RECONOCIMIENTO AUTOMÁTICO DE MODOS MUSICALES EN VOZ CANTADA

AUTOR: D. ENRIQUE TOMÁS CALDERÓN

TUTOR: D. JUAN FRANCISCO GÓMEZ MENA

TRIBUNAL:

PRESIDENTE: D. JUAN MIGUEL SANTOS SUÁREZ
VOCAL: D. JUAN FRANCISCO GÓMEZ MENA
SECRETARIO: D. JOSÉ FERMÍN PARERA BERMÚDEZ
SUPLENTE:: D. RAMÓN MARTÍNEZ RODRIGUEZ - OSORIO

FECHA DE LECTURA Y DEFENSA:

CALIFICACIÓN:

RECONOCIMIENTO AUTOMÁTICO DE MODOS MUSICALES EN VOZ CANTADA

D. ENRIQUE TOMÁS CALDERÓN

ETSI TELECOMUNICACIÓN UPM MADRID

Como abajo es arriba;
como arriba es abajo.

A Nuria

Resumen:

El Reconocimiento de Modos Musicales es una de las tareas más comunes de clasificación musicológica, especialmente cuando se dedica al estudio de fenómenos musicales muy alejados en forma y construcción de los postulados del mundo occidental.

En la presente memoria se parte de las raíces del problema, se explican las principales características de la percepción de la frecuencia en un entorno musical, y después se abordan las principales hipótesis que guían la construcción de las escalas musicales.

Además, en esta memoria se resumen las principales características de una aplicación de reconocimiento mediante análisis en tiempo real, creada por el autor. Se explican la arquitectura software utilizada y los diferentes problemas a los que el algoritmo debe responder. Por último, se muestra la calidad y robustez de la aplicación para su uso generalizado.

Palabras clave:

Modos musicales, análisis de tono, analizador prosódico, inteligencia artificial, procesado en tiempo real, función autodisimilitud, análisis de espectros, autocorrelación, suavizado de histogramas, construcción de escalas musicales, percepción de la frecuencia.

Glosario:

ADF = Auto Dissimilarity Function

AADF = Adaptative Auto Dissimilarity Function

API = Aplication Program Interface

CENT = corresponde a la relación 2^(1/1200)=1.000578

EWMA = Media Aritmética con Ponderación Exponencial

JND= Número de Diferencias Notorias de Frecuencia

MAPATONE = Programa de generación automática de música tonal y modal (propiedad de Dr. Francisco Javier Sánchez)

MIDI = Musical Instrument Digital Interface

PD = Pure Data

RAMM = Reconocimiento Automático de Modos Musicales

SIFT = Simplified Inverse Filtering

INDICE

1. Introducción al problema del reconocimiento del modo musical.

1.1 Acerca del título del Proyecto Fin de Carrera.

Entendemos por *Reconocimiento de Modos Musicales* como una de las tareas rutinarias que puede realizar un especialista o investigador en Musicología, es decir, un estudioso de la historia y características organológicas de la música con el fin de estudiar y clasificar cualquier manifestación musical. Estas tareas, consisten en un estudio profundo de las particularidades de la estructura compositiva e interpretativa, con el afán de encontrar puntos de encuentro o diferencias entre diferentes culturas. Particularmente, y de forma aproximada, reconocer un *modo musical* consistiría en encontrar la escala musical que utiliza, la estructura que combina las notas de esa escala, y la forma de organizarlas en el tiempo.

Estas manifestaciones culturales pueden ser de origen folklórico, de transmisión popular o culta. Por este tipo de música, entendemos aquella que es propia y originaria de cada región del planeta, aquella que está entroncada con la cultura y forma de ser de los pueblos de cualquier región y de cualquier continente. A menudo, diferentes pueblos han absorbido las manifestaciones de otros más influyentes, y les han ido sumando sus propias formas interpretativas o compositivas, originando de esta forma un conjunto de culturas propias a lo largo de los siglos. Este conjunto, puede verse a alto nivel diferenciando por ejemplo músicas por continentes, o a muy bajo nivel, pudiendo hablar de *micromúsicas*, como por ejemplo músicas propias de comarcas, de pueblos, de barrios, de clubes, etc.

Pues bien, esa tarea rutinaria de algunos musicólogos, se fundamenta en gran parte en el reconocimiento y anotación de las frecuencias que se perciben en la música. Para ello, se viene utilizando demasiado a menudo la propia capacidad de reconocimiento musical mediante el oído y aplicaciones informáticas no dedicadas. Como tal, es un procedimiento totalmente subjetivo y falto de la exactitud necesaria para el análisis. Más aún, cuando la comparación previa entre modos para su clasificación ha de hacerse sobre bases de datos muy extensas, las diferencias pueden ser muy sutiles, para lo cual se hace necesaria la utilización de herramientas de análisis.

Por tanto, el objetivo principal de este trabajo es la investigación en las posibilidades de realizar un *Reconocedor Automático de Modos Musicales*, el cual podrá realizar la tarea de análisis musicológico de forma automática y objetiva, lo cual facilitará el avance en las investigaciones y abrirá las puertas a la posibilidad de comparación exhaustiva y objetiva de resultados entre diferentes estudios e investigadores.

1.2 Los modos musicales y la voz cantada.

El hecho de restringir este trabajo al análisis de la voz cantada y no al de cualquier sonido musical, tiene una doble explicación; por un lado, las características específicas de la voz hacen que el análisis deba plantearse con técnicas diferentes del caso de analizar música instrumental; por otro lado, en casi todas las culturas con conocimientos

musicales no demasiado evolucionados, la principal manifestación musical es expresada mediante la voz, y los instrumentos tienen un carácter poco protagonista.

En todo caso, para no ver esta restricción como un defecto, podemos añadir los siguientes comentarios como puntos de partida en el trabajo:

1. El folklore musical se basa principalmente en la voz, y pocas veces es necesaria la inclusión de instrumentos musicales para dotar de significado global a la música.

2. La utilización de la voz es anterior a la de instrumentos, por tanto, los matices de cada modo podrán reflejarse mejor en música cantada que en música interpretada con instrumentos. Veremos entonces la música instrumental como una evolución de la cantada.

3. Para analizar un modo hay que estimar las frecuencias fundamentales que se utilizan, y para ello:
   - Hay que construir un estimador de frecuencia fundamental robusto

   - Puesto que la prosodia del habla con la que se canta es la que organiza la forma de cantar, hay que analizar las características fonológicas del habla regional, para comprender mejor los rasgos prosódicos diferenciadores.

     Por tanto, es necesario comprender el modo en el que se canta, y hay que comprender también los rasgos fonológicos de su habla, lo cual hace el análisis aún más profundo, pero también más inteligente.

Por todo lo anterior, considero que la investigación en el habla cantada se presenta como una disciplina con mayor interés, por lo cual y de aquí en adelante, en ella nos centraremos únicamente en este trabajo.

1.3 Los modos musicales y los modos de música.

El ser humano manifiesta su forma de ser y actuar en todos los aspectos de la vida cotidiana. Si bien la forma de ver el mundo que le rodea puede ser muy similar a la de otros de sus semejantes, el entorno particular de cada pueblo durante siglos y siglos, ha ido caracterizando unas costumbres, unos idiomas y unas manifestaciones culturales propias y distinguibles de las del resto de los pueblos del mundo. El habla es un ejemplo fundamental de esta diferenciación cultural, la cual se va independizando progresivamente del núcleo idiomático común en muchos casos, adquiriendo matices fonéticos propios, así como nuevos significados para su vocabulario.

En el arte, por ejemplo, tenemos otro ejemplo claro de la importancia que tiene el ámbito cultural que rodea a la persona para la elección de una u otra vía de expresión, y ahora según diferentes regiones, podemos separar diferentes movimientos artísticos según las culturas que las han originado: arte azteca, arte africano, arte musulmán, arte bizantino, etc, con todas sus subclases.

La música no es un caso aislado. La música es una de las manifestaciones que más ha sufrido la dependencia del ambiente cultural en el que se crea. Quizá, es una de las dependencias que hoy en día podemos identificar más claramente, puesto que el ambiente globalizado hace que ya adoptemos ritmos o sonoridades lejanas por su origen, e incluso nos parezcan cercanas y atrayentes. La música ha sufrido además un alto grado de evolución y de especialización, sobre todo en occidente, y en muchos casos una alta difusión.

Sin embargo, la evolución no es sólo propia de la música de las sociedades más ricas, la cual suele ser la de mayor difusión, sino que también afecta a las micromúsicas de las regiones. Este cambio suele ser mucho más lento, y suele estar originado en cambios socio-políticos más que culturales. Tal es el nivel de progreso e independencia musical, que entre dos municipios cercanos de la misma comarca, con historia idéntica y alto grado de dependencia, podemos encontrar diferencias sustanciales en la forma de concebir la interpretación musical de su folklore, de su creación y de su consideración.
Además, si no ha habido cambios grandes en el entorno, podemos encontrar regiones en las que la música se interpreta de la misma forma que hace siglos. Su estudio es muy importante, porque puede hacer que encontremos nuevas hipótesis para los orígenes de la música y los orígenes de las sociedades.

Todos los matices del entorno, que durante siglos pueden ir cambiando y acentuándose, pasan a formar parte del patrimonio de su cultura intrínseca, así como de su forma de comunicación y educación.

Hemos de preguntarnos por el origen de estas diferencias, pues es imposible que tal diversidad de músicas haya sido desarrollada de la casualidad.
Lo primero que hay que tener en cuenta es que la primera manifestación musical humana no pudo ser otra más que el canto, como una evolución del habla, si es que ya había sido formada. Esta hipótesis es apoyada por Terhardt. La voz, por tanto, tuvo que ser el primer instrumento musical de la historia, pues el conocimiento de la estructura musical necesaria para construir instrumentos especializados, no llegó hasta bien avanzada la cultura del ser humano. De hecho, la creación de un sistema de anotación musical, aunque muy básico, no llegó hasta la época del Imperio Mesopotámico, y siempre estaba unida a la voz. Lo que sí que se conoce, es que la voz era acompañada por otros instrumentos de viento muy primitivos, e instrumentos de percusión construidos con conchas de crustáceos, muchas veces para actos de naturaleza religiosa.
Por tanto, la música tuvo que estar grandemente influenciada por el habla de cada pueblo, puesto que al cantar a la vez se están diciendo palabras, y sólo aquella música que servía para cantar en un idioma determinado podía utilizarse, y el resto de posibilidades, sin utilidad por el momento, se rechazaba.
De esta forma, como el habla de cada región es diferente, también lo será el modo de cantar. Si sólo se podía cantar, la creación musical girará entorno a un sistema que funcione bien con el idioma y con la forma de decir las palabras. Por otro lado, es posible encontrar diferentes variedades fonológicas dentro de, por ejemplo, una misma comarca, y por tanto es normal que encontremos también diferentes variedades musicales.

Con el paso de los siglos y el avance en la investigación musical, cada pueblo va adoptando unas estructuras musicales propias que los individuos reconocen como propias, defendiéndolas y transmitiéndolas por transmisión oral. El desarrollo de estas ideas musicales, termina creando sofisticados sistemas de composición en muchos casos, pero que son adoptados por la gente con toda naturalidad.
Bien, pues estos sistemas de composición musical pueden ser los modos musicales. En realidad, un modo musical se compone de un conjunto de escalas musicales propias, un sistema de afinación y un conjunto de valores rítmicos, fruto de que la música se compone esencialmente de la combinación de unas alturas determinadas, en un orden establecido y sobre unos valores rítmicos regulares.

Por escala musical comprendemos un conjunto de notas consecutivas que mantienen una relación musical propia y unificada. Las escalas son utilizadas para la composición en gran parte de la música de muchas regiones, dejando su sonoridad propia. Hablando en términos de ingeniería, constituyen todo el espectro posible de frecuencias utilizables en una obra.

El sistema de afinación, está muy relacionado con la escala musical, puesto que para poder repetir siempre las mismas frecuencias, se han de establecer unas normas de distancias entre las notas. Cada cultura ha podido ir identificando ciertas distancias entre notas, y las han ido utilizando, pero no tienen por qué coincidir con las de otras regiones o culturas del planeta. Tampoco esta claro que sea necesario un sistema de afinación, puesto que la mera distinción perceptiva entre consonancia y disonancia, como luego veremos, puede ser el principio de construcción de la escala musical.

Por último, el origen de los valores rítmicos es otro enigma sobre el cual poco podemos aventurar. La explicación clásica suele estar asociada con la teoría de los biorritmos humanos o de la naturaleza circundante, los cuales infringirían una influencia rítmica sobre nuestros actos. La combinación de los mismos a gusto del intérprete, va dando lugar a diferentes opciones de figuración rítmica, también independientes por regiones. La concepción del ritmo, es algo muy particular de las sociedades: los ritmos musicales de una cultura no suelen ser compartidos por las de otras, aunque sí admirados, o estudiados. Sin embargo, una persona que crece en una determinada cultura y se sumerge en sus ritmos, ha de derrochar en general mucho esfuerzo personal para aprender los de otras culturas si estos son realmente distintos en concepción. Como ejemplos, se puede destacar la admiración de los compositores del siglo XX (como Ligeti, Stockhaussen, etc) en Occidente por los ritmos africanos, o los problemas que tiene un intérprete europeo por comprender los ritmos caribeños.

Pues bien, una vez explicado lo que constituye un modo musical, este concepto subyace a la teoría y se manifiesta en la percepción que tenemos de la música que escuchamos. Como ejemplo, podemos identificar la música del País Vasco y diferenciarla de la de los pueblos de África central, y a la vez podemos comparar sus ritmos con los de los salmos cristianos del siglo III , o la variación prosódica con los *ragá* de la India, aunque nunca hayamos leído ningún libro sobre cualquiera de sus teorías.
El hecho de por qué una música de la que desconocemos el autor, su base de creación e interpretación podemos clasificarla, diferenciarla, o localizarla geográficamente, tiene mucho que ver con los timbres de los instrumentos que la interpretan y por la forma de cantar y alcanzar las notas, pero fundamentalmente con los modos musicales que emplean. El hecho objetivo es que reconocemos sonoridades, escalas, ritmos, de forma

inconsciente en el momento de la escucha, y que en ese momento, nos sentimos influenciados: un modo musical nos puede provocar sentimientos como rechazo, tristeza, melancolía, agresividad, fortaleza, etc. Aparece por tanto, cierto componente filosófico que se mezcla con las raíces humanas, estudiado desde la antigüedad clásica (es el caso de los *ethos* griegos) y desde los orígenes de las sociedades orientales (como el caso de la elección de escalas musicales diferentes en India para cada momento del día).

