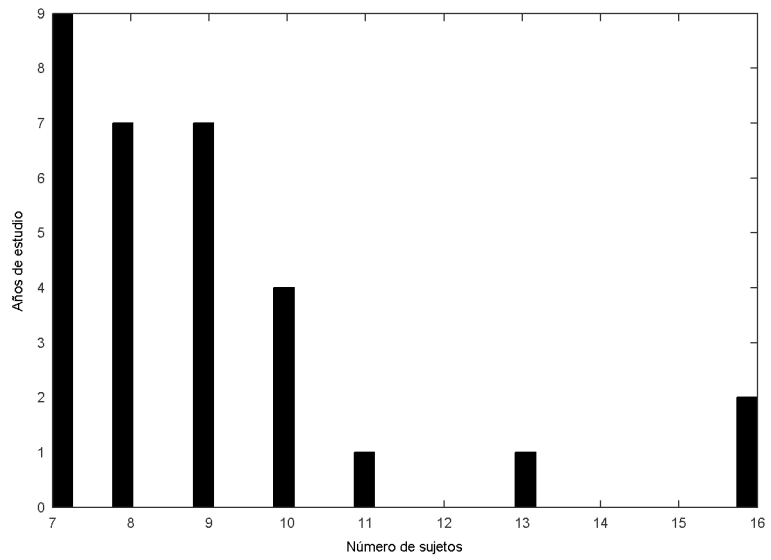


Figura 6.2: Histograma de años de formación musical de los sujetos.

En la figura 6.3 se muestra la cantidad de sujetos que estudiaban cada instrumento. Se observa que existen instrumentos de todo tipo, algo interesante teniendo en cuenta que este es un factor que podría sesgar los resultados.

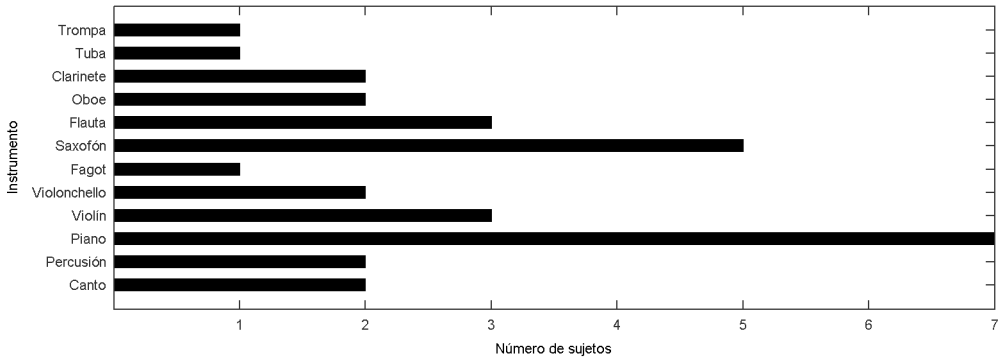


Figura 6.3: Histograma de instrumentos que han estudiado los sujetos.

Cuestionario

Para la evaluación de resultados, se les ha pedido a los usuarios cuantificar una serie de sonidos en base a dos criterios: consonancia y naturalidad. Los tres experimentos llevados a cabo comprenden 8 muestras diferentes, cada una de ellas con 3 variantes. Para cada uno de las 8 muestras, a los usuarios, se les ha pedido responder este cuestionario:

1. **Sonido x.A:** Califícalo según los siguientes adjetivos:
   - Disonante - `1 2 3 4 5 6 7 8 9 10` - Consonante
   - Antinatural - `1 2 3 4 5 6 7 8 9 10` - Natural
2. **Sonido x.B:** Califícalo según los siguientes adjetivos:
   - Disonante - `1 2 3 4 5 6 7 8 9 10` - Consonante
   - Antinatural - `1 2 3 4 5 6 7 8 9 10` - Natural
3. **Sonido x.C:** Califícalo según los siguientes adjetivos:
   - Disonante - `1 2 3 4 5 6 7 8 9 10` - Consonante
   - Antinatural - `1 2 3 4 5 6 7 8 9 10` - Natural
4. ¿Cuál consideras más apropiado en un contexto musical?
   - A
   - B
   - C

Las variantes A, B y C representan respectivamente el sonido sin procesar, procesado con el sistema desarrollado y procesado con Melodyne Editor. No obstante, los usuarios no sabían a cuál correspondía cada uno de ellos.

Evaluación estadística de los resultados

Para cada variante, se ha obtenido:

- La media de las valoraciones de los usuarios sobre el nivel de consonancia.
- La desviación típica de las valoraciones de los usuarios sobre el nivel de consonancia.
- La media de las valoraciones de los usuarios sobre el nivel de naturalidad.

- La desviación típica de las valoraciones de los usuarios sobre el nivel de naturalidad.
- El porcentaje de veces que dicha variante ha sido escogida como el mejor resultado de entre las tres variantes propuestas.

La comparación se ha llevado a cabo observando la diferencia entre las valoraciones medias de consonancia y naturalidad. Para asegurar que dicha diferencia entre las medias no es fruto de la casualidad, se ha llevado a cabo la prueba *t-Student* para cada comparación. Esta prueba se aplica cuando la población se asume ser normal pero el tamaño muestral es demasiado pequeño como para que el estadístico en el que está basada la inferencia esté normalmente distribuido, utilizándose una estimación de la desviación típica en lugar del valor real.

Aunque los detalles estadísticos de la prueba t-Student son complejos, Matlab permite realizar la prueba t-Student adecuada para este caso a través de la función `ttest(X,Y)`, donde X e Y son los dos vectores de muestras que se están comparando. Esta función devuelve un valor p, el cual indica la probabilidad de que las medias poblacionales sean iguales ($\mu_X = \mu_Y$), aunque las medias muestrales no lo sean ($\overline{X} \neq \overline{Y}$). En general, se considera que un valor $p < 0{,}05$ es suficiente para considerar que la diferencia de las medias muestrales ($\overline{X} - \overline{Y}$) es debido a una diferencia en las medias poblacionales ($\mu_X - \mu_Y$).

Análisis objetivo

Varios artículos publicados proponen un modelo perceptual para la cuantificación del *roughness* ([7, 20, 21, 22, 23, 24, 36, 38, 41, 48, 49, 51]). La palabra *roughness* significa "aspereza", y está directamente relacionado con lo agradable o desagradable que resulta un sonido desde un punto de vista puramente perceptual.

En general, todos estos algoritmos se basan en el siguiente esquema:

1. Obtener la posición y la amplitud de los picos armónicos más importantes.
2. Comparar todos ellos entre sí y cuantificar la aspereza que produce cada par de armónicos.
3. Realizar una suma de la aspereza de cada par de armónicos.

La cuantificación de la aspereza de dos armónicos depende principalmente de la separación entre ellos, de las amplitudes relativas y de las amplitudes absolutas. Sin embargo, otros algoritmos tienen en cuenta otros aspectos como la fase, la posición absoluta, etc. Además, en la suma final puede realizarse algún tipo de ponderación

en función de las características propias del par de armónicos que se estén considerando. En estos detalles se diferencian unos algoritmos de otros.

El algoritmo propuesto en [49] por Vassilakis ha sido implementado por el autor en una herramienta web (ver figura 6.4), y éste es el que se ha utilizado para la cuantificación objetiva de la disonancia. Los detalles de este algoritmo pueden ser encontrados en [49], y se trata de un algoritmo que no introduce novedades muy importantes sobre el esquema anteriormente expuesto. Esta herramienta proporciona el nivel de aspereza sonora estimado a lo largo de toda la señal.

Esta utilidad puede ser encontrada en la siguiente dirección:

http://musicalgorithms.ewu.edu/algorithms/Roughness.html [49]

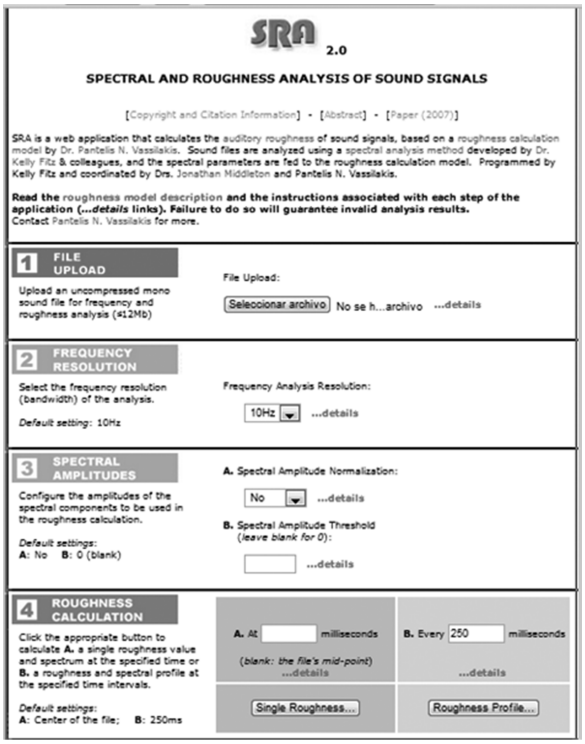


Figura 6.4: *SRA: Spectral and Roughness Analysis of sound signals* online application

6.2. Experimento 1: Señales sintéticas

En este experimento se pretende poner a prueba el sistema en un caso sencillo, que represente cláramente el funcionamiento general del sistema. Para ello se han generado artificalmente tres acordes, cada uno formado por 4 notas. Cada nota está compuesta por 6 armónicos con una distribución paso-bajo. Estos sonidos son totalmente estables en frecuencia, y tienen una envolvente en amplitud que decrece exponencialmente (como suele suceder en los sonidos reales).

Los acordes generados constan de las siguientes notas:

- Acorde sintético 1 (Acorde perfecto mayor desafinado): C4, E4 + 11 cents[1], G4 - 21 cents, C5 + 30 cents
- Acorde sintético 2 (Acorde perfecto menor desafinado): C4, E♭4 + 13 cents, G4 + 17 cents, C5 - 32 cents
- Acorde sintético 3 (Acorde mayor en primera inversión desafinado): E4, G4 + 16 cents, C5 - 30 cents, E5 + 17 cents

Estos acordes tienen sonoridades claramente disonantes, ya que existen fuertes desviaciones de cada nota con respecto a su posición en la escala temperada. A pesar de ello, el Sistema no encuentra dificultades para trabajar con los parciales, ya que se encuentra en un caso realmente favorable (timbre estable y armónico, ausencia de parciales espúreos, armonía sencilla, etc.). A modo de ejemplo, en la figura 6.5 se muestra la componente sinusoidal en cada etapa del Sistema para el acorde sintético 1 (acorde perfecto mayor).

El Sistema ha sido configurado para ajustar a la escala de afinación perfecta (Zarlino) los acordes sintéticos 1 y 3, y a la escala temperada el acorde sintético 2. La razón es que los acordes mayores son especialmente consonantes dentro de la afinación perfecta debido a la coincidencia exacta de parciales. La detección de f_0 ha sido establecida de forma automática con éxito. Cada etapa del sistema, en general, ofrece los resultados esperados. Tan sólo existe cierta dificultad en hacer el seguimiento temporal de armónicos en los que existen batidos, pero tras eliminar los armónicos espúreos el resultado son parciales limpios libres de oscilaciones indeseadas.

[1] Un *cent* es la centésima parte de un semitono.

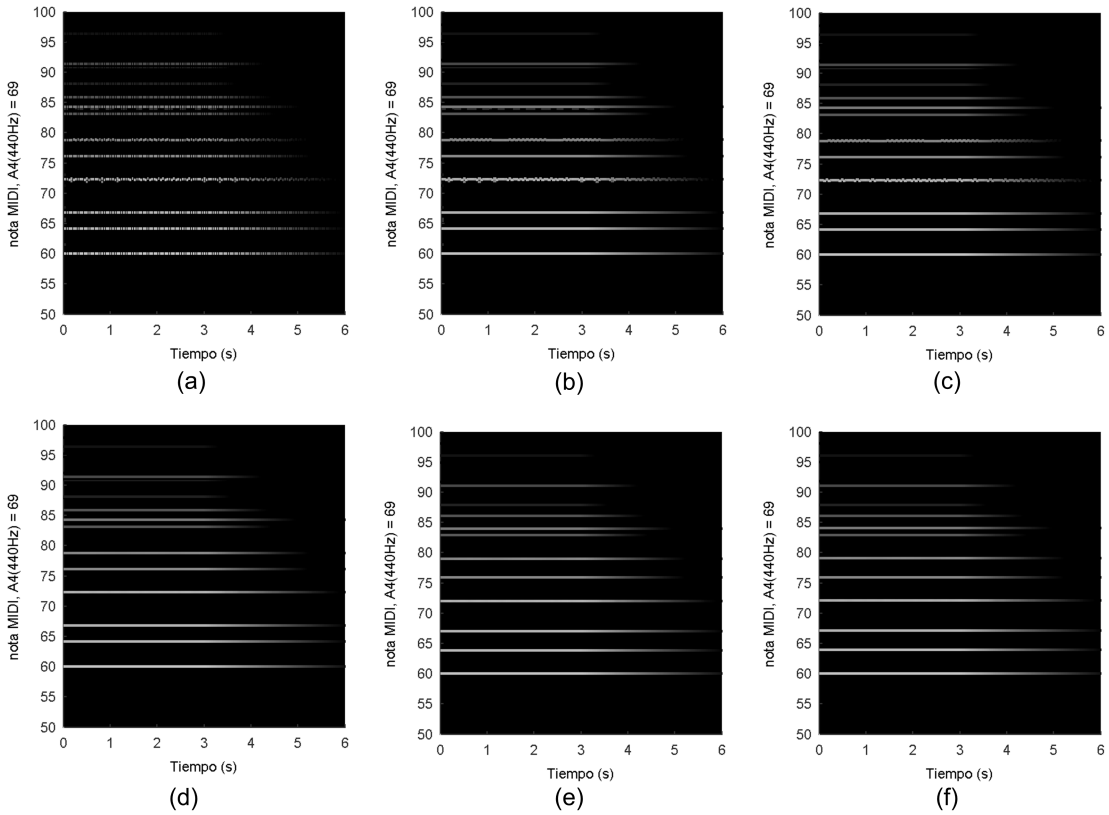


Figura 6.5: Acorde Sintético 1 (Notas: [C4, E4, G4, C5] , Números midi: [60, 64, 67, 72]): **(a)** Resultado de la estimación sinusoidal **(b)** Seguimiento temporal de los parciales **(c)** Eliminación de parciales de corta duración **(d)** Estabilización de parciales **(e)** Ajuste a la nueva estructura armónica **(f)** Traslación a la afinación 440Hz.

6.2.1. Análisis subjetivo de los resultados

Cada acorde ha sido evaluado en sus tres variantes:

- A) Original
- B) Procesado con el Sistema
- C) Procesado con *Melodyne Editor*

Obteniendo en total 9 muestras sintéticas a analizar. Los resultados obtenidos en los cuestionarios se muestran en la tabla 6.2.

Variante	**Consonancia subj. media y desv. típica**	**Naturalidad subj. media y desv. típica**	**Escogido como mejor resultado**
Sintético 1.A	mean(c) = 3,48 std(c) = 1,48	mean(n) = 3,03 std(n) = 1,68	3.2 %
Sintético 1.B	**mean(c)=6.64 std(c)=2.05**	**mean(n)=5.03 std(n)=2.41**	**77.4 %**
Sintético 1.C	mean(c) = 5,48 std(c) = 1,80	mean(n) = 4,67 std(n) = 2,38	19.35 %
Sintético 2.A	mean(c) = 2,67 std(c) = 1,30	mean(n) = 2,90 std(n) = 2,07	6.45 %
Sintético 2.B	**mean(c)=5.35 std(c)=2.25**	**mean(n)=4.35 std(n)=2.41**	**74.2 %**
Sintético 2.C	mean(c) = 3,96 std(c) = 1,87	mean(n) = 3,51 std(n) = 2,03	19.3 %
Sintético 3.A	mean(c) = 3,22 std(c) = 1,81	mean(n) = 3,51 std(n) = 2,17	3.23 %
Sintético 3.B	**mean(c)=6.19 std(c)=2.42**	**mean(n)=4.64 std(n)=2.48**	**45.16 %**
Sintético 3.C	mean(c) = 6,35 std(c) = 1,90	mean(n) = 5,06 std(n) = 2,32	51.6 %

Tabla 6.2: Resultados de los cuestionarios en el caso de acordes sintéticos

A partir de los resultados de los cuestionarios, se extraen una serie de conclusiones interesantes sobre el experimento y sobre el funcionamiento del Sistema:

- La mejora producida entre el sonido sin procesar (variante A) y el sonido procesado con el Sistema (variante B) es, en todos los casos, muy notable. Existen casi 3 puntos de diferencia en la valoración media de consonancia en una escala del 1 al 10. La prueba t-Student ofrece un valor $p < 0{,}01\,\%$ al comparar la variante A con la variante B.

- Al comparar los resultados que ofrece el Sistema desarrollado con los que ofrece *Melodyne Editor*, se observa que en los casos 1 y 2 la variante B se muestra significativamente mejor que la variante C (valor $p < 1\,\%$ al realizar el test t-Student). En el caso 3, la prueba t-Student determina que no es posible afirmar que las diferencias estadísticas no sean fruto de la casualidad ($p > 5\,\%$).

- Resulta interesante comprobar cómo la naturalidad suele estar bastante correlada con la consonancia percibida. Aunque no es algo que se haya estudiado

con detalle, es un fenómeno que se ha observado e invita a la reflexión sobre el significado de "consonante" que manejan los músicos habitualmente.

- Se ha observado que la percepción de consonancia es pequeña incluso en el caso de sonidos sintéticos perfectamente afinados. Este hecho parece indicar los timbres reales suenan "mejor" para los sujetos que los sintéticos, debido que influye positivamente la familiaridad tímbrica.

- La conclusión global es que el sistema desarrollado funciona de forma ideal con este tipo de sonidos, consiguiendo exactamente los resultados esperados. Sin embargo, el caso de sonidos sintéticos es demasiado favorable, y es necesario poner a prueba el Sistema con sonidos reales.

6.2.2. Análisis objetivo de los resultados

La herramienta propuesta en [49] para la cuantificación objetiva de la "aspereza sonora" ofrecen otro interesante punto de vista para valorar el buen funciomiento del Sistema (análisis objetivo). Esta herramienta ofrece como resultado la evolución temporal de la aspereza (relacionado directamente con la disonancia). En las figuras 6.6, 6.7 y 6.8 se muestran los resultados obtenidos en cada caso.

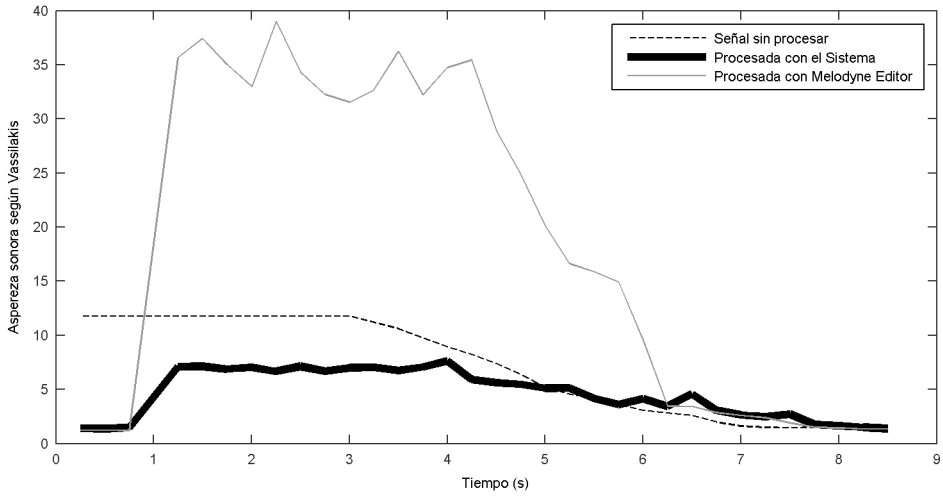


Figura 6.6: Resultado del análisis del acorde Sintético 1 con la herramienta online *SRA: Spectral and Roughness Analysis of sound signals*, basada en el algoritmo de Vassilakis [49].