Este es el momento, por tanto, de buscar los parecidos entre la voz hablada y la voz cantada. A modo de resumen explicativo, y utilizando las nociones sobre modos musicales que hemos ofrecido anteriormente, podemos decir que:

- El habla hace uso de cambios en frecuencia a nivel de sílaba y palabra para dar énfasis o acentuar sílabas, y que se repetirían monótonamente en ausencia de rasgos prosódicos de más alto nivel. La voz cantada utiliza un conjunto de alturas de frecuencia para dar sentido al discurso, que es el sistema de afinación.

- El habla progresa en frecuencia de forma diferente según el tipo de oración, de forma que el oyente es capaz de identificar si se trata de una afirmación, de una pregunta, etc y que son parte de la base de la prosodia. La voz cantada organiza las alturas y crea frases reconocibles, crea tensión y distensión con cambios en la amplitud y frecuencia, y organiza el discurso en más que sucesiones de notas, que es el sistema de escalas musicales.

- El habla progresa con un ritmo propio, que se va incrementando o ralentizando en ciertos puntos, con unas reglas adoptadas por todos los hablantes. En la voz cantada, el ritmo puede venir impuesto únicamente por el ritmo del texto, como ocurre en los coros de la antigua Grecia, o ser modificado con la inclusión de instrumentos de percusión, pero nunca se permite el desplazamiento de los acentos o del énfasis en las sílabas fundamentales o importantes.

Por tanto, un musicólogo que se plantea el reto de buscar analogías entre el flamenco andaluz y la música de Sri Lanka, puede estudiar sus teorías de composición y sus modos, y descubrir que hay grandes parecidos (como actualmente se sabe). Su tarea ha de abordarla de forma objetiva, y no subjetivamente como suele hacerse a menudo actualmente, y es ahí donde recae la importancia de buscar un método de estudio científico que no se base en la percepción de quien realiza el estudio, sino en la objetividad.

A continuación, vamos a estudiar algunos aspectos perceptivos del ser humano, para así afrontar la tarea de estimar la frecuencia fundamental desde un punto de vista más profundo.

2. Percepción de la frecuencia.

Puesto que en este Proyecto Fin de Carrera el objetivo principal es encontrar un método para reconocer la frecuencia de la voz cantada, era necesario un pequeño esfuerzo investigador en cuestiones referentes a la psicoacústica especial del ser humano, especialmente en los apartados referentes a cómo tenemos constancia de las relaciones interválicas, de la consonancia y disonancia entre frecuencias, y de cómo este problema se ha abordado desde muy antiguo, dando lugar a teorías de afinación y a las escalas musicales.

2.1 Fisiología y Percepción.

Se sabe que Pitágoras, en siglo VI aC, ya conocía que los sonidos producidos por varias cuerdas vibrantes sonaban consonantes cuando las relaciones entre sus longitudes mantenían relaciones sencillas formadas por números enteros. Sin embargo, no pudo completar su demostración del fenómeno porque aún no conocía lo suficiente acerca del contenido de los armónicos de cada sonido.

Más tarde hacia 1638, Galileo descubre que las consonancias "agradables" se basan en una cierta regularidad en la vibración del aire, de forma que, y como escribe: "los pulsos producidos por dos tonos, al mismo tiempo, deben ser conmensurables en número, de forma que no mantengan al tímpano en un estado de tormento perpetuo".

Hasta que en el siglo XIX fueron identificados los diferentes armónicos que componen el sonido, no se pudo obtener una explicación adecuada respecto al origen de la consonancia y la disonancia.

Para comprender bien las conclusiones sobre consonancia y disonancia, hemos de realizar previamente un estudio sobre el oído y la percepción del sonido por parte del sistema nervioso humano, para más tarde adentrarnos en la forma de agrupar los sonidos.

2.1.1 Acústica fisiológica.

Cuando hay un fenómeno acústico, la onda de presión sonora viaja por el aire, sufriendo atenuaciones en amplitud, reflexiones y refracciones. Si al final es captada por un ser humano, incidirá en su pabellón auditivo y se transmitirá por el oído externo a través del canal o conducto auditivo hasta incidir en el tímpano. Se sabe que, al pasar por el conducto auditivo, se generan ciertas resonancias que harán más sensible al tímpano para recibir la onda de presión. Se puede decir que es una primera cavidad resonante, pero eso sí, con una ganancia muy pequeña.

El tímpano es la fina membrana que separa el oído externo del oído medio, donde se aloja la cadena de huesecillos, y a la que está unida solidariamente. Por una parte actúa como transductor de la vibración que le transmite la onda de presión sonora incidente

desde el exterior, y al mismo tiempo que impide que cualquier partícula la penetre en el oído medio.

El oído medio está herméticamente cerrado del exterior, excepto por la trompa de Eustaquio, comunicada con la garganta, que se abre durante breves lapsos de tiempo para compensar excesos o defectos de presión y servir como conducto de drenaje.

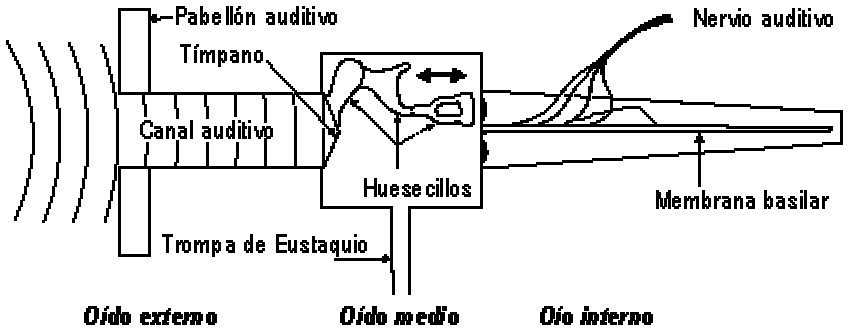


figura 2.1 Esquema del oído.

En el oído medio, encontramos los huesecillos: martillo, yunque y estribo. Actúan como un sistema de palancas que comunica la onda mecánica desde el tímpano, sobre el que se apoya el martillo, hasta la cóclea, por medio de la ventana oval, sobre la cual se apoya el estribo. Estos huesecillos realizan una importante labor de *adaptación de impedancias* que permite que la presión que imprime el estribo en la membrana oval, sea 30 veces mayor que la ejercida sobre el tímpano.

La ventana oval supone el punto de entrada de esta señal en el oído interno. El único órgano responsable de la audición el oído interno es la cóclea, donde se llevará a cabo la conversión de la señal analógica de entrada, en miles de señales nerviosas que viajan por el nervio auditivo hacia el cerebro. Estas señales, que por la naturaleza neuronal son de tipo *digital* (trenes de pulsos), tienen que transmitir toda la información de interés que haya en la señal original, y además codificarla adecuadamente para que el cerebro pueda interpretarla.

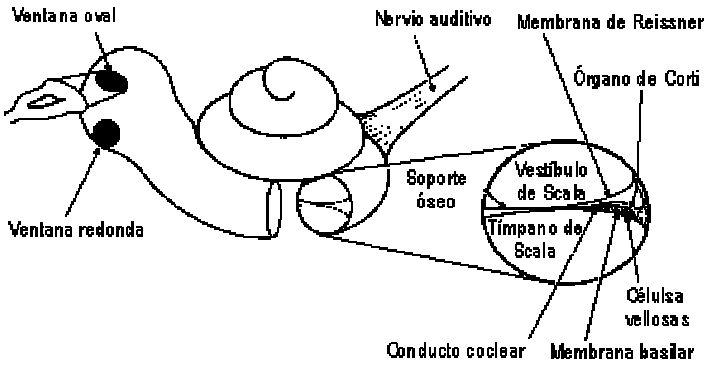


figura 2.2 Esquema del oído interno.

La cóclea es una estructura con forma de tubo cónico alargado, que se encuentra enrollada sobre sí misma en forma de espiral. Observando una sección transversal se aprecian tres diferentes cámaras que la recorren en toda su longitud: dos canales y el conducto coclear. La cóclea está llena de líquido y rodeada por paredes óseas rígidas. En los canales hay líquidos de diferentes densidades, que se encuentran separados por dos membranas, una de ellas, la membrana de Reissner [ver ilustración] es extraordinariamente delgada. Apoyados en la membrana basilar se encuentra el complejo y delicado órgano de Corti, que contiene varias filas de diminutas células vellosas a las cuales se conectan las fibras nerviosas. Cada fila de células vellosas contiene unas 7000 células, habiendo un total de 24.000 células en varias filas. Cada célula vellosa posee numerosos cilios, que se doblan cuando la membrana basilar responde a un sonido, desencadenando una señal nerviosa en el nervio auditivo.

Georg von Békésy recibió el premio Nobel de Fisiologia y Medicina en 1961 por su descubrimiento del funcionamiento de la cóclea y su membrana basilar. Realizó numerosos experimentos con cócleas de animales y cadáveres, y construyó modelos que imitaban su funcionamiento con el objetivo de desentrañar el patrón de vibración de la membrana basilar ante las diferentes frecuencias y amplitudes de las señales entrantes.

Para comprender cómo vibra la membrana basilar imaginamos la cóclea desenrollada, con forma de cilindro estrecho dividido en dos secciones por la membrana basilar. En el extremo más grande del cilindro, se encuentran las ventanas oval y redonda, cada una de ellas cerrada por una membrana. En el otro extremo de la membrana basilar hay un pequeño orificio denominado *helicotrema* que comunica las dos secciones. La membrana basilar acaba a poca distancia del extremo del cilindro, con lo cual el fluido puede transmitir ondas de presión de vuelta desde el final de la membrana.

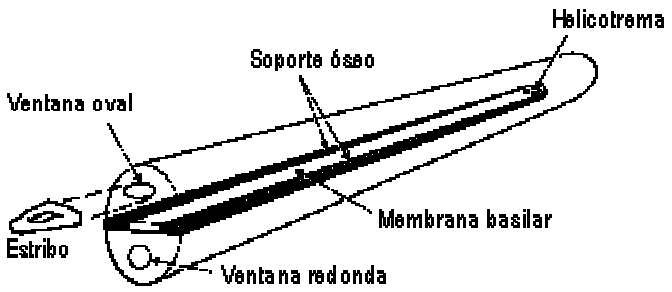


figura 2.3 Esquema de la cóclea.

Cuando el estribo vibra contra la ventana oval, se producen oscilaciones en la membrana basilar según la vibración viaja a través del interior de la cóclea. Los tonos agudos hacen vibrar la membrana basilar cerca de las ventanas (donde es delgada y rígida), mientras que los tonos graves hacen vibrar la membrana basilar cerca del *helicotrema*, donde es más flácida.

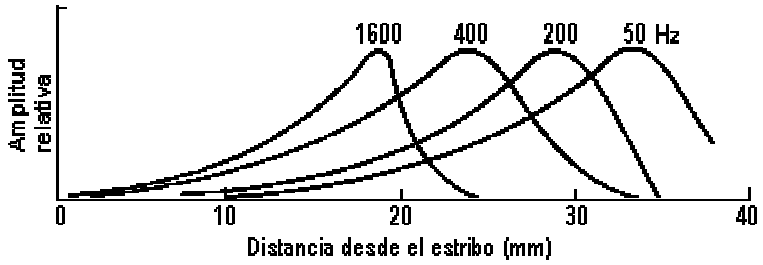


figura 2.4 Vibración de la cóclea.

De este modo, en la cóclea tiene lugar un análisis espectral inicial. La conversión de las vibraciones mecánicas de la membrana basilar en impulsos eléctricos del nervio auditivo se lleva a cabo en el ya mencionado Órgano de Corti. En función de la frecuencia de la señal la vibración se encontrará localizada a lo largo de la membrana basilar excitando unas u otras fibras nerviosas, que corresponderán a las diferentes frecuencias.

Asimismo, en función de la amplitud de la vibración, más fibras nerviosas serán estimuladas.

Hermann von Helmholtz (1821-1894) identificó que el patrón de vibración en el interior de la cóclea hacía que la respuesta de cada fibra nerviosa correspondiese a la de un resonador selectivo, sintonizado a diferentes frecuencias en función de la posición que ocupase cada terminación nerviosa a lo largo de la membrana basilar. Es lo que se conoce como *teoría del lugar*.

Los experimentos posteriores de Georg von Békésy demostraron que esta aproximación es cierta, pero que las bandas de filtrado no son lo suficientemente estrechas como para justificar la precisa percepción del tono del oído humano, aunque, como veremos en el siguiente apartado, sí son de vital importancia para permitir que cada fibra nerviosa se capaz de transmitir correctamente una señal restringiendo su actuación a una banda de frecuencias, que denominaremos *banda crítica*.

2.2 Psicoacústica

Como hemos mencionado anteriormente, una célula nerviosa es excitada cuando la vibración de la membrana basilar supera un cierto umbral. Existen células con diferentes umbrales para permitir un gran rango dinámico de respuesta al oído. Asimismo, la señal que viaja por cada fibra nerviosa individual del nervio auditivo, aún tratándose de un tren de pulsos, posee la misma periodicidad que la envolvente temporal de la señal filtrada en banda correspondiente a la vibración de la membrana basilar allí donde esta su terminación nerviosa. Estas consideraciones hay que basarlas en conocimientos neurosensoriales, y sobre todo en la teoría del *Potencial de Acción* sobre células sensoriales. Un ejemplo de esto se observa en la siguiente ilustración, donde se ve cómo se genera esta información.

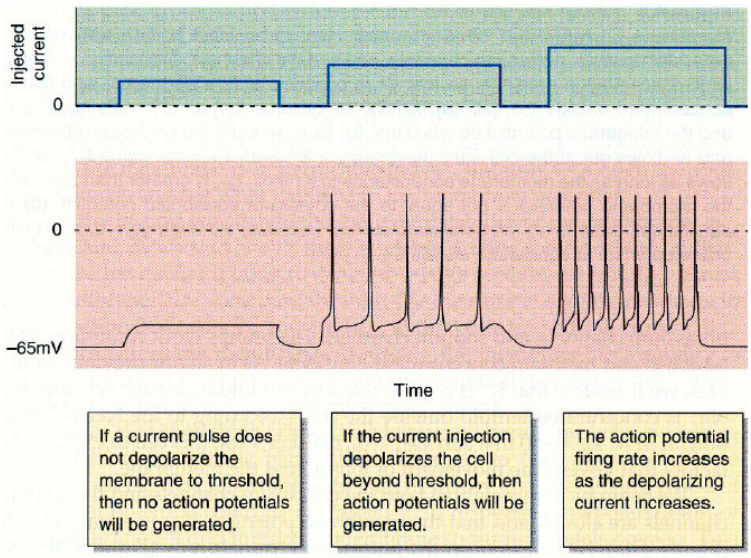


figura 2.5. Principio de transmisión sensorial.