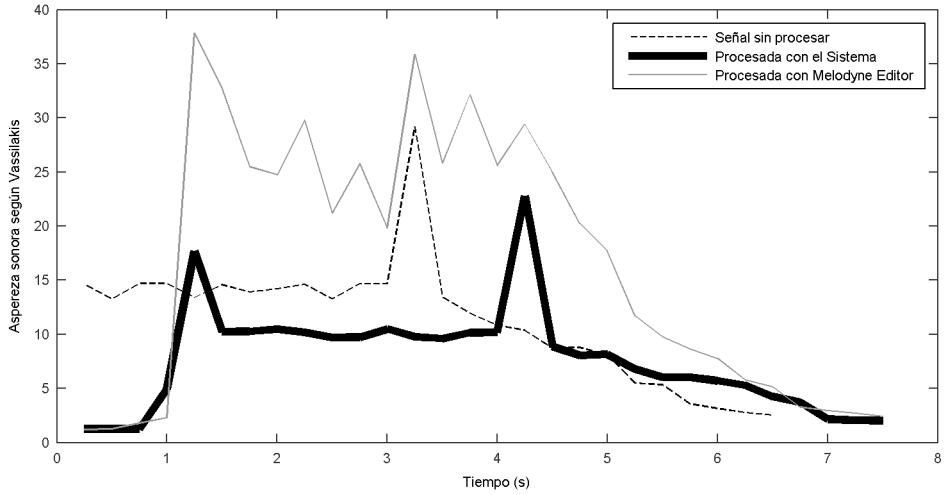


Figura 6.7: Resultado del análisis del acorde Sintético 2 con la herramienta online *SRA: Spectral and Roughness Analysis of sound signals*, basada en el algoritmo de Vassilakis [49].

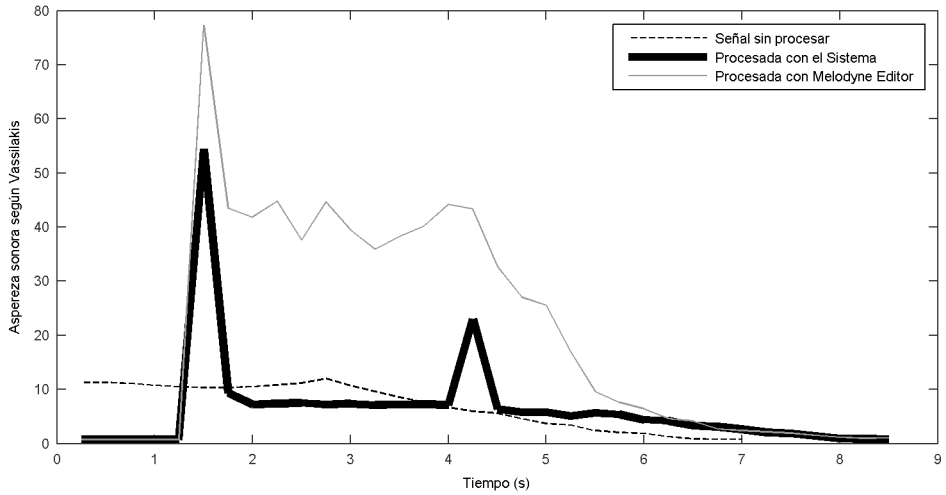


Figura 6.8: Resultado del análisis del acorde Sintético 3 con la herramienta online *SRA: Spectral and Roughness Analysis of sound signals*, basada en el algoritmo de Vassilakis [49].

A partir de los resultados que ofrece el algoritmo, se extraen una serie de conclusiones:

- Los sonidos procesados por el sistema desarrollado, en general, ofrecen una menor aspereza que el resto de sonidos. Este resultado es lógico, teniendo en cuenta que la forma de reducir la disonancia está inspirada en el procedimiento de cuantificación que se utiliza en el algoritmo utilizado.

- Resulta llamativo el resultado que ofrece Melodyne Editor. En general, ofrece una aspereza objetiva incluso mayor que el sonido sin procesar, a pesar de que los cuestionarios ofrecen resultados contradictorios. La razón radica en que el algoritmo [49] utiliza una curva similar a la expuesta en la figura 6.9 para la cuantificación de la aspereza producida por un par de parciales.

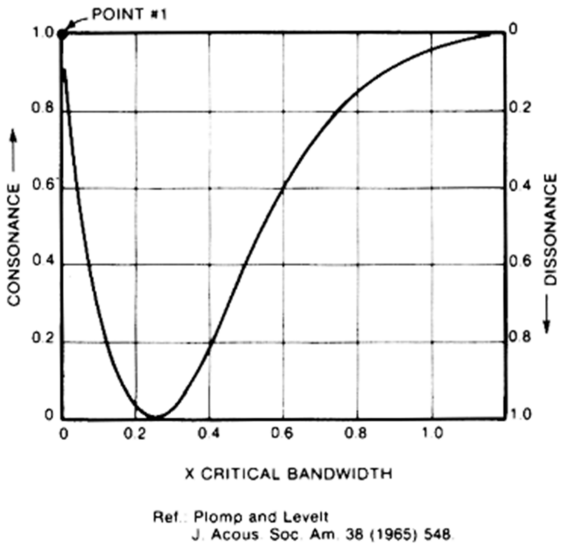


Figura 6.9: Curva propuesta por Plomp en [40] para la cuantificación de la aspereza producida por un par de parciales.

En el acorde original, la desviación introducida a los parciales es bastante grande, provocando el efecto musical de una gran desafinación. Sin embargo, esta gran desviación es considerada por el algoritmo como poco disonante, al no provocar interferencias dentro de la misma banda crítica. Al procesar el sonido con Melodyne, los parciales se acercan entre sí y reducen la desafinación, aunque se produce un incremento de la aspereza computada. Las diferencias entre los cuestionarios y el algoritmo (especialmente notable en 6.6) por tanto se debe a la pureza del sonido (basado en tonos perfectos) y a las consideraciones de tipo musical (el acorde utilizado es de uso común). Como se observará en experimentos posteriores, en general los resultados de la aspereza sí que están muy relacionados con la disonancia musical.

- En los resultados se observan picos de aspereza en el inicio y el final del acorde procesado (especialmente en las curvas 6.7 y 6.8). Aunque no resulta perceptible, los cambios de amplitud son puntos en los que la componente residual y sinusoidal no están perfectamente separadas, dando lugar a superposiciones de armónicos disonantes durante un brevísimo instante de tiempo.

6.3. Experimento 2: Guitarra Acústica

En este experimento se han procesado sonidos grabados de una guitarra acústica real desafinada. En concreto se ha trabajado con dos acordes diferentes, La Mayor y Re Mayor. Cada acorde está compuesto por las siguientes notas:

- Acorde 1 (La Mayor desafinado): A2 - 33 cents, E2 - 33 cents, A3 + 27 cents, C#4 + 16 cents, E4 - 20 cents

- Acorde 2 (Re Mayor desafinado): D3 - 30 cents, A3 + 28 cents, D4 + 15 cents, F#4 + 3 cents

Las desviaciones con respecto a la frecuencia ideal han sido medidas directamente sobre el espectrograma, analizando las frecuencias fundamentales de cada cuerda.

Ambos acordes tienen una sonoridad disonante, plagada de batidos y de intervalos que se desvían de los comúnmente utilizados. Las grabaciones han sido realizadas a 44100Hz / 16 bits, en mono y con un micrófono de condensador de alta calidad en un estudio de grabación.

Para el procesado, se ha utilizado como referencia la escala temperada a 440Hz. Las f_0s han sido introducidas manualmente, ya que de forma automática el sistema no era capaz de detectar adecuadamente todas las notas del acorde. En general, el procesado se realiza adecuadamente, suprimiéndose los batidos con éxito y ajustando a las posiciones correctas cada parcial. En la figura 6.10 se muestra el caso del La Mayor desafinado en cada etapa del Sistema a modo de ejemplo.

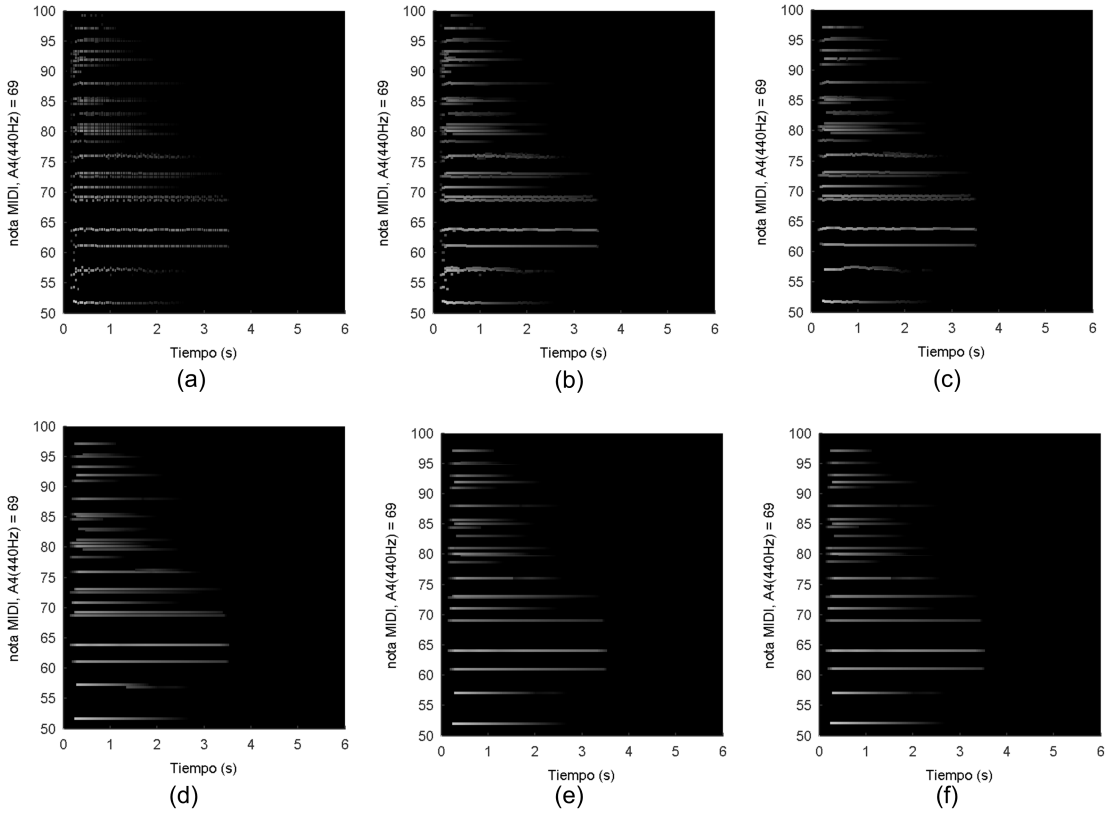


Figura 6.10: Acorde La Mayor desafinado con guitarra acústica (Notas: [A2, E2, A3, C#4, E4], Números MIDI: [45, 52, 57, 61, 64]): **(a)** Resultado de la estimación sinusoidal **(b)** Seguimiento temporal de los parciales **(c)** Eliminación de parciales de corta duración **(d)** Estabilización de parciales **(e)** Ajuste a la nueva estructura armónica **(f)** Traslación a la afinación 440Hz.

6.3.1. Análisis subjetivo de los resultados

En la tabla 6.3.1 se muestra un resumen de los resultados obtenidos en los 31 cuestionarios pasados a músicos con formación de conservatorio. Las tres variantes utilizadas para cada acorde son:

- G. Acú 1.A: Acorde La Mayor desafinado sin procesar.
- G. Acú 1.B: Acorde La Mayor desafinado procesado con el Sistema.
- G. Acú 1.C: Acorde La Mayor desafinado procesado con Melodyne Editor.
- G. Acú 2.A: Acorde Re Mayor desafinado sin procesar.

- G. Acú 2.B: Acorde Re Mayor desafinado procesado con el Sistema.
- G. Acú 2.C: Acorde Re Mayor desafinado procesado con Melodyne Editor.

Variante	**Consonancia subj. media y desv. típica**	**Naturalidad subj. media y desv. típica**	**Escogido como mejor resultado**
G. Acú. 1.A	mean(c) = 4,61 std(c) = 1,90	mean(n) = 6,54 std(n) = 2,23	3.2 %
Guit. Acú. 1.B	**mean(c)=7.19** **std(c)=1.83**	**mean(n)=7.29** **std(n)=1.86**	**83.9 %**
G. Acú. 1.C	mean(c) = 5,83 std(c) = 2,35	mean(n) = 6,64 std(n) = 2,18	9.7 %
G. Acú. 2.A	mean(c) = 4,32 std(c) = 1,81	mean(n) = 5,77 std(n) = 1,85	3.2 %
G. Acú. 2.B	**mean(c)=7.09** **std(c)=1.68**	**mean(n)=6.77** **std(n)=1.78**	**71 %**
G. Acú. 2.C	mean(c) = 6,19 std(c) = 1,99	mean(n) = 6,58 std(n) = 1,70	25.8 %

Tabla 6.3: Resultados en el caso de la guitarra acústica

De esta tabla de resultados se extraen las siguientes conclusiones:

- El sistema desarrollado tiene un comportamiento especialmente bueno en el caso de la guitarra. Ello queda demostrado por la gran diferencia en la valoración media de la consonacia entre la variante A y B (diferencias de 2.58 y 2.77 para el acorde 1 y 2 respectivamente). Existe además una diferencia significativa entre la variante B y C, lo cual quiere decir que el Sistema funciona bastante mejor que Melodyne Editor para este tipo de sonidos. Especialmente interesante resulta el dato que indica la cantidad de veces que los usuarios consideran la variante B como el mejor resultado, llegando al 83.9 % en el caso del La Mayor. En todas las comparativas, las diferencias en la valoración media han sido avaladas con la prueba t-Student.
- Vuelve a existir una correlación entre la naturalidad y la consonancia. Este resultado es sorprendente, ya que los usuarios no sólo no son capaces de detectar la falta de naturalidad que provoca el Sistema, sino que llegan a considerar que es más natural el acorde procesado.
- La conclusión global es que el sistema implementado ofrece resultados muy aceptables, de gran utilidad práctica y con posibilidad de ser aplicados en el ámbito de la grabación profesional de estudio.

6.3.2. Análisis objetivo de los resultados

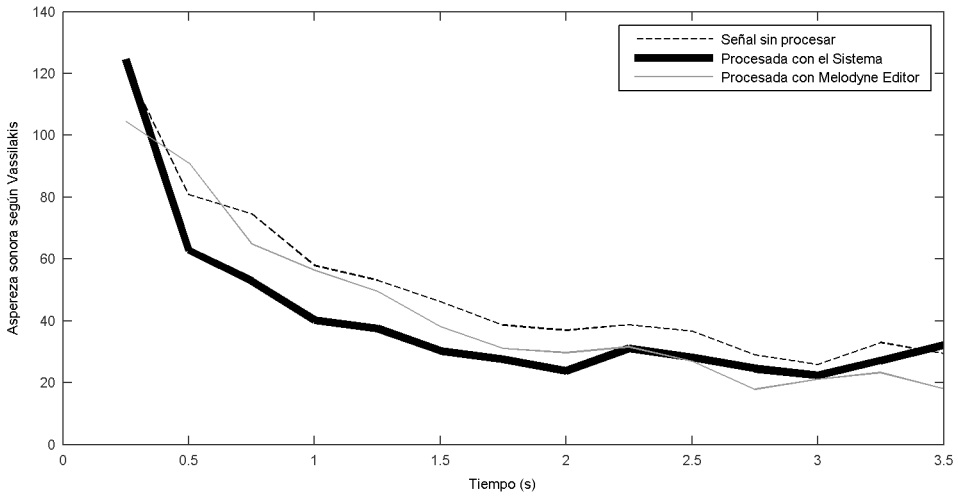


Figura 6.11: Resultado del análisis del acorde La Mayor de guitarra acústica con la herramienta online *SRA: Spectral and Roughness Analysis of sound signals*, basada en el algoritmo de Vassilakis [49].

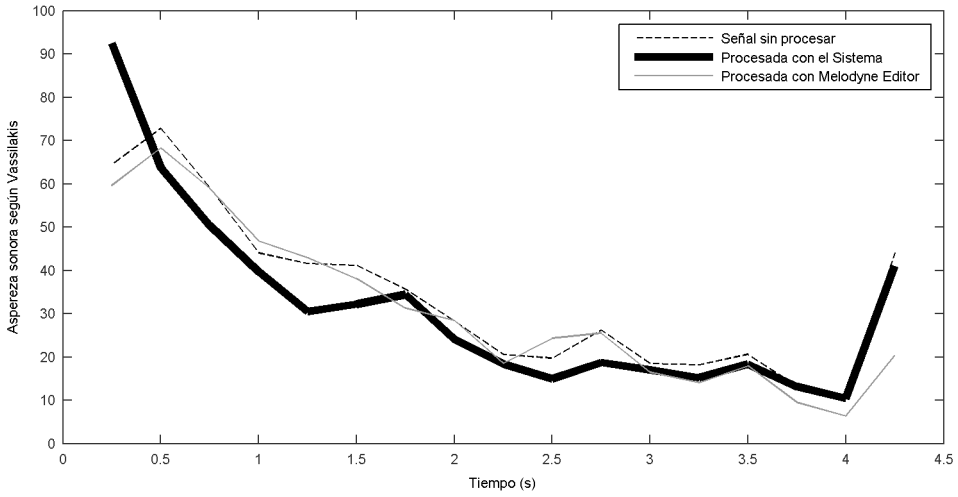


Figura 6.12: Resultado del análisis del acorde Re Mayor de guitarra acústica con la herramienta online *SRA: Spectral and Roughness Analysis of sound signals*, basada en el algoritmo de Vassilakis [49].

Como se observa en las figuras 6.11, 6.12, el análisis objetivo resulta totalmente coherente con el análisis subjetivo. En general, la reducción de la aspereza con respecto al original es notable en los tramos de mayor energía de la señal (teniendo en cuenta la envolvente en forma de exponencial decreciente).

Sorprende la poca influencia que tiene sobre la aspereza sonora el procesado a través de Melodyne, algo que puede explicarse por una mala resolución de armónicos cercanos entre sí. Definitivamente, el caso de la guitarra es un buen ejemplo de instrumento apropiado para ser procesado con el sistema desarrollado.

6.4. Experimento 3: Conjuntos instrumentales

En este experimento se han diseñado sonidos de conjuntos intrumentales desafinados a partir de una grabación multipista de las notas de cada instrumento del conjunto. Concretamente se han utilizado los siguiente acordes:

- Acorde Do Mayor desafinado tocado por cuarteto de cuerda frotada (Dos violines, viola y violoncello): C3 - 6 cents, E3 - 7 cents, C4 + 30 cents, G4 - 73 cents.
- Acorde La Mayor desafinado tocado por cuarteto de viento metal (Tuba, Trombón, Trompa y Trompeta): A2, E3 + 49 cents, A3 - 5 cents, C#4 - 33 cents.
- Acorde Si♭ Mayor desafinado tocado por cuarteto de viento madera (2 fagots, clarinete y oboe): B♭2, F3 - 44 cents, B♭3 - 50 cents, D5 + 31 cents.