La *teoria de la periodicidad* afirma que el cerebro es capaz de decodificar los patrones temporales de cada fibra nerviosa, analizando la autocorrelación dentro de cada señal individual, así como la correlación cruzada entre fibras correspondientes a distintas bandas críticas buscando patrones temporales de vibración cuya periodicidad está relacionada por números enteros. Según esta teoría es deseable que cada fibra reciba sólo una señal armónica, pues si el patrón de vibración de cada fibra nerviosa fuese demasiado complicado, conteniendo componentes de diferentes periodicidades, las búsquedas de correlación serían demasiado complicadas. Así pues, y hablando en términos de procesado de señal, para el cerebro surge la necesidad de un filtrado en bandas (bandas críticas como veremos) previo de cada señal nerviosa para luego poder relacionarlas, que se lleva a cabo en la cóclea, como hemos visto ya en el apartado anterior.

Llamamos *Banda crítica* a la banda de frecuencias que excita cada terminación nerviosa. Mediante estudios fisiológicos de cócleas, así como de experimentos perceptivos, sabemos que el ancho de banda de estos filtros paso banda centrados en cada terminación nerviosa es tal que se abarcarían la totalidad del espectro audible con 24 de ellos. Pero no debemos olvidar que en realidad no existen 24 filtros, sino uno continuo, pues el filtrado es el resultado de las propiedades mecánicas de la cóclea sobre cada punto de la membrana basilar.

El ancho de banda crítica es función de la frecuencia, como se observa en el gráfico. En frecuencias, aproximadamente, y anticipando la notación de intervalos musicales que más tarde veremos, un intervalo de tercera menor corresponde a 1/5 de la frecuencia central de cada banda, un tono entero a 1/8, y un semitono a 1/16 de la frecuencia central.

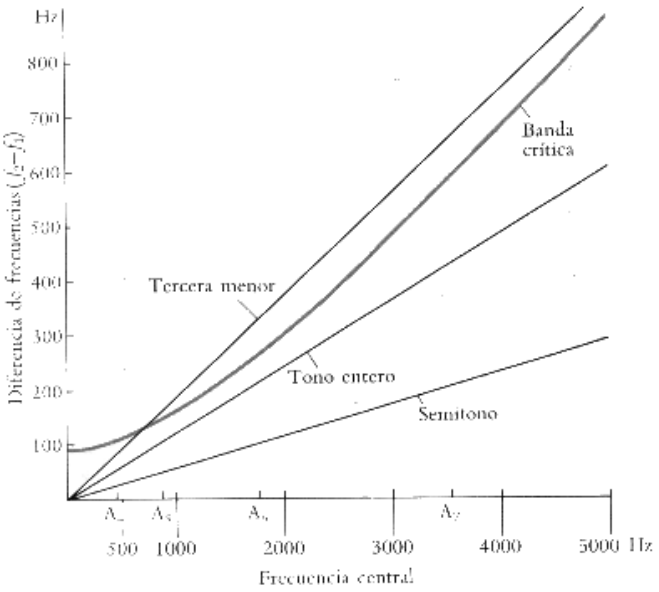


figura 2.6. Intervalos musicales en las bandas críticas

2.2.1. Percepción de sonidos simultáneos: batimientos y aspereza

A continuación se tratarán los fenómenos perceptivos que ocurren como resultado la escucha de dos sonidos simultáneos, o lo suficientemente próximos como para que el cerebro los relacione.

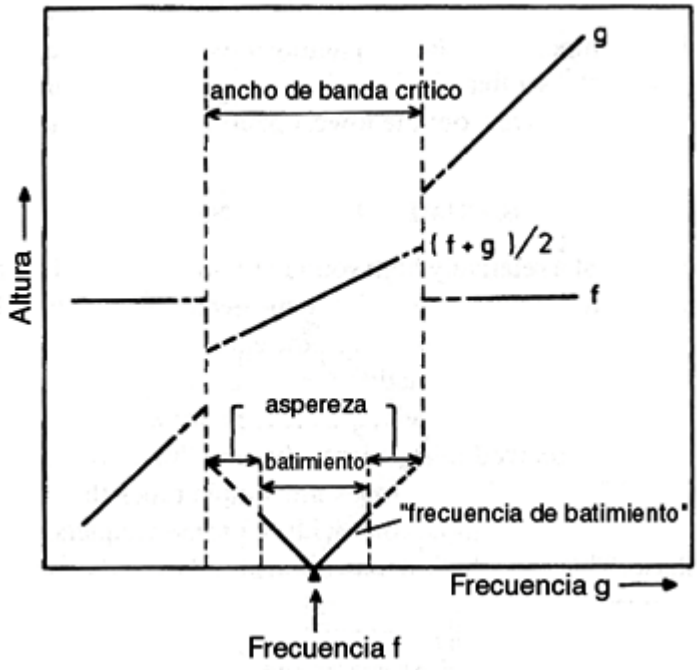


figura 2.7 Ancho de banda de batimiento

Consideraremos primero el caso de dos sonidos simples, sin espectro complicado, que suenan simultáneos. Pueden distinguirse varias condiciones dependiendo de la diferencia de frecuencia (fig. 2.7).

-Si los sonidos tienen frecuencias iguales, éstos se fusionan en un sonido, cuya intensidad depende de la relación de fase entre los dos sonidos primarios.

-Si los dos sonidos primarios difieren en sus frecuencias, el resultado es una señal con amplitud periódica y variaciones de frecuencia, con una frecuencia igual a la diferencia de frecuencia. Las variaciones de amplitud pueden ser considerables y resultar en una intensidad fluctuante y sonoridad percibida.

Estas fluctuaciones de sonoridad son llamadas *batimientos*, si es que pueden ser percibidos por el oído, lo cual ocurre si su frecuencia es menor a 20 Hz.

Por ejemplo, un estímulo igual a la suma de dos sonidos simples con amplitudes iguales y frecuencias f y g:

$$p(t) = \text{sen } 2\pi ft$$

puede ser descrita como:

$$p(t) = 2 \cos 2\pi \, 1/2 \, (g - f)t \cdot 2 \cos 2\pi \, 1/2 \, (g + f)t$$

Esta es una señal con una frecuencia que es el promedio de frecuencias primarias originales, y una amplitud que fluctúa lentamente con una frecuencia de batimiento de g - f Hz (fig. 2.8). La variación de amplitud es menor si los dos sonidos primarios tienen amplitudes diferentes.

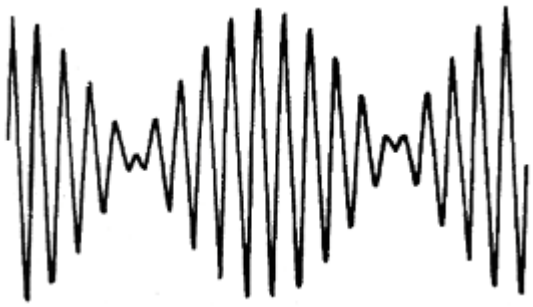


figura 2.8 Modulación en amplitud

Cuando la diferencia de frecuencia es mayor de 20 Hz, el oído ya no es capaz de seguir las rápidas fluctuaciones individualmente. En lugar de la sensación de sonoridad fluctuante, hay una especie de pulsación que llamaremos *aspereza*.

En la práctica musical, como veremos después, los batimientos pueden ocurrir cuando hay armónicos que no coinciden al tener intervalos consonantes desafinados. Si las

frecuencias fundamentales de los sonidos de una octava o quinta difieren en algo de la relación teórica (1:2, 2:3), habrá armónicos que difieren algo en su frecuencia lo cual causará batimientos. Estos batimientos juegan un rol muy importante en el proceso de afinación de los instrumentos.

En general, no se han hecho investigaciones psicoacústicas sobre intervalos desafinados de sonidos complejos, pero en gran medida las observaciones hechas con respecto a los sonidos simples, se aplican para los complejos.

2.2.2 Efectos no lineales. Sonidos diferenciales.

Es difícil encontrar un sistema con una respuesta puramente lineal. Para Helmholtz, la respuesta del oído no era lineal, de modo que el oído fortalecía o reconstruía el armónico fundamental de un sonido a partir de sus armónicos superiores. Como hemos visto en párrafos anteriores, el oído tiene una respuesta bastante lineal a la hora de combinar frecuencias, de modo que no se lleva a cabo el fenómeno descrito por Helmholtz con suficiente intensidad. Sin embargo sí que aparecen otros fenómenos más tenues, pero importantes.

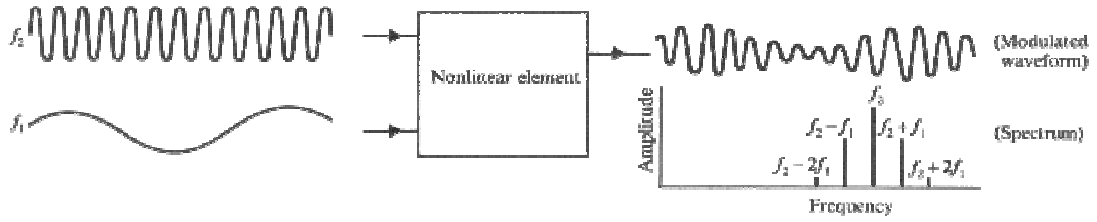


figura 2.9 Esquema de respuesta no lineal del oído.

La *distorsión interarmónica* es uno de los fenómenos más importantes: ¿Cómo afecta la presencia de un armónico en el resto de armónicos?. Podemos analizar la respuesta de un sistema mediante su desarrollo en serie de Taylor. Resulta muy sencillo de estudiar el caso de un sistema con una componente cuadrática en su respuesta, al cual introducimos dos funciones armónicas.

$$\left\{\begin{array}{l}\cos(f_1 t)+\cos(f_2 t)= \\ \frac{1}{2}\left(e^{jf_1 t}+e^{-jf_1 t}+e^{jf_2 t}+e^{-jf_2 t}\right)\end{array}\right\} \xrightarrow{ax+bx^2} \left\{\begin{array}{l} a\left[\underbrace{\frac{1}{2}\left(e^{jf_1 t}+e^{-jf_1 t}\right)}_{\cos(f_1)}+\underbrace{\frac{1}{2}\left(e^{jf_2 t}+e^{-jf_2 t}\right)}_{\cos(f_2)}\right] \\ +b\left[\underbrace{\frac{e^{j(f_1+f_2)t}+e^{-j(f_1+f_2)t}}{2\cos(f_1+f_2)}}+\underbrace{\frac{e^{j(f_1-f_2)t}+e^{-j(f_1-f_2)t}}{2\cos(f_1-f_2)}}\right. \\ \left.+\underbrace{\frac{1}{2}\left(e^{j2f_1 t}+e^{-j2f_1 t}\right)}_{\cos(2f_1)}+\underbrace{\frac{1}{2}\left(e^{j2f_2 t}+e^{-j2f_2 t}\right)}_{\cos(2f_2)}+4\right]\end{array}\right\}$$

figura 2.10 Resultado de la distorsión interarmónica

Como se puede observar en la fórmula anterior y se puede comprobar escuchando una señal de estas características, el oído genera términos de diferencia así como armónicos superiores que antes no existían, y por tanto, es la prueba de que el oído no es un sistema lineal. Predominan los tonos de diferencia cuadráticos y cúbicos, debido a los términos de orden 2 y 3 en el desarrollo de Taylor de la respuesta del oído.

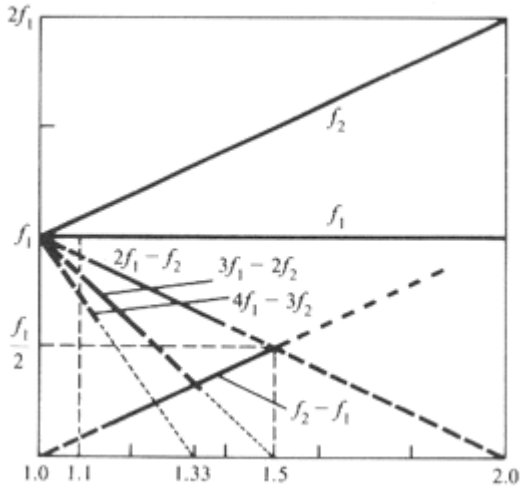


figura 2.11 Mapa de frecuencias generadas por la distorsión

Dos sonidos simples a un nivel de presión sonora relativamente alta y con una diferencia de frecuencia no demasiado amplia puede dar lugar a la percepción de sonidos que podemos llamar diferenciales. Como hemos visto, estos sonidos aparecen en el oído como un producto de características de transmisión no lineal. Los diferenciales no están presentes en la señal sonora, sin embargo son percibidos como si lo estuvieran: el oído no puede distinguir entre sonidos "reales" (presentes en el estímulo) y aquellos que no lo son (diferenciales). Los diferenciales son sonidos simples que pueden ser anulados sumando un sonido simple "real" con la misma frecuencia y amplitud pero con fase opuesta.

Investigaciones psicoacústicas sobre sonidos diferenciales han mostrado que las alturas de los diferenciales coinciden con las frecuencias predichas por transmisión no lineal. Sin embargo, la correspondencia entre la amplitud relativa predicha y la sonoridad subjetiva medida no es nada perfecta. El fenómeno de los diferenciales es más complicado de lo que puede describirse con una simple fórmula.

A pesar de que los diferenciales fueron descubiertos por músicos en un contexto musical (Tartini y Sorge en el siglo XVIII), su significación musical no es muy alta. Pueden ser fácilmente evocados tocando sonidos fuertes en el registro alto en dos flautas o dobles cuerdas en el violín. En una situación auditiva normal, su sonoridad es demasiado débil como para llamar la atención, y en muchos casos aparecen enmascarados por los sonidos de los instrumentos más graves. Algunos maestros de

violín (siguiendo a Tartini) los utilizaban como herramienta de control de afinación de intervalos de dobles cuerdas.

Los tonos de diferencia también son responsables de la aparición de muy tenues batidos cuando se escuchan intervalos de quinta (3/2) y octava (2/1) ligeramente desafinados. Algunos compositores en el siglo XX, como E. Varese, han aprovechado los tonos de diferencia para construir melodías que suenan en registros mas graves que los que pueden alcanzar los intérpretes. Un ejemplo lo vemos en la siguiente figura.

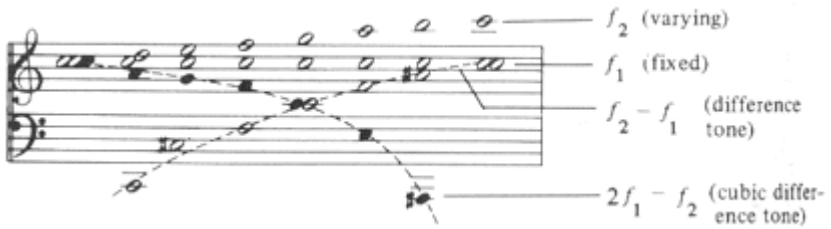


figura 2.12 Melodía creada a partir de una diferencia de tonos

2.3 La percepción del tono o altura.

La altura o tono es la propiedad más característica de los sonidos, tanto con bajo contenido espectral como aquellos con contenido complejo. Los sistemas de alturas se encuentran entre los más elaborados e intrincados jamás desarrollados tanto en la cultura occidental como no occidental. La altura tiene relación con la frecuencia de un sonido simple y con la frecuencia fundamental de un sonido complejo. La frecuencia de un sonido es una propiedad cuya producción puede a menudo controlarse, y se mantiene durante su propagación hacia los oídos del oyente.

En lo que nos respecta, la altura puede describirse como un atributo unidimensional, es decir que todos los sonidos pueden ser ordenados a lo largo de una sola escala con respecto a la altura. Los extremos de esta escala son grave (sonidos con frecuencia baja) y agudo (sonidos con frecuencia alta). A veces, puede dificultarse la tarea de comparar la altura de dos sonidos distintos por factores tales como la diferencia tímbrica entre ellos o el componente de ruido en cada uno. Hay varias escalas subjetivas de altura:

1. La escala mel. Un sonido simple de 1000 Hz tiene una altura definida de 1000 mels. La altura en mels de otros sonidos con otra frecuencia debe ser determinada por experimentos de escalado comparativo. Un sonido con una altura que dobla subjetivamente a la de 1000 Hz, es de 2000 mels; la "altura media" serán 500 mels, etc. Ya que el significado subjetivo de "altura el doble de aguda" o "altura el doble de grave" es invariablemente ambiguo, la escala mel es poco confiable: no es frecuentemente utilizada.

2. La escala de altura musical (C 1, C4, es decir Do 1, Do 4, etc.). Estas indicaciones son utilizables tan sólo en situaciones musicales.

3. La escala de frecuencia física en Hz. En literatura psicoacústica la altura de un sonido es a menudo indicado por su frecuencia o, en el caso de sonidos complejos, por su frecuencia fundamental. Dado el tipo de

correspondencia entre frecuencia y altura, la frecuencia es una indicación aproximada de nuestra percepción de altura. Debe notarse, sin embargo, que nuestra percepción opera, más o menos, de acuerdo a una escala de frecuencia logarítmica.

La notación musical occidental, basada en el pentagrama, puede comprenderse como si de un papel milimetrado con el eje de ordenadas (altura del sonido, frecuencia) en escala logarítmica se tratara. La separación entre líneas consecutivas corresponde bien a intervalos de tres semitonos (tercera menor), o de cuatro semitonos (tercera mayor). El eje de abscisas representa el tiempo.

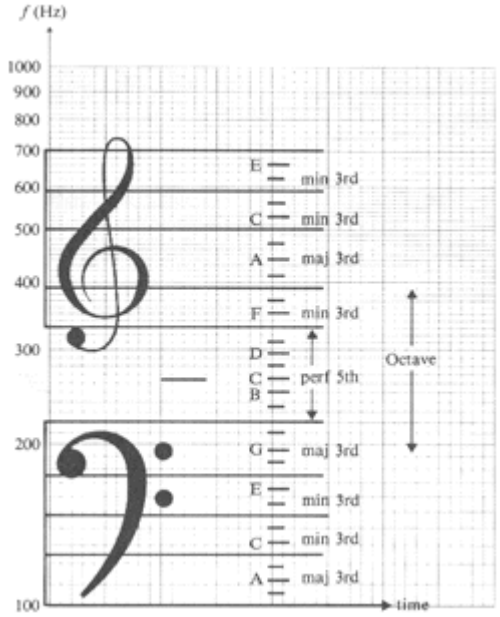


figura 2.13 Analogía logarítmica del pentagrama

Cada una de las líneas o espacios corresponde a una de las teclas blancas del piano, mientras que para designar las teclas negras se utilizan alteraciones, a saber: sostenido # (subir un semitono), bemol *b* (bajar un semitono).

Nota: En el sistema de notación anglosajona las notas se llaman así	la	si	do	re	mi	fa	sol
	A	B	C	D	E	F	G

La altura en su sentido musical tiene un rango de alrededor de 20 a 5000 Hz, más o menos el rango de las fundamentales de las cuerdas de un piano o los tubos de un órgano. Sonidos con frecuencias más altas son audibles pero sin una sensación definida de altura. Sonidos bajos en el rango de los 10 a 50 Hz pueden describirse como

pulsaciones (rattling sound). La transición de la percepción de las pulsaciones a una verdadera sensación de altura, es gradual. La altura puede ser percibida después de que algunos períodos de la onda sonora fueron expuestos al oído.

Los sonidos simples tienen alturas definidas que pueden ser indicadas por medio de su frecuencia. Estas frecuencias pueden servir como frecuencias de referencia para las alturas de sonidos complejos. La sensación de altura en el caso de sonidos complejos es más difícil de entender que en el caso de los sonidos simples. Además, los primeros cinco a siete armónicos de un sonido complejo pueden ser distinguidos individualmente si la atención del oyente es alertada sobre su posible presencia. De todos modos, en la práctica musical, el sonido complejo es caracterizado por una sola altura, la altura del primer parcial. Algunos experimentos han demostrado que la altura de un sonido complejo con una frecuencia fundamental "f0" es algo más grave que la de una onda sinusoidal (sonido simple) con la misma frecuencia. La existencia de la altura grave de un sonido complejo genera dos preguntas:

1) ¿por qué el total de parciales de un sonido complejo es percibido como una sola altura?

2) ¿por qué dicha altura es la propia de la frecuencia fundamental (primer parcial)?

La primera pregunta puede ser contestada haciendo referencia a la *teoría de la Gestalt*. Una explicación basada en dicha teoría puede ser formulada de la siguiente manera:

Los diferentes parciales de un sonido complejo son siempre presentados simultáneamente. Nos familiarizamos con los sonidos complejos de las señales del habla a una edad muy temprana (tanto de nuestra propia habla como de la de los demás). Podemos decir que no sería eficiente percibirlos separadamente. Todos los componentes indican una misma fuente y significado, de forma que percibirlos como unidad da una idea más simple del entorno que la percepción por separado. Esta forma de percepción debe ser vista como un proceso de aprendizaje perceptivo. La psicología de la *Gestalt* ha formulado varias leyes que describen la percepción de estímulos sensoriales complejos. La percepción de la frecuencia más baja de los sonidos complejos puede ser clasificado dentro de la categoría de la "ley del destino común". Los armónicos de un sonido complejo presentan "destino común".

La segunda pregunta también puede ser contestada con la ayuda de un proceso de aprendizaje dirigido hacia la eficiencia perceptiva. La periodicidad de un sonido complejo es la característica más constante en su composición. Las amplitudes de los parciales son objeto de variación causada por la reflexión selectiva, absorción, el pasaje a través de objetos, etc. El enmascaramiento también puede oscurecer ciertos parciales. La periodicidad, sin embargo, es un factor muy estable y constante de los sonidos complejos. La periodicidad de un sonido complejo es, a la vez, la periodicidad del primer parcial del sonido. La percepción de sonidos complejos puede ser vista como un proceso de reconocimiento de modelos. Sin embargo, la presencia de una serie completa de armónicos no es una condición necesaria para el éxito del proceso de reconocimiento de la altura. Es suficiente que al menos un par de armónicos adyacentes estén presentes para que pueda determinarse la periodicidad. Por tanto, es concebible la existencia de un proceso de aprendizaje perceptivo que haga posible el reconocimiento de la periodicidad fundamental a partir de un número limitado de parciales armónicos. Las

teorías de reconocimiento de modelos y su aplicación a la percepción de la altura grave son relativamente recientes, posiblemente pase algún tiempo antes de que las preguntas sobre la llamada altura grave de los sonidos complejos sean contestadas.

La literatura clásica sobre la percepción del sonido abunda en teorías basadas en la idea de von Helmholtz (1863), que explica que la altura grave de un sonido complejo está basada en la fuerza relativa del armónico fundamental. Los armónicos más altos se supone que simplemente influyen sobre el timbre de los sonidos, careciendo de suficiente fuerza como para afectar la altura. Sin embargo, la percepción de la altura grave también ocurre cuando el primer armónico no está presente en el estímulo sonoro. Esto ya había sido observado por Seebeck (1841) y puesto ante la atención de los psicoacústicos modernos por Schouten (1938). Estas observaciones llevaron a Schouten a formular la teoría de la periodicidad de la altura. En esta teoría, la altura deriva de la periodicidad en forma de onda de los armónicos más altos del estímulo, del residuo. Esta periodicidad no cambia si se quita un armónico. Con esta teoría las observaciones de Seebeck y Schouten sobre los sonidos sin armónicos fundamentales podían ser explicadas.

A mediados del siglo XVIII, A. Seebeck realizó una serie de experimentos sobre la percepción del tono que produjeron sorprendentes resultados. Como fuente de sonido, Seebeck utilizó una sirena consistente en un disco rotatorio con orificios periódicamente espaciados por los que atravesaba una ráfaga de aire. Cuando los orificios estaba regularmente distribuidos (a), la sirena producía un sonido con un tono muy bien definido, correspondiente al periodo entre ráfagas. Duplicando el número de orificios (b) el tono se elevaba exactamente una octava. Sin embargo, utilizando un disco con orificios espaciados distancias t1 y t2 (c) se produjo un resultado inesperado: el tono percibido era el mismo que en el caso (a), aunque cambiaba el timbre, la calidad del sonido.

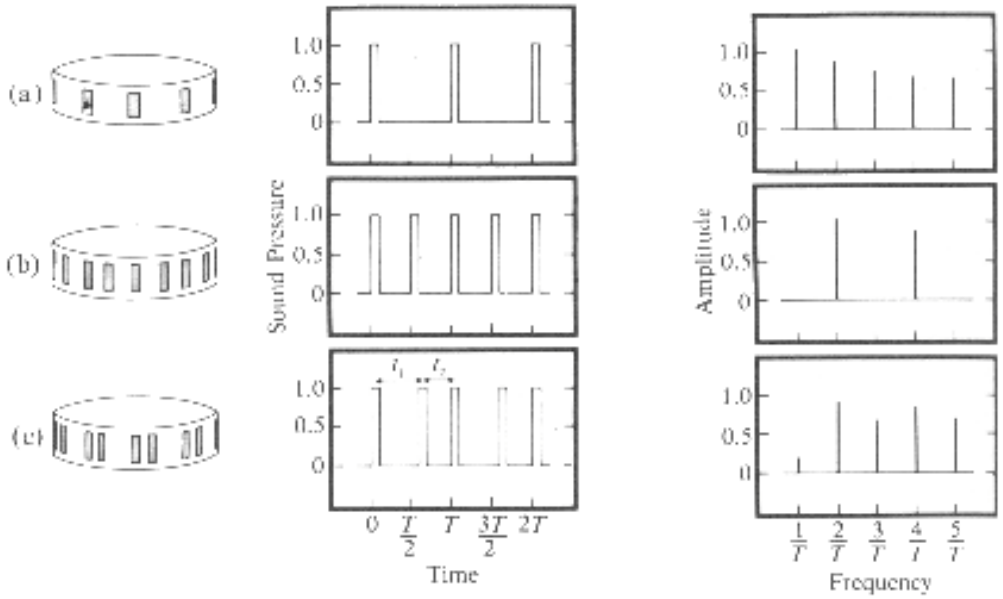


figura 2.14 Experimento de Seebeck

Casi simultáneamente, G. S. Ohm adaptó teorema de análisis espectral de Fourier a la acústica, formulando una hipótesis (ley de la acústica de Ohm, también llamada 2ª ley de Ohm) según la cual sólo podría percibirse el tono correspondiente a una determina frecuencia si la onda acústica tenía una potencia apreciable en dicha frecuencia. Ohm fue muy crítico con la interpretación de Seebeck según la cual es la periodicidad, y no la potencia en la frecuencia fundamental lo que determina el tono.

En el siglo XIX, H. von Helmholtz apoyaba la idea de Ohm, añadiendo importancia a la llamada distorsión armónica, que generaría un fuerte fundamental a partir de los armónicos superiores (tonos de suma y diferencia). Sin embargo, es fácil construir un experimento para refutar el punto de vista de Helmholtz: tomamos un sonido con fuertes armónicos parciales en frecuencias múltiplos de una fundamental. Entonces filtramos la fundamental y añadimos un armónico en una posición ligeramente desplazada respecto a la posición de la fundamental original. Si el oído generase una fuerte fundamental a partir de los armónicos superiores deberían percibirse batidos entre el armónico añadido y la fundamental generada por el oído, lo cual no sucede.

De hecho, no es necesario que una señal tenga nada de potencia en su frecuencia fundamental, para que tenga una periodicidad de valor inverso a su periodo. En la figura se observa un tren de pulsos al cual hemos restado su armónico fundamental. Es sencillo construir experimentos desplazando todas las componentes armónicas de un sonido, donde se observa que la periodicidad de, incluso la envolvente de una onda, es la que nos determina la percepción del tono.