Al igual que el caso de la guitarra, las desviaciones de frecuencia han sido medidas sobre el espectrograma, en el cual se visualiza bastante bien la posición exacta de las diferentes frecuencias fundamentales.

Los acordes han sido procesados para ser ajustados a la afinación temperada centrada en 440Hz. La detección automática de f_0s, en general estima las notas más importantes de los acordes, pero no todas ellas. Por ello, las notas que forman los acordes han sido introducidas manualmente para mejorar el funcionamiento del sistema. Esta introducción manual de las notas del acorde está contemplada en el contexto de uso real de la herramienta.

Para mostrar gráficamente el funcionamiento del Sistema con este tipo de sonidos, en la figura 6.13 se muestran los parciales de un acorde Do mayor desafinado tocado por un cuarteto de cuerda frotada en cada etapa del Sistema. Se observa una gran densidad de parciales y de picos espúreos, lo que hace que el procesado provoque

una mayor antinaturalidad en el sonido final, aunque la sensación de afinación y consonancia se vea mejorada.

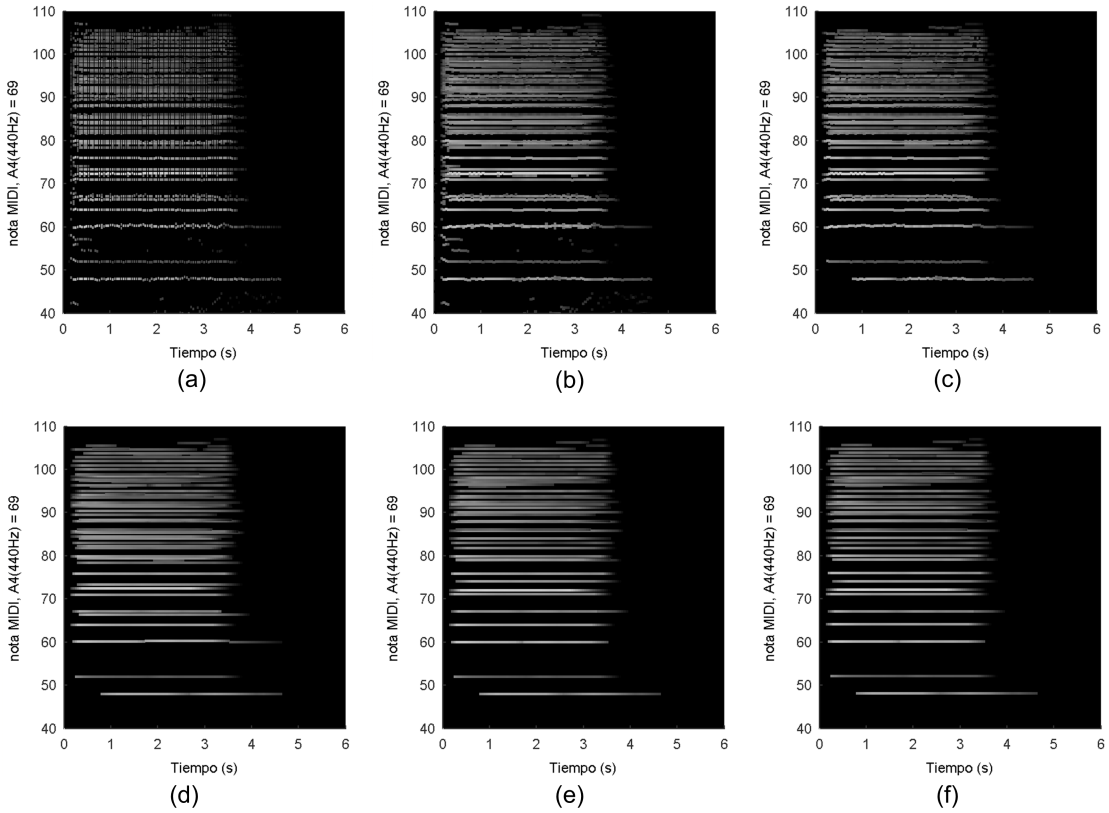


Figura 6.13: Do mayor desafinado tocado con cuarteto de cuerda (Notas: [C3, E3, C4, G4] Números MIDI: [48, 52, 60, 67]): **(a)** Resultado de la estimación sinusoidal **(b)** Seguimiento temporal de los parciales **(c)** Eliminación de parciales de corta duración **(d)** Estabilización de parciales **(e)** Ajuste a la nueva estructura armónica **(f)** Traslación a la afinación 440Hz.

6.4.1. Análisis subjetivo de los resultados

Al igual que en los dos anteriores experimentos, se han trabajado con 3 variantes de cada acorde:

- Cuerdas 1.A: Do Mayor con cuarteto de cuerda sin procesar.
- Cuerdas 1.B: Do Mayor con cuarteto de cuerda procesado con el Sistema.
- Cuerdas 1.C: Do Mayor con cuarteto de cuerda procesado con Melodyne Editor.

- V-metal 2.A: La Mayor con cuarteto de viento metal sin procesar.
- V-metal 2.B: La Mayor con cuarteto de viento metal procesado con el Sistema.
- V-metal 2.C: La Mayor con cuarteto de viento metal procesado con Melodyne Editor.
- V-madera 3.A: Si♭ Mayor con cuarteto de viento madera sin procesar.
- V-madera 3.B: Si♭ Mayor con cuarteto de viento madera procesado con el Sistema.
- V-madera 3.C: Si♭ Mayor con cuarteto de viento madera procesado con Melodyne Editor.

Variante	**Consonancia subj. media y desv. típica**	**Naturalidad subj. media y desv. típica**	**Escogido como mejor resultado**
Cuerdas 1.A	mean(c) = 1,54 std(c) = 0,80	mean(n) = 3,22 std(n) = 1,85	0 %
Cuerdas 1.B	**mean(c)=5.54** **std(c)=2.15**	**mean(n)=4.83** **std(n)=2.19**	**77.4 %**
Cuerdas 1.C	mean(c) = 4,77 std(c) = 1,96	mean(n) = 4,70 std(n) = 2,14	22.6 %
V-metal 2.A	mean(c) = 3,54 std(c) = 1,23	mean(n) = 4,77 std(n) = 1,83	6.45 %
V-metal 2.B	**mean(c)=6.51** **std(c)=1.99**	**mean(n)=6** **std(n)=1.91**	**31 %**
V-metal 2.C	mean(c) = 7,16 std(c) = 2,08	mean(n) = 6,9 std(n) = 1,61	62.6 %
V-madera 3.A	mean(c) = 2,19 std(c) = 1,27	mean(n) = 3,38 std(n) = 1,80	0 %
V-madera 3.B	**mean(c)=4.03** **std(c)=2.33**	**mean(n)=4.41** **std(n)=2.48**	**32 %**
V-madera 3.C	mean(c) = 4,64 std(c) = 2,38	mean(n) = 4,67 std(n) = 2,05	68 %

Tabla 6.4: Resultados en el caso de acordes sintéticos

A la vista de los resultados, se pueden extraer una serie de conclusiones:

- En el caso de conjuntos intrumentales con notas sostenidas como los estudiados, el Sistema mejora sustancialmente la sensación de disonancia producida por el sonido sin procesar (hasta 5 puntos de diferencia en el cuarteto de cuerda). Sin embargo, Melodyne Editor demuestra ser un competidor importante para el caso de instrumentos de viento, donde alrededor del 60 % de los usuarios prefieren el sonido procesado por este software. Las diferencias entre las medias han sido testeadas con la t-Student obteniendo $p < 5\,\%$ en todos los casos.

- Al igual que en los casos anteriores, la correlación entre consonancia y naturalidad sigue existiendo.

- La conclusión es que el Sistema ofrece un comportamiento aceptable, pero no excelente debido a las oscilaciones naturales de un instrumento de viento. El Sistema, al estabilizar los parciales suprime estas oscilaciones e introduce una antinaturalidad que es percibida de alguna forma como disonancia. Melodyne Editor, para instrumentos de viento resulta ser más adecuado.

6.4.2. Análisis objetivo de los resultados

En las figuras 6.14, 6.15 y 6.16, se muestra la aspereza sonora según Vassilakis en [49] para cada sonido analizado. Como se observa, el sonido procesado por el Sistema suele ofrecer una aspereza menor que el resto de variantes.

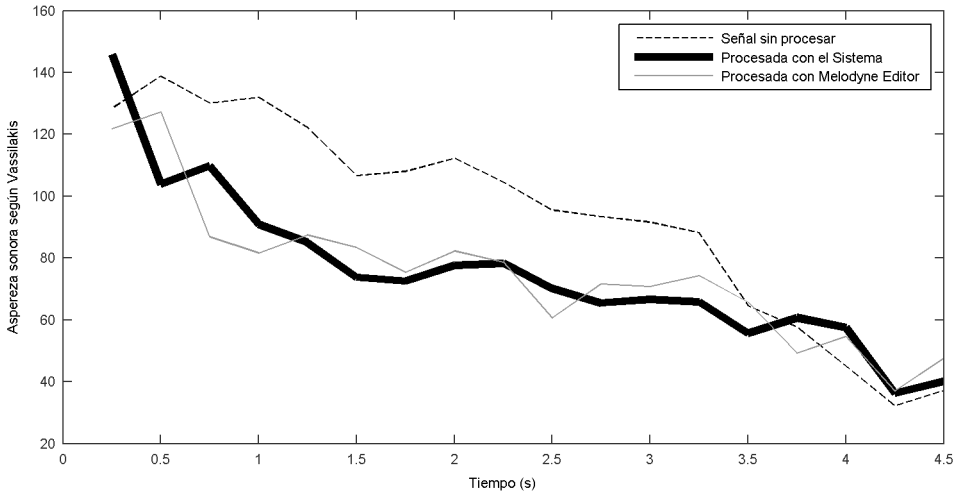


Figura 6.14: Resultado del análisis del acorde Do Mayor de un cuarteto de cuerda frotada con la herramienta online *SRA: Spectral and Roughness Analysis of sound signals*, basada en el algoritmo de Vassilakis ([49]).

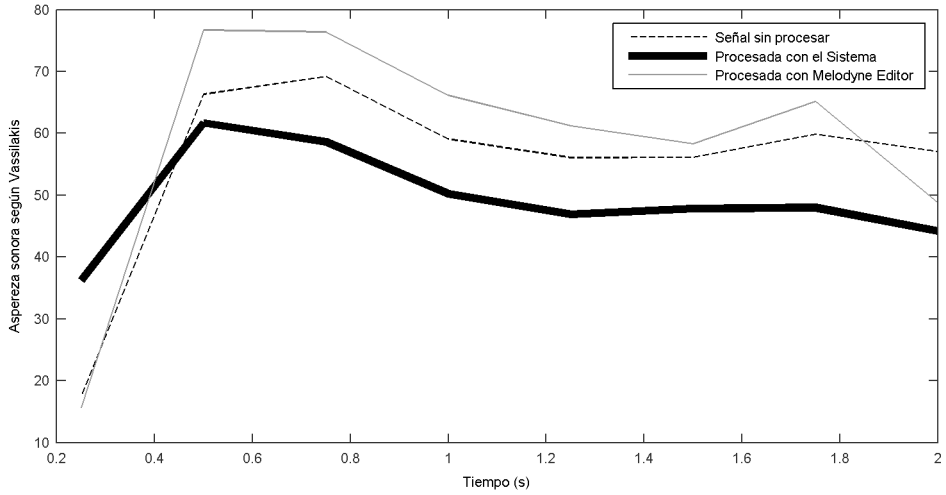


Figura 6.15: Resultado del análisis del acorde La Mayor de un cuarteto de viento-metal con la herramienta online *SRA: Spectral and Roughness Analysis of sound signals*, basada en el algoritmo de Vassilakis ([49]).

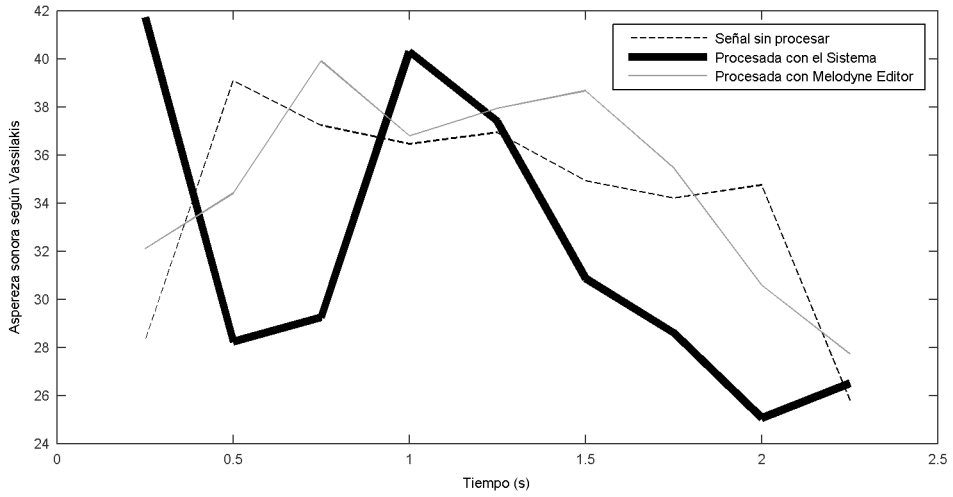


Figura 6.16: Resultado del análisis del acorde Sib Mayor de un cuarteto de viento-madera con la herramienta online *SRA: Spectral and Roughness Analysis of sound signals*, basada en el algoritmo de Vassilakis ([49]).

A partir de los resultados, se extraen las siguientes conclusiones:

- En el caso del La Mayor tocado por un cuarteto de viento-metal (figura 6.15) se produce un fenómeno similar al caso de los sonidos sintéticos (apartado

6.2). La curva del sonido procesado con Melodyne Editor ofrece un valor de aspereza mayor que las demás. La razón, al igual que sucedía con sonidos sintéticos, es la presencia de batidos en el sonido procesado. Estos batidos, perceptualmente "chocan" entre sí y contribuyen a aumentar la aspereza, sin embargo el sonido resultante tiene más coherencia musical (se acerca más a una sonoridad familiar). De ahí las discrepancias entre el análisis subjetivo y objetivo.

- La oscilación que se muestra en el sonido procesado por el Sistema de la figura 6.16 parece estar relacionada con la aparición de parciales espúreos, fruto de un análisis en el cual los parciales están registrados de forma intermitente. La aparición y desaparición de parciales incrementa la curva de aspereza debido a las irregularidades espectrales que crea.

Capítulo 7

Conclusiones y líneas futuras de trabajo

Durante la realización de este Proyecto, ha sido necesario estudiar y profundizar en conceptos provenientes de diferentes áreas temáticas. Por un lado, ha sido necesario comprender con detalle el concepto de *disonancia*, tanto a nivel musical como a nivel perceptual, encontrando aquella definición que permita su modelado computacional para su posterior tratamiento. Por otro lado, se han estudiado e implementado las técnicas de procesado de la señal necesarias para lograr una manipulación adecuada del contenido musical. Dada la componente subjetiva que existe tras el concepto de disonancia, ha sido necesario realizar una evaluación rigurosa del Sistema a través de un análisis, tanto objetivo como subjetivo, de los resultados.

En este capítulo se comentan las conclusiones más importantes extraídas de cada fase del desarrollo, y además se comentan las líneas futuras de trabajo más interesantes.

7.1. Conclusiones

En este Proyecto Fin de Carrera, se ha desarrollado un sistema capaz de procesar material polifónico disonante, analizar la causa de dichas disonancias y resolverlas para conseguir una versión consonante del mismo sonido. A continuación se recopilan de forma desglosada las conclusiones extraídas a lo largo del desarrollo.

Aspectos generales

- La idea global del Proyecto aporta una novedad en el campo del post-procesado en estudio de grabación, ya que a diferencia de otros sistemas, éste es capaz de procesar la polifonía como un todo. La implementación comercial del sistema desarrollado daría lugar a un producto útil que no es posible encontrar en el mercado hoy día.

- Se ha comprobado que el modelo de señal sinusoidal + residual definido por Serra en [43] se ajusta muy bien a las necesidades del Proyecto. Este modelo ha sido utilizado con éxito para obtener una representación versátil de la componente armónica de la señal de entrada. La comprensión de este modelo ha requerido profundizar en numerosos conceptos asociados a la Transformada Corta de Fourier (STFT).

- Se ha realizado un estudio detallado del concepto de disonancia, del cual se ha concluído que se trata de un término difícil de definir por su importante componente subjetiva. Tras este estudio se ha utilizado la definición de disonancia propuesta por Plomp y Levelt en [40], la cual contempla esta subjetividad pero a la vez permite un manejo computacional del mismo.

- Se ha diseñado un algoritmo novedoso para el procesado de la señal, que consigue manipular la estructura armónica de un sonido polifónico de forma conjunta para mejorar su consonancia. La hipótesis inicial que inspiró este algoritmo es que la disonancia puede ser atenuada mediante el reajuste de los parciales a posiciones con relaciones armónicas, evitando así los desagradables batidos. Los resultados demuestran que la hipótesis inicial estaba correctamente planteada, con una serie de matices que ya se han comentado a lo largo del Proyecto.

- Los resultados obtenidos reflejan que la herramienta desarrollada, en general, logra mejorar sustancialmente la consonancia de acordes tocados con gran variedad de instrumentos. Especialmente en el caso de sonidos sintéticos y acordes de guitarra, alrededor del 75 % de los usuarios consideraron que los sonidos procesados con el sistema desarrollado eran los que mejor sonaban en terminos de consonancia y naturalidad.

Percepción de la disonancia

- Se ha concluído que no existe una definición óptima de *disonancia*, ya que depende del contexto y de la percepción subjetiva. Es por ello que ha sido necesario estudiar la bibliografía al respecto, para así encontrar una definición que fuera realmente útil para este Proyecto.

- Se ha visto que las ideas propuestas por Plomp en [40] ofrecen un modelo para la cuantificación de la disonancia que, combinado con algunas consideraciones musicales, ha ofrecido buenos resultados.

Análisis

El análisis llevado a cabo para modelar la señal de entrada se basa en el esquema de un modelo sinusoidal + residual estándar, planteado por Serra en [43]. El objetivo

es parametrizar de la forma más útil posible el contenido armónico, y extraerlo de la señal original de la forma más efectiva posible. Este método de análisis generalmente obtiene una parametrización aceptable de la señal. A continuación se comentan una serie de conclusiones extraídas de esta etapa del Proyecto.

- Se ha realizado un estudio analítico del caso en el que existen armónicos demasiado cercanos que no consiguen resolverse de forma independiente. Cuando dos armónicos "chocan" entre sí, el resultado es una serie de picos espectrales oscilantes en magnitud y frecuencia, que en ocasiones no son posibles de identificar como pertenecientes a una única entidad. En el subapartado 6.1.1 de este Proyecto, se incluyen algunas expresiones matemáticas que explican el resultado de la STFT en este tipo de casos.
- Se ha conseguido un conjunto de parámetros (expuestos en el subapartado 3.4.4) que ofrecen resultados muy aceptables para una gran variedad de sonidos. Esto evita el constante ajuste manual por parte del usuario. No obstante, en ciertos casos muy concretos es conveniente realizar dicho ajuste si se desean resultados prácticos.
- La inharmonicidad de ciertos timbres no está contemplada en la detección de f_0s, pero la detección polifónica de notas no es una etapa crítica del sistema, ya que puede ser ajustada manualmente.