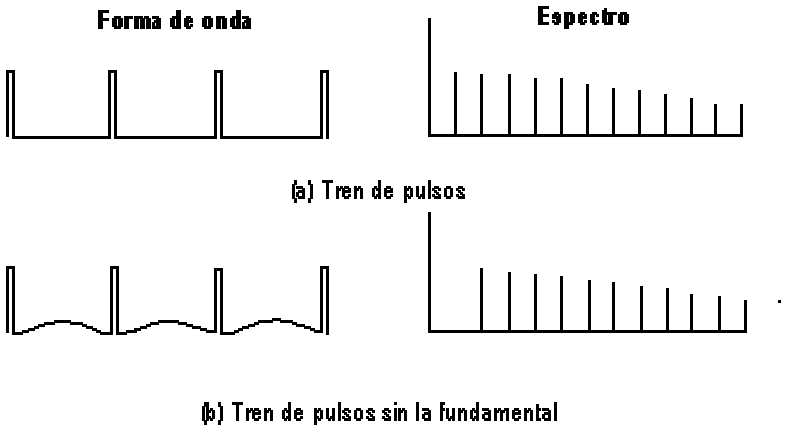


figura 2.15 Eliminación del armónico fundamental

En resumen, la percepción del tono en última instancia es realizada por el cerebro en base a patrones de correlación temporal de las señales de las fibras nerviosas. Sin embargo, no sería posible para el cerebro desentramar esta información si no fuese porque diferentes fibras adquieren la periodicidad de armónicos de diferentes frecuencias, los que recaen dentro de su correspondiente banda crítica.

En la práctica, los sonidos complejos con fundamentales débiles o ausentes son muy comunes. Pensemos por ejemplo, en las conversaciones telefónicas por líneas básicas, las cuales están filtradas por debajo de 400 Hz. Más aun, los sonidos musicales son a menudo parcialmente enmascarados por otros sonidos. Estos sonidos pueden, sin embargo, poseer alturas graves muy claras. Los estímulos sonoros musicales efectivos resultan a menudo incompletos cuando se comparan al sonido producido por la fuente (instrumento, voz).

Experimentos sobre la percepción sonora han apuntado a una región de dominio para la percepción sonora, básicamente de 500 a 2000 Hz. Los parciales dentro de la región de dominio son los más influyentes con respecto a la altura. Una forma de demostrar esto, es trabajar con sonidos con parciales inarmónicos. Supongamos que tenemos un sonido con parciales de 204, 408, 612, 800, 1000 y 1200 Hz. Los primeros tres parciales de forma aislada darían una altura de "204 Hz". Los seis juntos dan una altura de "200 Hz" debido al peso relativo de los parciales más altos, los que se encuentran en la región de dominio. La altura percibida de los sonidos complejos con frecuencias fundamentales graves (menos de 500 Hz) depende de los parciales más altos. La altura percibida de los sonidos con frecuencias fundamentales agudas es determinada por la fundamental, porque se encuentra en la región de dominio.

Los sonidos con parciales inarmónicos han sido usados con frecuencia en la investigación de la percepción sonora. Una aproximación de la altura evocada por ellos es la fundamental de la serie armónica más cercana. Supongamos que tenemos un sonido con parciales de 850, 1050, 1250, 1450, 1650 Hz. La serie armónica más cercana es 833, 1042, 1250, 1458 y 1667 Hz, la cual contiene los armónicos 4, 5, 6, 7 y 8 de un sonido complejo cuya fundamental es 208,3 Hz. Esta fundamental puede ser usada como una aproximación de la sensación de altura del complejo inarmónico (fig. 2.16). Consideremos un sonido inarmónico con parciales de 900, 1100, 1300, 1500, 1700 Hz. Este sonido tiene una altura ambigua, ya que son posibles dos aproximaciones de series armónicas: una con fundamental de 216,6 Hz (el parcial de 1300 Hz es el armónico 6 en este caso) y otra con fundamental de 185,9 Hz (1300 Hz es el armónico 7).

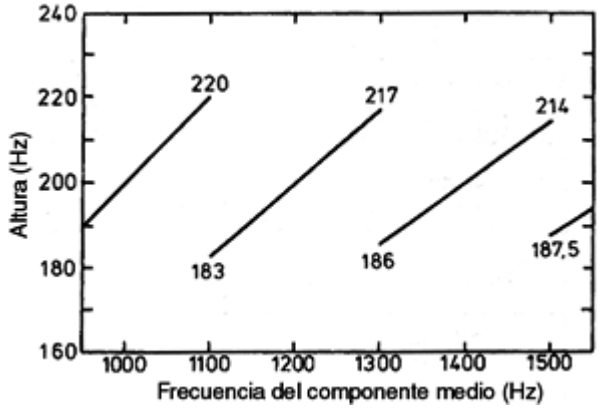


figura 2.16 Sensación de frecuencia en sonidos con parciales inarmónicos

Si no todos los parciales de un sonido complejo son necesarios para percibir la altura grave, ¿qué cantidad es suficiente?

La siguiente serie de investigaciones experimentales muestran un número progresivamente decreciente (fig. 2.17) de parciales necesarios. De Boer (1956) trabajó con cinco armónicos en la serie de dominio; Schouten, Ritsma y Cardozo (1962) con tres; Smoorenburg (1970) con dos; Houtsma y Goldstein (1972) con uno más uno, es decir, un parcial para cada oído. A través de este último experimento, los autores concluyeron que la altura percibida es un proceso del sistema nervioso central, no ocasionado por el órgano sensitivo periférico (los oídos). El último paso en la serie de experimentos sería la altura percibida evocada por un parcial. La posibilidad de esto último también fue demostrada por Houtgast (1976); para esto es necesario cumplir con ciertas condiciones: rellenar con ruido la región de frecuencia de la altura grave, una relación señal-ruido baja, y el deber de dirigir la atención del oyente hacia la región de frecuencia de la fundamental mediante estímulos previos. Estas condiciones, crean una situación perceptiva irreal, de forma tal que somos llevados a la idea de que debería estar por deducciones fomentadas por los estímulos anteriores.

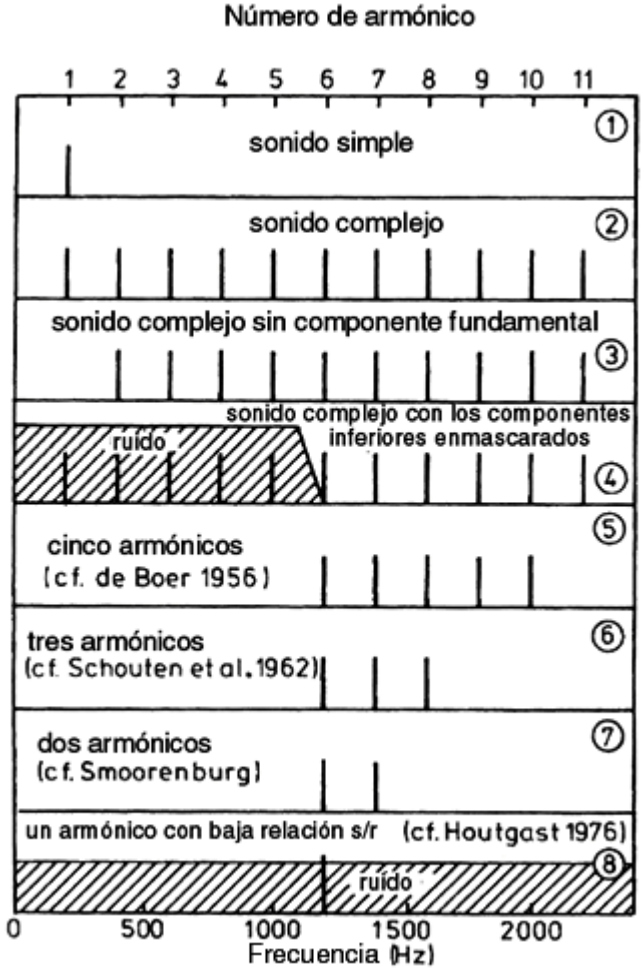


figura 2.17 Experimento de percepción restando armónicos

2.4 Teorías de consonancia y disonancia

El sonido simultáneo de varios sonidos puede resultar agradable o desagradable. Por darle un nombre fácil, el sonido agradable se llama consonante y el desagradable o áspero, disonante. Los términos consonancia y disonancia fueron usados en sentido perceptivo o sensorial; este aspecto ha sido llamado consonancia tonal o consonancia sensorial, lo cual debe distinguirse del concepto de consonancia en una situación musical, o de a lo que consideramos consonancia y disonancia en diferentes estéticas.

La consonancia perceptiva de un intervalo consistente de dos sonidos simples depende directamente de la diferencia de frecuencia entre los sonidos, no de la relación de frecuencia (o intervalo musical). Si la separación de frecuencia es muy pequeña o grande (más que el ancho de banda, sin que los sonidos interfieran unos con otros), ambos sonidos juntos suenan consonantes. La disonancia ocurre si la separación de frecuencia es menor que un ancho de banda (fig. 10). Veamos la evolución histórica de estos conocimientos.

Helmholtz, en el siglo XIX, descubre que el cerebro realiza algo parecido a un análisis espectral de la señal sonora, separando un sonido en sus diferentes armónicos parciales. Experimentando con sus teorías, llega a la conclusión de que hay disonancia cuando la diferencia de frecuencia entre dos parciales (tonos puros que componen el sonido musical) es tal que aparecen entre 30 y 40 batidos por segundo al combinarse.

Investigaciones actuales han llegado a la misma conclusión: la consonancia o disonancia de dos tonos puros (y de sus armónicos) que suena juntos depende de la diferencia de frecuencias, y no de su cociente. La máxima disonancia se produce cuando la diferencia de frecuencias es aproximadamente ¼ del ancho de banda crítico. También, depende ligeramente del nivel de presión sonora. Como el ancho de banda crítico depende de la frecuencia central de la banda, disponemos de una fórmula empírica para calcular la diferencia de frecuencias a la que se obtiene la mayor disonancia:

$$\Delta f = 2.27\,(1+ (Lp - 57)/ 40\,)\; f^{0.447}$$

siendo Lp el nivel de presión sonora y *f* la frecuencia del tono a analizar. De esta forma, podemos obtener una relación de disonancia o consonancia según nos movemos por el ancho de la banda crítica. En la figura 2.18 se muestran gráficamente estas conclusiones.

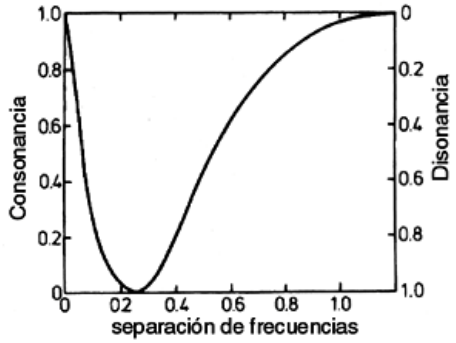


figura 2.18 Ancho de banda de consonancia-disonancia

La explicación de este fenómeno en términos neurológicos se basa en el hecho de que cuando dentro de la banda crítica correspondiente a una fibra nerviosa, tan sólo recae una componente armónica, y la señal neuronal posee una perfecta periodicidad, se facilita al cerebro la búsqueda de las correlaciones temporales. Sin embargo, si dentro de la banda crítica recaen diferentes componentes armónicas, la señal neuronal no posee una correlación fácil de desentrañar para el cerebro, lo que provoca el stress o "tormento perpetuo" que relataba Galileo.

Por tanto, la percepción de la consonancia y disonancia es realizada por el cerebro en función de patrones de comparación de la correlación temporal de las señales que provienen de las fibras nerviosas. Se podría decir entonces, que es fruto de una anomalía propia de la limitación en el cálculo de la frecuencia fundamental del sonido, mediante la división en bandas críticas de frecuencia: el origen de la percepción consonancia-disonancia sería fruto del método de análisis de la señal, no de la naturaleza del estímulo.

El intervalo más disonante ocurre con una separación de frecuencia de alrededor de un cuarto del ancho de banda crítica: alrededor de 20 Hz en secciones de frecuencia grave, alrededor de 4% (algo menos de un semitono) en las regiones más agudas (fig. 2.19). La separación de frecuencia de la tercera menor (20%), tercera mayor (25%), cuarta (33%), quinta (50%), etc., es generalmente suficiente para dar una combinación consonante de sonidos simples. Sin embargo, si las frecuencias son graves, la separación de frecuencias de terceras (y eventualmente también quintas) es menor al ancho de banda crítico, de forma tal, que estos intervalos causan un batido disonante. Por esta razón, estos intervalos consonantes no son usados en el registro de bajo en las composiciones musicales.

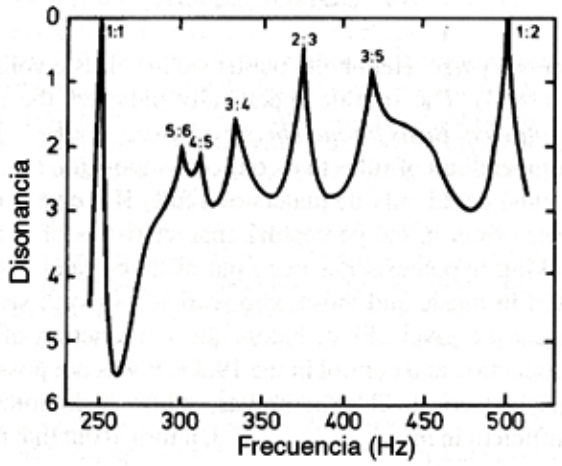


figura 2.19 Relación de disonancia para un tono de 250 Hz

La consonancia de los intervalos entre sonidos complejos puede derivarse de las consonancias de las combinaciones de sonidos simples contenidos en ellos. En estos

casos la disonancia es el elemento aditivo; la disonancia de todas las combinaciones de parciales vecinos puede ser determinada y añadida para alcanzar la disonancia total (o la consonancia total) entre los sonidos. Sonidos con parciales espaciados, por ejemplo el clarinete (toca solo los parciales impares) son más consonantes que los sonidos parciales muy juntos.

La consonancia de un intervalo musical, definido como la suma de dos sonidos complejos con una cierta relación en su frecuencia fundamental, es altamente dependiente de la simplicidad de la relación de frecuencia. Intervalos con relaciones de frecuencia que puedan ser expresados en números pequeños (digamos, menos de 6) son relativamente consonantes porque los componentes más graves (los más importantes) de ambos sonidos están o muy separados o coinciden. Si la relación de frecuencia es menos simple, habrá un número de parciales de ambos sonidos que difieran sólo un poco en su frecuencia, y estos pares de parciales dan lugar a la disonancia. Parece que intervalos con el número 7 en sus proporciones de frecuencia (7/4, 7/5, etc.) están en el límite entre consonancia y disonancia.

La consonancia musical en la música polifónica y armónica occidental, está claramente basada en la consonancia perceptiva de sonidos complejos (armónicos). Intervalos con relaciones de frecuencia simples son consonantes; intervalos con relaciones de frecuencia no simples serán disonantes.

Hay esencialmente tres explicaciones en la actualidad para el concepto de consonancia y disonancia y su asociación con relaciones de frecuencia de denominador bajo. La primera está basada en el hecho de que la mayoría de los sonidos musicales (y en el lenguaje hablado) son sonidos complejos periódicos cuyos parciales están -al menos aproximadamente- armónicamente relacionados con la fundamental. Esta explicación afirma que aprendemos a reconocer las relaciones entre los parciales armónicos de los sonidos complejos y a considerar esas relaciones consonantes o naturales. Un exponente actual de esta teoría es Terhardt.

La segunda explicación también se basa en la estructura armónica de los sonidos complejos. Para el caso de sonidos complejos simultáneos, coincidirán en frecuencia relativamente más armónicos de los sonidos, así como también los productos de la distorsión primaria no lineal resultante de la interacción entre los armónicos, en la medida que los sonidos estén relacionados por fracciones de denominador bajo. Por ejemplo, si los dos sonidos están relacionados por una octava, todos los armónicos del sonido de frecuencia más aguda coinciden con armónicos del sonido de frecuencia más grave, resultando en un sonido suave (consonante). A la inversa, si los sonidos complejos están levemente desafinados, se producirá una sensación de batimiento o aspereza, la cual, se presume, está ligada a la disonancia.

La tercera explicación se basa en el supuesto de que el cerebro prefiere combinaciones de frecuencias cuyos patrones de excitación neural contengan una periodicidad común. Esta idea es sostenida por Boomsliter y Creel (1961) y Roederer (1973). Esencialmente, predice la existencia de "detectores" de relaciones de frecuencia de denominador bajo.

Las dos últimas explicaciones se basan originalmente en sonidos presentados simultáneamente. Habitualmente se argumenta que, ya que las escalas musicales que contienen estos intervalos naturales preexisten a la armonía y polifonía, las explicaciones de desarrollo de escalas basadas en presentaciones simultáneas de sonidos

son inapropiadas. Este argumento es cuestionable, ya que varios sistemas musicales tribales utilizan acompañamientos en octavas, quintas y otros intervalos. En todo caso, ambas explicaciones han sido extendidas a los intervalos melódicos. La explicación de batimientos se extiende a través de lo que podríamos llamar la "hipótesis Picapiedra" (Wood, 1961), esto es, que la música primitiva era ejecutada en cavernas altamente reverberantes que proveían una presentación pseudosimultánea. La versión moderna de esta hipótesis es sostenida por Benade (1976), quien asume que la guía provista por la interferencia (reverberante) entre armónicos desafinados son el principal criterio de entonación usado por los músicos en una situación musical real. La explicación basada en aprender las relaciones armónicas de sonidos complejos es obviamente también aplicable a intervalos melódicos.

3. Construcción y uso de las escalas en modos musicales

Dado que la música práctica se limita a un grupo relativamente pequeño de relaciones discretas de altura, ¿cómo se eligen los valores específicos de estas relaciones? Esto es, ¿existen relaciones de frecuencia "naturales" inherentes en la forma en que el sistema auditivo procesa el estímulo sonoro, que son siempre únicos, y como tales definen los intervalos de la escala? Según la teoría tradicional de la música occidental, tales intervalos no existen, están asociados con el concepto de consonancia-disonancia, y los define el denominador de la relación de frecuencia, por ejemplo: la relación 2:1 (octava) es el intervalo más consonante, la relación 3:2 (quinta) el segundo intervalo más consonante, etc. Debe destacarse que el concepto de consonancia discutido anteriormente se definía -circularmente- por la teoría musical como la sensación asociada con relaciones de frecuencia de denominador pequeño presentadas simultáneamente. Esto puede ser o no, sinónimo de consonancia según los procedimientos psicoacústicos, diseñados para minimizar la influencia del entrenamiento musical. El origen del concepto de consonancia en términos de relaciones de denominador pequeño es usualmente atribuido al estudioso griego Pitágoras. Sin embargo, la preferencia de los antiguos griegos por las fracciones de denominador bajo se basaba posiblemente más en la metafísica que en la psicofísica.

3.1 Los intervalos musicales

Si para pintar un cuadro hemos de recurrir a una paleta de colores, para construir un sistema de composición musical modal, se necesitan una sucesión de sonidos producidos por un instrumento musical (o voz), que constituyan la base de las obras musicales, a la cual denominaremos escala musical.

En las escalas musicales, se asocia una nota musical a cada sonido, en función de la frecuencia relativa del sonido. La variable física que determina la nota es la frecuencia fundamental del sonido.

Entenderemos por intervalo musical, la distancia que separa dos notas. Los intervalos musicales pueden medirse en términos de la relación de frecuencias de los sonidos, aunque en música reciben nombres propios, cuya correspondencia física depende del tipo de escala utilizada. De esta forma, el intervalo formado de Do a Re es una segunda, de Do a Mi es una tercera, a Fa una cuarta, y así sucesivamente.

Por otro lado, y como fruto de las necesidades de la música, a los intervalos a cada distancia les añadimos un apellido que les caracteriza: mayor, menor, justo, aumentado y disminuído (tercera mayor o menor, cuarta aumentada, justa o disminuída, etc), que más tarde explicaremos.

Tal y como hemos visto en el apartado anterior, por las propiedades espectrales de los sonidos, ciertos intervalos resultan más consonantes que otros en virtud de la proximidad entre armónicos aportados por cada una de las notas que suenan simultáneamente.

En el siguiente espectograma, analizamos el sonido emitido por un piano que toca las

notas más próximas a la serie de armónicos de la nota do1, (el "1" identifica a la octava a la que pertenece el do) de frecuencia fundamental 32.703 Hz. Nos fijaremos en las relaciones de proximidad entre los armónicos.

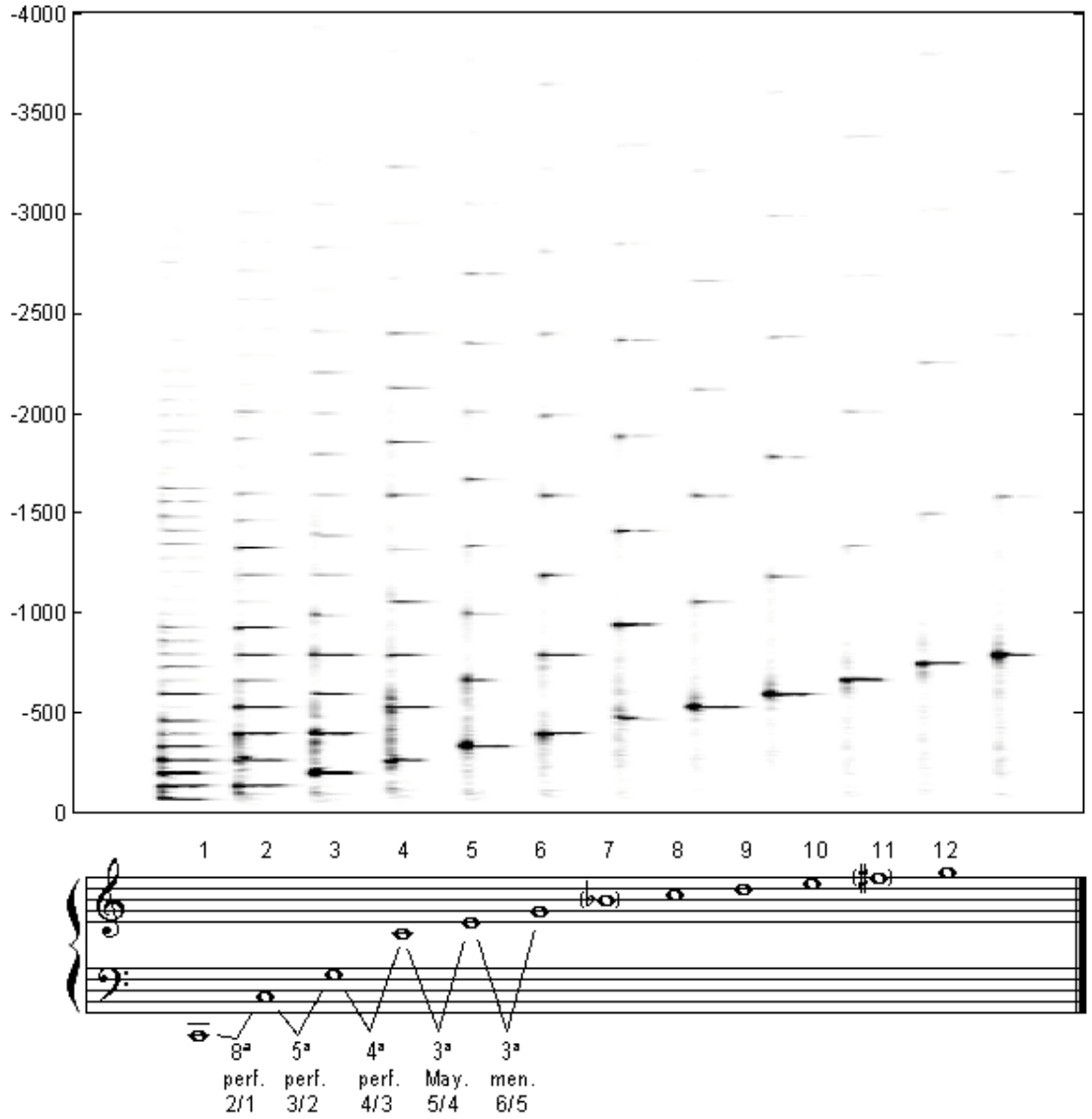


figura 2.20 Espectograma de aparición de armónicos

El intervalo de octava resulta muy consonante por la perfecta coincidencia entre los armónicos. Asimismo, si un sonido es consonante con otro, también lo será con el sonido una octava más alto, pues este no añade ningún armónico capaz de producir disonancia. El intervalo de quinta perfecta también resulta muy consonante pues los armónicos que no coinciden perfectamente, quedan a la distancia más alejada posible, perfectamente intercalados. En menor grado esto también es cierto para los intervalo de cuarta perfecta y tercera mayor y menor. Nos fijamos en estos intervalos exclusivamente, pues como veremos a continuación son los que utilizaremos como punto de partida para la construcción de las diferentes escalas.

Por otro lado, es justo advertir que, hasta bien entrado el siglo XIII, las únicas consonancias perfectas consideradas en la música occidental fueron la octava, el unísono (misma nota), la quinta y la cuarta. Fue precisamente la modificación de los

sistemas de afinación (con gran influencia del español Bartolomé Ramos de Pareja) lo que llevó a la incorporación de las terceras y sextas, que hasta entonces no podían considerarse consonancias. Para terminar este inciso, hay que remarcar que la introducción de las sextas y terceras, dio lugar a una nueva sonoridad, fruto de la nueva posibilidad de agrupar verticalmente notas diferentes de la octava, quinta y cuarta, lo cual llevó a la aparición de los acordes por terceras o sextas, base de nuestro sistema armónico actual.

¿Qué podemos tomar como punto de partida para la construcción de una buena escala?. La idea esencial es que la escala posea suficientes notas y que los intervalos entre ellas sean lo más consonantes posibles, pues al fin y al cabo las notas de la escala constituyen los ladrillos con los que se construir la música a interpretar. En función de a qué intervalos demos mayor prioridad, surgen muchas maneras de construir escalas. Vamos a estudiar las escalas justa, pitagórica y de igual temperamento.

3.1.1 La Escala de Entonación Justa

La escala de entonación justa (diatónica justa) nace a partir de la tríada mayor, un grupo de tres notas que suenan particularmente armoniosas (p.ej. do-mi-sol). El intervalo entre do-mi es una tercera mayor, entre mi-sol una menor y entre do-sol una quinta perfecta. La triada se denomina triada mayor y posee relaciones de frecuencias 5/4 (mi) y 3/2 (sol) respecto al do.

La escala justa es la llamada de Aristógenes, de Zarlino o de los físicos, por sus connotaciones más teóricas que prácticas. En la escala justa (primer tipo que surge al recorrer las notas: do, re, mi ,fa, sol, la, si, do) podemos ir superponiendo una tercera y una quinta a cada nota, y comprobamos que hay tres triadas mayores, que se denominan acordes de tónica, de subdominante y de dominante (I, IV y V grado, sobre do, fa y sol). Están construidas siempre partiendo de las primera, cuarta y quinta notas de la escala respectivamente.

Para construir la escala justa seguimos el siguiente procedimiento: fijamos primero las notas del acorde de tónica (do-mi-sol). Partimos de do1 normalizado a 1, entonces mi1=5/4 y sol1=6/4=3/2. Ahora procedemos con las notas del acorde de dominante (sol-si-re), por tanto, si1=3/2*5/4=15/8 y re2=3/2*3/2 = 9/4, trasponemos este re2 una octava a re1=9/8. Ahora tomamos el acorde de subdominante (fa-la-do), bajamos desde do2, a fa1=2/(3/2)=4/3, y a la1=2/(6/5)=5/3.

Desarrollamos las relaciones de la escala justa de do mayor, observándose tres diferentes relaciones entre notas sucesivas.

do	re	mi	fa	sol	la	si	do
1	9/8	5/4	4/3	3/2	5/3	15/8	2
	9/8	10/9	16/15	9/8	10/9	9/8	16/15
	Tono entero mayor	Tono entero menor	Semitono	Tono entero mayor	Tono entero menor	Tono entero mayor	Semitono

Aparte de las tres triadas mayores, la escala justa tiene dos triadas con las relaciones 10:12:15, que se denominan triadas menores. Al igual que las triadas mayores, también poseen un intervalo de tercera menor y uno de tercera mayor, pero cambiados de orden. Es decir, el intervalo más grave (10/12=6/5) es una tercera menor, y el intervalo más agudo (15/12=5/4) es una tercera mayor.