Procesado

Una vez que el análisis ha sido llevado a cabo, idealmente se debería disponer de una componente sinusoidal correctamente parametrizada y una componente residual sin rastros de contenido armónico. Si las f_0s han sido correctamente extraídas, los parciales están debidamente diferenciados y corresponden a diferentes sonidos armónicos, se dispone de información suficiente para manipular adecuadamente el sonido original. A continuación se exponen las conclusiones extraídas de la fase de procesado.

- La principal novedad de todo el Proyecto, como ya se ha comentado, reside en el Subsistema de Procesado. Las técnicas utilizadas han sido desarrolladas especialmente para la funcionalidad de esta herramienta, e introducen una novedad sobre los sistemas existentes al considerar el material polifónico como un todo.
- Se ha comprobado que los resultados obtenidos tras el procesado son generalmente buenos si los parámetros están debidamente ajustados y el análisis se llevó a cabo exitosamente.

Síntesis

- Se ha implementado un método de síntesis basado en [43], donde el contenido armónico de cada ventana temporal es sintetizado en el dominio de la frecuencia, para luego trasladarlo al dominio del tiempo mediante la IFFT. Posteriormente, estas ventanas se superponen siguiendo el procedimiento superposición-suma con un factor de superposición del 75 %. Su principal ventaja es el bajo coste computacional que supone.

- Se ha comprobado que este procedimiento ofrece una limitada resolución temporal para sonidos com cambios bruscos. Sin embargo, el uso de una componente residual compensa estos problemas, ya que en ella sí están registrados los ataques de las notas, así como las componentes ruidosas.

- Se ha estudiado detalladamente la bibliografía existente en referencia a la síntesis, concluyendo que existen otros procedimientos para la síntesis de sinusoides no estacionarias [14], que sin duda mejorarían los ataques y otros aspectos de la señal sintetizada. Sin embargo, dada la complejidad que supone la implementación de estos métodos, no se ha considerado que merezca la pena su implementación, ya que con el sencillo método propuesto por Serra en [43] ya se consiguen resultados bastante buenos.

- Mientras que el Subsistema de Análisis y el Subsistema de Procesado sí que pueden suponer problemas en el caso de señales reales de características complejas, el Subsistema de Síntesis tiene un comportamiento aceptable en la mayoría de las ocasiones.

Evaluación de resultados

La evaluación rigurosa del funcionamiento del sistema ha sido una de las etapas más laboriosas de todo el desarrollo del Proyecto. Tal y como se planteó en un principio, se ha utilizado evaluación por triangulación, al combinar un análisis subjetivo basado en cuestionarios con un análisis objetivo. A continuación se comentan las conclusiones más importantes sobre la evaluación de resultados.

- La valoración global de los resultados es buena o muy buena, y de ella se concluye que el sistema desarrollado cumple con éxito su función. Excepto casos concretos, los usuarios consideran que los sonidos mejoran sustancialmente al ser procesados con el sistema desarrollado (ver capítulo 6).

- Se ha diseñado un conjunto de sonidos representativo, que contempla los distintos casos en los que puede utilizarse la herramienta desarrollada.

- Se ha realizado una evaluación de resultados rigurosa, en la cual se han utilizado 31 sujetos con más de 7 años de formación musical oficial de edades y género diversos. Además de comparar la media y la desviación típica de los resultados, también se ha realizado la prueba t-Student para garantizar su validez estadística.

- Se ha realizado un análisis objetivo de los resultados basado en el algoritmo para la cuantificación de la disonancia implementado en [49]. Tras una revisión de dicho artículo, se concluye que está basado en fuentes importantes y por tanto es un algoritmo fiable.

A la vista de los resultados tan interesantes que se han obtenido, así como de la novedad inherente al sistema desarrollado, actualmente se está elaborando un artículo que pretende ser publicado en la revista *IEEE Transactions on Audio, Speech, and Language Processing*.

7.2. Líneas futuras de trabajo

En este apartado se exponen las numerosas líneas de trabajo que se proponen para mejorar y hacer más funcional el Sistema. En primer lugar se proponen mejoras que se pueden aplicar a los bloques funcionales existentes. Esto mejoraría el comportamiento global del Sistema, y sin duda es una interesante forma de trabajar sobre el diseño propuesto. Posteriormente, se plantea la estandarizaración como un plugin *VST* del Sistema desarrollado. Esto permitiría integrarlo en los entornos de grabación comúnmente usados en los estudios de grabación. Por último, se propone un esquema que permite aplicar el procesado a diferentes armonías consecutivas. Este esquema añade algunos bloques funcionales al Sistema, que no han sido desarrollados porque se trata de un problema complejo que requeriría un Proyecto de mayor envergadura.

Mejoras de los bloques funcionales

El buen comportamiento del Sistema está determinado por el funcionamiento de cada uno de los bloques funcionales que lo forman. Dado el diseño modular que se ha utilizado, mejorar el comportamiento del Sistema se traduce en depurar cada bloque funcional separadamente.

- **STFT + Estimación de sinusoides:** Estos dos bloques tienen como objetivo hacer una estimación efectiva de la componente sinusoidal. El sistema utilizado es el propuesto por Serra en [42], el cual es sencillo de implementar y ofrece resultados aceptables en señales de características favorables. Sin

embargo, en señales reales, este procedimiento no siempre consigue una componente sinusoidal pura ni una componente residual limpia de restos sinusoidales. Existen métodos más sofisticados que pueden mejorar la estimación sinusoidal [10, 11, 13, 46, 28, 29, 32, 34], cuya implementación se propone como línea futura.

- **Seguimiento temporal de sinusoides:** El seguimiento temporal es otro punto clave para que el resultado del procesado sea natural y tenga relación directa con el modelo físico que produce el sonido. El sistema utilizado es el mencionado en [42], el cual también es fácil de implementar pero no es robusto a problemas en el análisis. En [9, 26, 52] se plantean diferentes métodos que pueden ser utilizados para mejorar el seguimiento temporal de los parciales.

- **Estimación de f_0s:** Este bloque puede considerarse un "cuello de botella" del comportamiento global del sistema. La estimación de distintas f_0s en un entramado complejo de parciales es un problema complicado, y en este Sistema se ha aplicado una solución sencilla (la cual se explica en el apartado 3.7). La línea futura más clara sería implementar dicho bloque según el algoritmo planteado por Klapuri en [25], el cual puede considerarse el sistema publicado más completo actualmente. El diseño del Sistema es robusto a una detección incompleta de las f_0s, pero obviamente el resultado con señales conflictivas mejoraría claramente con una correcta estimación de las f_0s que lo forman.

- **Estabilización de parciales:** Este bloque pretende simplificar la parametrización de la componente sinusoidal, a la vez que eliminar los batidos en parciales de baja frecuencia que no han sido correctamente resueltos. Para ello se busca que todos los parciales sean estables en frecuencias, sin oscilaciones periódicas de amplitud. La forma de corregir las oscilaciones periódicas de amplitud puede ser mejorada, y marca una clara línea futura de trabajo. En primer lugar, no está preparada para más de un ataque, y la envolvente de amplitud no está del todo conseguida. Para mejorar la naturalidad del sistema sería necesario mejorar el cálculo de la envolvente de amplitud para casos más generales.

- **Generación de nueva estructura armónica:** En la solución implementada, no se considera la posibilidad de timbres no armónicos. Sería una interesante mejora estimar de alguna forma la inharmonicidad del timbre y generar una estructura armónica más natural a partir de las f_0s.

- **Traslación de parciales a nueva estructura armónica:** Este bloque permite implementaciones más o menos simples. La versión más sencilla consiste en acercar cada parcial a su posición más cercana de la "rejilla" generada (solución implementada). La línea futura propuesta en este bloque consiste en

identificar a qué f_0 corresponde cada parcial, y desplazarlo en consecuencia. Esta tarea no es fácil, ya que implica numerosos conceptos de *source separation* (separación de fuentes sonoras), y es un problema aún no solucionado del todo.

- **Síntesis de la componente sinusoidal:** Aunque este bloque tiene un buen funcionamiento tal y como está implementado, en ocasiones la resolución temporal de la componente sinusoidal provoca efectos indeseados en los ataques de las notas. Sería interesante combinar un análisis más detallado de la componente sinusoidal con una síntesis más efectiva, que garantice una resolución temporal adecuada a las características de la señal.

Implementación como plugin VST

El Sistema ha sido implementado en Matlab como prototipo. Se propone como línea futura de trabajo realizar la implementación en C++ según el estándar VST [50], ya que eso permitiría su integración con los entornos multipista más comúnmente usados en estudios de grabación.

Tras leer con detalle la documentación del *VST Software Development Kit*, se llega a la conclusión que el paradigma de este tipo de efectos está planteado para ser usados a tiempo real. Sin embargo, existen ejemplos que rompen esta norma como puede ser *Melodyne Editor*, que utiliza ficheros auxiliares internos para implementar un procesado *offline* (a posteriori) en el formato VST. La línea de trabajo futura propuesta consiste en realizar una implementación similar sobre un protocolo VST del Sistema desarrollado.

Detección de ataques y de cambios armónicos

Todo el planteamiento del Sistema está orientado a ser aplicado a un breve intervalo de tiempo, de no más de 10 o 15 segundos, donde la armonía sea estable pero existan desafinaciones. En este apartado se plantea de qué forma se podría extender este Sistema para ser aplicado a señales que contengan distintos acordes y/o ataques.

El Sistema en sí no tendría que ser modificado, tan sólo habría de incluirse en un sistema mayor que incluyera algunos bloques funcionales nuevos. En la figura 7.1 se muestra un diagrama de bloques del esquema que se propone.