Este esquema de formación de escalas, manifiesta sin embargo bastantes problemas, lo históricamente obligó a buscar nuevos métodos de construcción, Para verlos, comprobamos algunos casos:

- Las quintas en la escala justa no son todas iguales: son todas perfectas (3/2) todas menos una, entre re y la.
- Una de las cuartas tampoco es perfecta (la-re), puesto que la-re no es perfecta.
- Si añadimos alteraciones (sostenidos y bemoles), encontramos el problema de que si requerimos que mi-sol# sea una tercera mayor, entonces sol# =5/4*5/4=25/16, pero si requerimos que la*b*-do sea una tercera mayor, entonces lab=2/(5/4)=8/5. Es decir, la*b* es un poco más aguda que sol#. Sin embargo, en el piano para sol# y lab hay una sola tecla: son las llamadas notas enarmónicas.

La construcción de instrumentos afinados en la escala justa no es muy práctica porque requeriría complicados teclados para las enarmonías mencionadas anteriormente y haría falta volver a afinar el instrumento completamente cada vez que se desease cambiar de tonalidad.

3.1.2 La escala Pitagórica

La escala pitagórica se basa en la creación del mayor número posible de cuartas y quintas perfectas. Para conseguirlo, sacrificamos la afinación de terceras mayores y menores, así como las sextas, respecto la entonación justa.

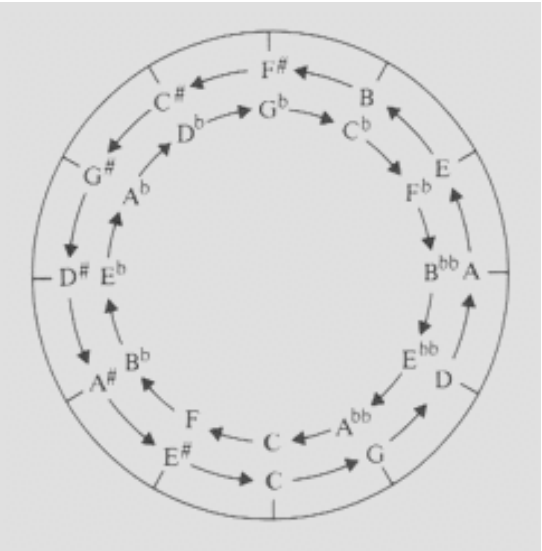


figura 2.21 La espiral pitagórica

Para empezar debemos tener en cuenta que:

- Una octava es una cuarta más una quinta (4/3*3/2=2), por tanto, subir una cuarta es lo mismo que basar una quinta y viceversa.
- Todas las notas de la escala (incluidos sostenidos y bemoles), pueden alcanzarse subiendo o bajando 12 quintas o doce cuartas sucesivamente.

Para construir la escala pitagórica vamos ascendiendo en intervalos de quintas justas (do1-sol1-re2-la2-mi3-si3; luego trasponemos las notas a la octava apropiada), si continuamos el proceso obtenemos los sostenidos, y ascendiendo cuartas justas, obtendríamos los bemoles. La escala pentatónica surge naturalmente mediante este procedimiento constructivo; basta detener el proceso una vez obtenidas cinco notas.

Vamos a poner unos ejemplos de cálculo de intervalos.

1) Semitono cromático:

Buscamos la diferencia entre Fa y Fa#: Siempre partimos de Do.

Fa# → $(3/2)^6$ /8 = 729/512, puesto que hay que pasar por seis quintas y cambiar 3 veces de octava (2^3=8), teniendo que normalizar.

Fa → 4/3, puesto que hay que pasar por una cuarta y no se cambia de octava.

Semitono cromático= 729/512 x 3/4 = 2187/2048

2) Semitono diatónico.

Buscamos la diferencia entre Si y Do. Partimos como siempre, de Do.

Do → 1, puesto que es el origen y no se cambia de octava.

Si → $(3/2)^5$ /8 = 243/256, puesto que hay que pasar por cinco quintas y se divide por 8 porque se cambia de octava tres veces

Semitono diatónico= 1 x 256/243 = 256/243

Sin embargo, si continuamos el procedimiento para obtener todos los sostenidos y bemoles, encontramos relaciones como las resultante entre fa#-fa es 2187/2048=1.068 (la denominaremos semitono cromático), que es menor que el semitono diatónico (256/243=1.053) que aparece entre notas sin alteraciones.

Si continuamos el proceso 12 veces intentando llegar a la misma nota de partida y así cerrar el círculo después de obtener todas las posibles notas alcanzamos la relación 3/2^12=129.7, próxima a siete octavas, pero con un error acumulado que nos aparta de nuestro objetivo, las siete octavas 2^7=128. Este intervalo resultante, se denomina coma pitagórica, y es igual al intervalo que separa un semitono cromático de un semitono diatónico, el intervalo que separa notas enarmónicas.

En la siguiente tabla. desarrollamos las relaciones de la escala pitagórica de do mayor, la cual tiene una única relación de tono, pero dos de semitono.

do	re	mi	fa	sol	La	si	do
1	9/8	81/64	4/3	3/2	27/16	243/128	2
	9/8	9/8	256/243	9/8	9/8	9/8	256/243

La principal ventaja de la escala pitagórica es el que las cuartas y las quintas son siempre perfectas. Sin embargo, las terceras tienen una afinación muy pobre. La relación por la cual las tercera mayores exceden y las terceras menores quedan por debajo de las terceras justas, es 1.0125, denominado coma sintónico.

También resulta interesante saber que muchos cantantes y violinistas tienden en favor de la entonación pitagórica en sus interpretaciones, sobre todo en música tradicional, lo que ratifica la importancia de las quintas y cuartas en la música.

3.1.3 La afinación temperada

Las afinaciones temperadas parten de la escala pitagórica, con la idea de alterar algunas notas en fracciones de coma sintónico antes mencionado, acondicionando un poco las terceras, que suenan muy desafinadas. Es justo decir que hubo muchísimos intentos para establecer un temperamento, sólo en España durante el siglo XVII Y XVIII, sabemos de la existencia de varias decenas. El procedimiento de construcción que triunfó, y que vamos a explicar, se lo debemos a Chladni, aunque ya en el Renacimiento, el teórico español Salinas especificaba en sus tratados las bases de construcción de la afinación temperada actual.

La escala de temperamento de tono medio de cuarto de coma, sube o baja varias notas en fracciones 1/4, 1/2, 3/4 o 5/4 del coma sintónico (el valor por el cual las terceras mayores y menores difieren de sus valores justos correspondientes). Sin embargo, entendemos que no se trata más que de una aproximación.

La escala de igual temperamento, habitualmente denominada escala temperada se basa en que todos los semitonos sean idénticos (lo que implica también tonos idénticos). Una octava está formada por doce semitonos, o bien cinco tonos y dos semitonos. La relación del semitono de igual temperamento es igual a 2^(1/12) = 1.05946.

El tono corresponde a 1.05946^2=1.12246. La quinta es 1.498 y la cuarta es 1.335, ambas muy próximas a los intervalos perfectos 1.500 y 1.333. La tercera mayor es 1.260 y la tercera menor 1.189,y por tanto no están demasiado próximas a los intervalos justos 1.250 y 1.200, pero tampoco están tan desafinados como las terceras de la escala pitagórica.

En lugar de trabajar con relaciones de frecuencias, es habitual comparar las notas utilizando cents. Un cent es 1/100 de un semitono de igual temperamento. Por tanto, una octava son 1200 cents, una quinta temperada 700 cents, una cuarta temperada 500 cents, y así sucesivamente. Un cent corresponde a la relación 2^(1/1200)=1.000578.

La principal ventaja de la escala de igual temperamento reside en que todas las notas enarmónicas poseen la misma afinación, imprescindible para la construcción de muchos de los instrumentos musicales), y que no es necesaria ninguna afinación para interpretar obras en diferentes tonalidades.

La importancia que ha tenido la instauración de una afinación igual para todos los instrumentos, ha sido tal, que el progreso de la música occidental sin él, habría estado lleno de problemas. Sabemos que en la Francia del Renacimiento, para que dos personas se pusieran de acuerdo en medio de una discusión, se les decía "Accordez vos flûtes" , algo así como “afinen sus flautas”. Ambos asuntos, muy difíciles de conseguir en la época. Los conceptos de afinación, como qué notas son consonantes entre sí, o no, y de temperamento, como las funciones melódicas y armónicas de cada nota en una escala, han ido variando a lo largo de los siglos hasta llegar a nuestro "temperamento igual": la división de una octava en doce semitonos de idéntico tamaño. Principios matemáticos, musicales, y culturales han acarreado más de un dolor de cabeza a músicos, teóricos y oyentes desde el famoso "Círculo de Quintas pitagórico".

3.2 Consideraciones prácticas sobre el uso de las escalas.

Dado que la música occidental actual utiliza un conjunto relativamente pequeño de relaciones discretas de altura, surge una pregunta obvia: ¿es general este uso de intervalos discretos? La evidencia de estudios etnomusicológicos indica que el uso de relaciones discretas de altura es esencialmente universal. Las únicas excepciones parecen ser ciertos estilos musicales primitivos, los cuales se encuentran en unas pocas culturas tribales. También el concepto de octava parece ser común a los sistemas musicales más avanzados.

Surge otra pregunta: ¿la escala cromática de 12 notas representa una norma o un límite al número de relaciones de alturas utilizables dentro de una octava? Hay varias composiciones occidentales basadas en cuartos de tono (aproximadamente 24 intervalos iguales por octava, como en Ligeti) y otras escalas microtonales, pero ninguna de estas escalas ha ganado mayor aceptación. Numerosas culturas usan escalas de menos de 12 notas por octava. Hay, sin embargo, aparentemente, solo hay dos culturas que, en teoría, usan más de 12 intervalos por octava: la hindú y la árabe-persa. Los dos sistemas hindúes (Hindustani y Karnático) están, según la tradición, basados en 22 intervalos posibles por octava; no son intervalos iguales (o similares). Hay evidencia que indica que en la práctica musical estas variaciones teóricas no son ejecutadas como intervalos discretos, sino como resultado de una variación controlada en la entonación (un vibrato lento).

El único sistema que podría utilizar verdaderos cuartos de tono es el árabe-persa. En este sistema hay varias versiones acerca del número posible de intervalos (15 a 24), así como cierta controversia acerca de si son verdaderos cuartos de tono o variaciones microtonales de ciertos intervalos (como en la India). Resulta claro que ni la escala hindú ni la árabe son cromáticamente microtonales.

Por tanto, la evidencia indica que la escala cromática de 12 intervalos puede, en efecto, reflejar algún tipo de limitación al número de intervalos por octava, que pueden tener un uso práctico en la música o, al menos, que el semitono es la separación utilizable más pequeña entre dos sonidos sucesivamente ejecutados. ¿Hay fundamentos perceptivos para el uso de intervalos discretos en general? Autores de libros sobre percepción musical a menudo contrastan el número de diferencias notorias de frecuencia (JND) con el semitono de la escala cromática: de 20 a 300 JNDs por semitono, dependiendo del rango de frecuencia y del paradigma experimental; de igual forma señalan la aparente discrepancia entre el número de sonidos que pueden ser "distinguidos" y el número de sonidos efectivamente utilizados en música. Hay varias razones por las cuales este contraste no es significativo.

Primero, es evidente a partir de la teoría musical, de la experiencia musical diaria, y de un número de experimentos, que la relación de frecuencia (más que la frecuencia per se) es el mediador primario de la información melódica. Es así que los JNDs de la relación de frecuencia más que de la frecuencia, son la comparación más relevante con el grado de la escala. Los JNDs de la relación de frecuencia son, para observadores altamente entrenados, un orden de magnitud mayor que JNDs de frecuencia obtenidos en la misma región de frecuencia, usando paradigmas psicológicos equivalentes. Trotter (1967) también indicó que existe poca correlación entre la inhabilidad para percibir relaciones melódicas o interválicas y la habilidad para discriminar la frecuencia.

La segunda, y más importante razón para la aparente discrepancia entre la magnitud de JND (ya sea de frecuencia o de relación de frecuencia) y el tamaño del grado de la escala es la limitación en el proceso de información por parte del sistema nervioso impuesta por niveles de procesado más altos (relacionados con la memoria y la atención). Un concepto útil en este respecto es el de la teoría de la información, en la cual la información transmitida se relaciona con la reducción de la incertidumbre. Los JNDs mencionados arriba fueron determinados con procedimientos psicofísicos de mínima incertidumbre y, como tales, probablemente reflejan el poder resolutivo del sistema auditivo periférico más que limitaciones impuestas por un procesado de mayor nivel. Numerosos estudios han indicado que al confrontarse con señales de alta información y/o rangos de alta información, los observadores tienden a codificar la información en categorías como medio de reducir la carga de información. Es posible que esos factores hayan determinado el uso de intervalos discretos.

En 1956, Miller revisó experimentos que probaban la habilidad de los observadores para categorizar estímulos en varias modalidades sensoriales. Concluyó que para los estímulos que varían en una sola dimensión física, la información transmitida al observador estaba en el orden de los 2,8 bits en todas las modalidades; esto es, eran capaces de colocar, sin error, el estímulo tan sólo en un máximo de $2^{2,8}$ categorías. Esto fue contrastado con la habilidad de los observadores para discriminar varios miles de estímulos a lo largo de un continuo en pruebas discriminativas de elección forzada. La discrepancia entre la magnitud de JNDs y el intervalo más pequeño, por tanto, parece ser otro caso de esta clásica discrepancia entre resolución en la identificación y habilidades discriminativas.

3.2.1 Ajuste de intervalos musicales aislados

Los procedimientos de ajuste también han sido usados para estudiar la percepción de los intervalos musicales, principalmente la octava, pero también otros intervalos. En el paradigma típico, al sujeto se le presentan pares de sonidos (secuenciales o simultáneos), uno de los cuales tiene una frecuencia fija, y otro cuya frecuencia está bajo el control del sujeto. A este sujeto se le instruye para que ajuste la frecuencia del sonido variable, de forma tal que la relación de altura de ambos sonidos corresponda a un intervalo musical determinado.

Para el caso de ajustes repetidos de intervalos, los sujetos generalmente demuestran una variabilidad bastante pequeña, relativa a la obtenida en experimentos de producción de magnitud. La desviación promedio de ajustes repetidos de octavas secuenciales o simultáneas compuestas por sinusoides, se encuentra en el orden de los 10 cents (Ward, 1953, 1954; Terhardt, 1969); es levemente inferior en el caso de octavas compuestas por sonidos complejos. Un rango de desviaciones promedio de 14 a 22 cents para el ajuste de los otros intervalos de la escala cromática (presentados simultáneamente) han sido observados por Moran y Pratt (1926).