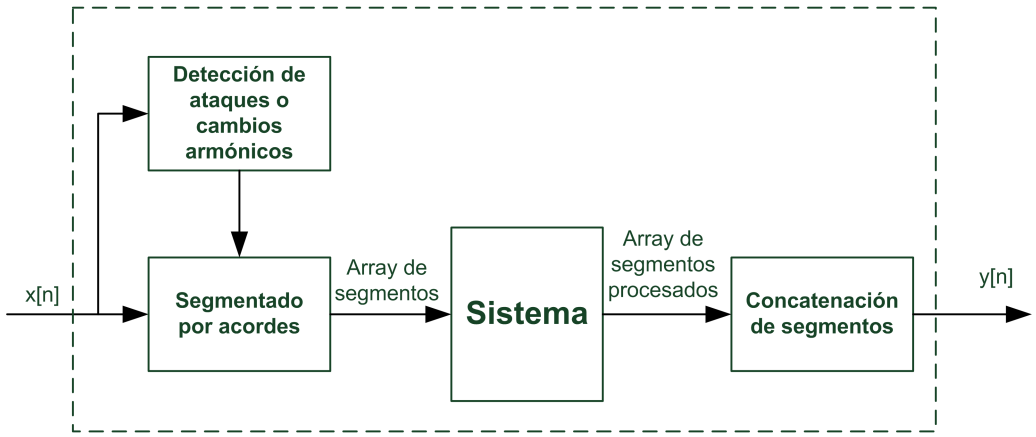


Figura 7.1: Diagrama de bloques de un posible sistema para procesar distintos acordes consecutivos.

La gran dificultad se encuentra en la implementación del bloque *Detección de ataques o cambios armónicos*, ya que resulta muy difícil establecer un método "inteligente" que sea capaz de determinar cuándo existe un cambio armónico y cuándo no.

Para implementar el detector de cambios armónicos, se propone utilizar el estimador de vector de croma implementado por Dan Ellis en [12]. Este algoritmo permite sintetizar en un sencillo vector de 12 posiciones con información sobre la armonía de una ventana temporal. Para detectar ataques, se puede tener un conocimiento general sobre las técnicas comúnmente usadas en [5].

Bibliografía

[1] Antares ATG-6: Auto-tune for guitar. Publicado: 16/05/2011, Último acceso: 18/11/2011 `http://www.harmonycentral.com/videos/2644`.

[2] Diccionario de la Real Academia Española. Último acceso: 30/01/2012, `http://www.rae.es/rae.html`.

[3] M. Abe and J.O. Smith. Design criteria for the quadratically interpolated fft method (i): Bias due to interpolation. *October*, 5(I):1, 2004.

[4] J.M. Barbour. *Tuning and temperament: a historical survey*. Dover Publications, 2004.

[5] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, September 2005.

[6] N. Cazden. Sensory Theories of Musical Consonance. *The Journal of Aesthetics and Art Criticism*, 20(3):301, 1962.

[7] S.H. Chon. *Quantifying the consonance of complex tones with missing fundamentals*. PhD thesis, Stanford University, 2008.

[8] R. Clark. *Mixing, recording, and producing techniques of the pros*. Thomson Course Technology PTR, 2006.

[9] P. Depalle, G. Garcia, and X. Rodet. *Tracking of partials for additive sound synthesis using hidden Markov models*, volume 1, pages 225–228. IEEE, 1993.

[10] P. Depalle and T. Hélie. Extraction of spectral peak parameters using a short-time Fourier transform modeling and no sidelobe windows. In *Proceedings of IEEE Workshop on Audio Mohonk*. Citeseer, Citeseer, 1997.

[11] K. Dressler. Sinusoidal extraction using an efficient implementation of a multi-resolution FFT. In *Proc of the Int Conf on Digital Audio Effects DAFx06 Montreal Quebec Canada*, pages 247–252, 2006.

[12] D. Ellis. Chroma feature analysis and synthesis, (Última actualización: 21/04/2007, Último acceso: 02/11/2011). `http://labrosa.ee.columbia.edu/matlab/chroma-ansyn/`.

[13] S.A. Fulop and K. Fitz. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *The Journal of the Acoustical Society of America*, 119(1):360, 2006.

[14] M. Goodwin and X. Rodet. *Efficient Fourier Synthesis of Nonstationary Sinusoids*, pages 333–334. Number 510. INTERNATIONAL COMPUTER MUSIC ACCOCIATION, 1994.

[15] O.R. Gurney. An old babylonian treatise on the tuning of the harp. *Iraq*, 30(2):229–233, 1968.

[16] K. Hahn and O. Vitouch. Preference for Musical Tuning Systems: How Cognitive Anatomy Interacts with Cultural Shaping. *cognitionunikluacat*, pages 757–760, 2002.

[17] F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

[18] H. Helmholtz. *Die Lehre den Tonempfindungen als physiologische Grundlage fur die Theorie der Musik / von H. Helmholtz*. F. Vieweg und Sohn, Braunschweig, 1877.

[19] APPLE INC., G. Steffen, S. Markus, and P. Fournier. Polyphonic note detection (patent), 10 2011.

[20] B. Jacobsson and J. Jerkert. Consonance of Non-Harmonic Complex Tones : Testing the Limits of the Theory of Beats. *Perception*, (1965):1–10, 1999.

[21] F. Jülicher, D. Andor, T. Duke, and F. Julicher. Physical basis of in hearing. *Sciences-New York*, 98(16):9080–9085, 2009.

[22] A. Kameoka and M. Kuriyagawa. Consonance theory part I: consonance of dyads. *The Journal of the Acoustical Society of America*, 45(6):1451–9, June 1969.

[23] A. Kameoka and M. Kuriyagawa. Consonance theory part II: consonance of complex tones and its calculation method. *The Journal of the Acoustical Society of America*, 45(6):1460–9, June 1969.

[24] D.F. Keislar. *Psychoacoustic Factors in Musical Intonation: Beats, Interval Tuning, and Inharmonicity*. PhD thesis, Standford University, 1992.

[25] A. Klapuri and M. Davy. Signal processing methods for the automatic transcription of music. *Tampere University of Technology Publications*, 460(March), 2004.

[26] M. Lagrange, S. Marchand, M. Raspaud, and J.B. Rault. *Enhanced partial tracking using linear prediction*, page 141146. Citeseer, 2003.

[27] Levelt and Plomp. The connotation of musical consonance. *Acta Psychologica*, 20:308319, 1962.

[28] S.N. Levine, T.S. Verma, and J.O. Smith. Alias-free, multiresolution sinusoidal modeling for polyphonic, wideband audio. *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, page 4.

[29] S.N. Levine, T.S. Verma, and J.O. Smith. Multiresolution sinusoidal modeling for wideband audio with modifications. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, pages 3585–3588, 1998.

[30] D. Maggiolo. Apuntes de acústica musical. Publicado: 2003, Último acceso: 30/01/2012, `http://www.eumus.edu.uy/docentes/maggiolo/acuapu/sap.html`.

[31] R.C. Maher. *An Approach for the Separation of Voices in Composite Musical Signals*. PhD thesis, University of Illinois, 1989.

[32] S. Marchand and P. Depalle. Generalization of the derivative analysis method to non-stationary sinusoidal modeling. *chaodyn9909042*, pages 1–8, 2008.

[33] B.C. Moore. Frequency difference limens for short-duration tones. *Journal of the Acoustical Society of America*, 54(3):610–619, 1973.

[34] D.E. Newland. Harmonic wavelet analysis. *Applied Optics*, 48(2):203–225, 1993.

[35] A. Nuttall. Some windows with very good sidelobe behavior. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(1):84–91, February 1981.

[36] N. Of. Calculating Sensory Dissonance : Some Discrepancies Arising from the Models of Kameoka & Kuriyagawa and Hutchinson & Knopoff. pages 65–84, 1965.

[37] A.V. Oppenheim, R.W. Schafer, and J.R. Buck. Discrete-Time Signal Processing, 1999.

[38] R. Parncutt. Commentary on Keith Mashinter â s â Calculating Sensory Dissonance : Some Discrepancies Arising from the Models of Kameoka & Kuriyagawa and Hutchinson & Knopoff â. *Perception*, 1(4):201–203, 2006.

[39] W. Piston and M. DeVoto. *Harmony*. Norton, 1987.

[40] R. Plomp and W.J.M. Levelt. Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America*, 38(4):518–560, 1965.

[41] D. Pressnitzer and S. McAdams. Two phase effects in roughness perception. *The Journal of the Acoustical Society of America*, 105(5):2773–82, May 1999.

[42] X. Serra. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition, PhD Thesis*. PhD thesis, University of Standford, 1989.

[43] X. Serra. Musical Sound Modeling with Sinusoids plus Noise. *Computer Music Journal*, pages 1–25, 1997.

[44] J.O. Smith. *Spectral Audio Signal Processing, October 2008 Draft*. `http://-ccrma.stanford.edu/~jos/sasp/`. online book.

[45] J.O. Smith and J.S. Abel. Bark and erb bilinear transforms. *Ieee Transactions On Speech And Audio Processing*, 7(6):697–708, 1999.

[46] H. Stephen and M. Malcolm. On sinusoidal parameter estimation. *Proceedings of the 6th International Conference on Digital Audio Effects (DAFx-03)*, 2003.

[47] P. Van Hengel. Cochlea model - conceptual documentation. Publicado 10/05/2001, Último acceso: 30/01/2012, `http://www.ai.rug.nl/acg/cpsp/docs/cochleaModel.html`.

[48] P.N. Vassilakis. *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance*. PhD thesis, University of California, Los Angeles, CA, 2006.

[49] P.N. Vassilakis, W. Belden, and A. Chicago. SRA : A Web-based Research Tool for Spectral and Roughness Analysis of Sound Signals. *Computing*, (July):11–13, 2007.

[50] Steinberg Virtual and Studio Technology. Software Development Kit. *Development*, pages 1–84.

[51] J. Vos. Commentary on â Calculating Sensory Dissonance : Some Discrepancies Arising from the Models of Kameoka & Kuriyagawa and Hutchinson & Knopoff â by Keith Mashinter. *Society*, 1(3):180–181, 2006.

[52] A.L.C. Wang. *Instantaneous and frequency-warped signal processing techniques for auditory source separation*. PhD thesis, Citeseer, 1994.

[53] G. Zarlino and F. Franceschi. *Istitvtioni harmoniche*. Appresso Francesco Senese, al segno della pace, 1562.