3.2.2. Evidencia experimental relevante para intervalos naturales y escalas

a. Escalas no occidentales. Tres de los sistemas musicales no occidentales más importantes (hindú, chino y árabe-persa) tienen escalas aproximadamente equivalentes a las escalas occidentales de 12 intervalos, y por tanto tienen la misma propensión hacia la consonancia "perfecta" (octava, cuarta y quinta). Hay, sin embargo, algunas culturas musicales que aparentemente emplean escalas de 5 y 7 intervalos aproximadamente igualmente temperados en los cuales las cuartas y quintas están significantemente desafinadas de sus valores naturales. Escalas con 7 intervalos son asociadas con culturas del sudeste asiático. Por ejemplo, Morton (1974) registra mediciones de la afinación del xilófono tailandés que "varía sólo +- 5 cents" de una afinación de 7 intervalos igualmente temperados. La escala de 5 intervalos de 240 cents, se asocia con las orquestas de "gamelan" de Java y Bali; sin embargo, dice McPhee (1966), "las desviaciones dentro de lo que se considera una misma escala son tan grandes que uno podría afirmar -con buena razón- que hay tantas escalas como gamelans".

Por tanto, parece haber una propensión hacia las escalas que no utilizan consonancias perfectas y que son en muchos casos altamente variables, en culturas que son o pre-instrumentales o cuyos instrumentos principales son del tipo del xilófono. Instrumentos de esta índole producen sonidos cuyos parciales son, en gran medida, inarmónicos y cuyas alturas son a menudo ambiguas.

b. Entonación en la ejecución. Un número de mediciones han sido hechas sobre la entonación de músicos tocando instrumentos de afinación variable en condiciones reales de ejecución; los resultados de estas mediciones fueron registrados por Ward (1970). Muestran una variación considerable en la afinación de un intervalo dado en una ejecución. La tendencia generalizada parece ser contraer el semitono y expandir levemente todos los otros intervalos relativos al temperamento igual. Hay también

alguna evidencia de efectos dependientes del contexto (por ejemplo: tocar Fa# más alto que Solb). Estos resultados reflejan, hasta cierto punto, los resultados de experimentos sobre identificación de intervalos aislados, los cuales mostraban una tendencia a comprimir la escala para los intervalos pequeños y agrandarla para los intervalos grandes, tanto en presentación ascendente como descendente.

3.2.3. Identificación y discriminación de intervalos musicales

Los resultados de experimentos sobre identificación de tres intervalos musicales armónicos, han sido analizados a la luz de categorías naturales de relación de denominador bajo (2/1,3/2,etc). Éstos han mostrado que la similitud de los intervalos se basa principalmente en el ancho del intervalo. Esto es, las confusiones más pronunciadas (por parte de los sujetos en los experimentos con pruebas de identificación absoluta) tuvieron lugar entre intervalos adyacentes o, equivalentemente, el tiempo de respuesta para determinaciones "diferentes" fue inversamente proporcional a la diferencia en el ancho de los intervalos. Hubo leves tendencias adicionales en todos los experimentos a considerar categorías equivalentes de nombres (por ej. tercera mayor y menor) e inversiones (segunda menor, séptima mayor) como más similares, pero no hubo tendencia generalizada a confundir entre relaciones de bajo denominador.

Los resultados de estos experimentos de identificación absoluta son consistentes con los resultados de los experimentos de categorización y ajuste de intervalos, los cuales muestran, en general, distribuciones unimodales de categorías de intervalos a lo largo de la dimensión de magnitud de relación de frecuencia.

3.2.4. Conclusiones: categorías aprendidas versus innatas

La evidencia que se ha presentado hasta ahora implica que las categorías de intervalos musicales son aprendidas, más que un resultado directo de las características del sistema auditivo. Esta evidencia incluye:

1) la variabilidad hallada en escalas medidas y entonación, incluso cuando los efectos contextuales posibles son tenidos en cuenta.

2) la variabilidad propia del sujeto, y desviaciones consistentes de categorías de relaciones de bajo denominador en experimentos de categorización y ajuste de intervalos.

3) la ausencia de singularidades de relaciones de bajo denominador en funciones de JND con relaciones de frecuencia de bajo denominador y ausencia de confusiones de relaciones de bajo denominador en experimentos de identificación absoluta.

4) la relativa inhabilidad de sujetos sin entrenamiento musical para identificar o discriminar intervalos musicales.

Asumiendo, entonces, que la entonación de los músicos se basa en su habilidad para reproducir categorías aprendidas, hay básicamente tres hipótesis alternativas para el origen de estas categorías:

1) las categorías son aprendidas de las escalas de una cultura determinada, cuyos intervalos fueron originalmente elegidos al azar,

2) las categorías son aprendidas de escalas de una cultura, cuyos intervalos derivaron de consideraciones sobre consonancia sensitiva (Plomp y Levelt, 1965), o

3) las categorías se basan en el temprano e inconsciente aprendizaje de las relaciones entre los parciales de sonidos ambientales, principalmente la voz hablada (Terhardt, 1977, 1977, 1978).

Dada la recurrencia generalizada hacia las consonancias perfectas, especialmente la octava, la hipótesis 1 parece insostenible. La variabilidad de las escalas en la música pre-instrumental o música cuyos instrumentos más importantes producen sonidos inarmónicos es consistente tanto con la 2, como con la 3. Un problema con la hipótesis de Terhardt, sin embargo, es que esta hipótesis predice que incluso observadores sin entrenamiento musical poseen un sentido de intervalos musicales básicos. La evidencia obtenida a partir de algunos experimentos indica que no es así; en efecto, incluso el concepto de similitud de octava y unísono parece ser una función del entrenamiento aprendido. También parece ser que los niños no poseen un sentido interválico innato.

Por tanto, basados en la evidencia existente, la mayor posibilidad parece ser que los intervalos naturales, según los dicta la consonancia sensitiva, han influenciado la determinación de las escalas de la mayoría de las culturas, pero que la entonación de los músicos individualmente es básicamente una función de su habilidad adquirida para reproducir las categorías de intervalos aprendidos de estas escalas. Por otro lado, la variabilidad de entonaciones, se vería originada por los sonidos ambientales y por las variedades dialectales propias de la voz hablada.

3.2.5 Generalización de octava.

3.2.5.1 Introducción

Como se ha mencionado, las escalas de la música occidental se basan, en parte, en el concepto de generalización de la octava (esto es, que los sonidos separados por una octava son, en cierto sentido, musicalmente equivalentes, y por tanto las escalas son únicamente definidas especificando los intervalos dentro de una octava). Esta generalización parece ser universal para las culturas musicales avanzadas.

3.2.5.2 Posibles explicaciones para la generalización de la octava

Hay varias explicaciones posibles para el carácter único de la octava como base de una supuesta circularidad de altura relativa. Hablando de consonancia sensitiva, se ha mencionado que para intervalos musicales simultáneos formados por sonidos complejos cuyos parciales guardan relación armónica, la octava exacta es única en el sentido de que todos los parciales de los sonidos coincidirán exactamente. Por tanto, el intervalo de octava no será más disonante que el sonido complejo de altura más grave.

Otra explicación es consecuencia de las alturas de los sonidos complejos. Modelos actuales de percepción de alturas de sonidos complejos asumen que la percepción de la altura de sonidos complejos es un proceso de reconocimiento, en el cual un "procesador central de altura" intenta hacer coincidir los parciales del sonido complejo con la serie armónica más adecuada. Una consecuencia de esta operación será un cierto grado de ambigüedad de octava en las predicciones de modelo de la altura fundamental.

3.2.5.3 Evidencia psicofísica con respecto a la generalización de la octava

Las representaciones bidimensionales de la altura implican que las manifestaciones de la equivalencia de octava deberían encontrarse en experimentos para los cuales el entrenamiento musical no es un requisito previo.

Sin embargo, hay evidencia (Shepard, 1964) de experimentos en los que se utilizan sonidos complejos, cuyos parciales consisten en octavas de la fundamental. Estos sonidos complejos son, esencialmente, pasados a través de un filtro de banda que sirve para mantener la altura sonora promedio constante, independientemente de la frecuencia fundamental. Los experimentos muestran que los juicios sobre relación de altura entre sonidos de este tipo con diferentes frecuencias fundamentales se basan en la proximidad relativa de los armónicos más que en las diferencias de frecuencia fundamental absolutas. El resultado sorprendente se da cuando un grupo de sonidos de esta índole, cuyas frecuencias fundamentales cubren un rango de una octava en semitonos, son tocados cíclicamente: la impresión es la de una altura constantemente ascendiendo (o descendiendo) sin los saltos de octava que uno podría esperar (y que se oyen si los sonidos no están debidamente separados). Esta ilusión se cita a menudo como evidencia de la circularidad de la octava.

Nuestra conclusión es que la generalización de octava es un concepto aprendido cuyos orígenes se encuentran en la posición única de la octava en el rango de consonancia sensitiva de intervalos de sonidos complejos.

3.3 Conclusiones al estudio sobre la percepción de la frecuencia.

Basados en la evidencia comentada, las siguientes conclusiones con respecto a la percepción de intervalos musicales y escalas parecen justificadas.

1. El uso de un número relativamente pequeño de relaciones discretas de alturas en la música es probablemente dictado por limitaciones inherentes en el procesamiento de estímulos de alta carga de información por parte de los sistemas sensitivos humanos.

2. Los intervalos naturales, en el sentido de intervalos que muestran disonancia sensitiva mínima (aspereza) en el caso de presentación simultánea de sonidos complejos, han probablemente influido la evolución de las escalas de muchas culturas musicales, pero los estándares de entonación para una cultura dada son las categorías interválicas aprendidas de las escalas de dichas culturas. Un corolario de esto es que la acción de entonación de un músico dado es básicamente determinada por su habilidad

para reproducir estas categorías aprendidas y es poco influida, en la mayoría de los casos, por señales psicofísicas (aspereza, batimientos o consonancias desafinadas, etc.).

3. El concepto de percepción categórica, también relacionado a las limitaciones del proceso de estímulos de alta carga de información, es probablemente una descripción razonable de la forma en que los intervalos son percibidos en todas las situaciones excepto las de incertidumbre mínima, una situación análoga a la percepción de fonemas en el habla.

4. Música de cuartos de tono puede ser teóricamente posible dado cierto acostumbramiento a ella, aunque la escala occidental actual de 12 intervalos ha supuesto, posiblemente un límite práctico para la percepción en las culturas occidentales. La división de la escala en intervalos en cuartos de tono puede contener información melódica sustancial en algunas culturas.

5. La generalización de la octava es posiblemente aprendida, y sus raíces se encuentran en la única posición de la octava en el espectro de la consonancia sensitiva de intervalos de sonidos complejos.

3.4 Reservas

La percepción de intervalos musicales aislados puede tener poco que ver con la percepción de la melodía. Hay considerable evidencia de que las melodías son percibidas como patrones, más que como una sucesión de intervalos sucesivos, y que la magnitud interválica es sólo un factor en el total percibido.

La habilidad de nombrar intervalos individuales no es crucial para la percepción (e incluso la producción) musical. Muchos músicos amateurs que aprenden y reproducen melodías "de oído" no pueden identificar intervalos aislados. La percepción categórica, al menos tal cual se la describe habitualmente, puede ser relevante tan sólo cuando los músicos están escuchando "analíticamente", por ejemplo para transcribir una melodía.

Hay obviamente una necesidad de experimentos sobre percepción de la entonación de notas individuales en frases melódicas desconocidas y familiares.

4. Una aplicación de reconocimiento automático de modos musicales

4.1 Introducción.

Una vez estudiados los problemas y especiales características de la percepción de la frecuencia por el ser humano en un contexto de articulación musical, nos adentramos en la tarea de diseñar de un reconocedor de modos musicales para voz cantada. En esta sección, describiremos el conjunto de las decisiones de alto nivel que fueron tomadas a lo largo del diseño, y nos centraremos más tarde en la descripción a bajo nivel de las técnicas utilizadas en cada bloque del reconocedor.

La forma de abordar la descripción elegida es la de ir estudiando en detalle cada un de los grandes bloques que se pueden identificar en la aplicación, aunque no haya sido necesariamente así su generación temporal.

El objetivo de nuestro reconocedor será estimar la frecuencia fundamental de la voz cantada y sacar ciertos parámetros de más alto nivel relacionando las frecuencias que han ido pareciendo. Para ello, tomamos la decisión de realizar un análisis del tono en tiempo real, alargado todo lo que el usuario pretenda, y después realizar el análisis interválico en diferido sobre unos histogramas de frecuencia o sobre ventanas que mostrarán ciertos resultados.

Por tanto, y a priori, el reconocedor debe constar de los siguientes módulos:

- Una etapa de muestreo de sonido y preprocesado, en la que recogemos el sonido por medio de un micrófono y la convertimos en digital con un conversor A/D, o bien reproduciendo un fichero de audio almacenado mediante un sencillo reproductor, que será dirigido al mezclador interno. Después, una etapa de preprocesado adecuará el sonido a unos valores normalizados. Tras esta etapa, tendremos las muestras almacenadas en nuestra memoria para poder empezar a estimar parámetros.
  En el caso de no necesitar tiempo real, se podría realizar todo el análisis en tiempo diferido, almacenando previamente las muestras de audio y procesándolas con estrategias diferentes, y quizá más eficientes, que las de tiempo real.

- Una etapa de estimación de la frecuencia fundamental y, si es necesario, de la energía de la señal. Para ello, hace falta la implementación de un algoritmo de cálculo de la frecuencia fundamental de la voz, especialmente diseñado para las cualidades de la voz cantada, y que también pueda darnos información acerca de la energía. Tras esta etapa, obtenemos los primeros parámetros de estudio: las alturas de las notas cantadas, la forma de alcanzar notas consecutivas (continua o a saltos) y la energía en el discurso musical.

- Una etapa que realice una proyección de los resultados de etapas anteriores sobre un espacio de parámetros, y que los agrupe mediante decisión por alguna técnica matemática o basada en reglas.

- Una etapa que compare los resultados con una base de datos almacenada en memoria de modos musicales, estime las funciones de coste y haga una estimación del resultado más probable.

El diseño parte de la autoimposición de crear una herramienta que no necesite requisitos especiales de potencia de cálculo o de hardware externo, para facilitar la movilidad en el entorno de adquisición de los datos sobre una plataforma tecnológica generalizada. La solución tomada es la programación de una aplicación para cargarse en un ordenador (portátil o no) con un sistema operativo MS Windows (elección que puede ser ampliada al entorno Apple o Unix, lo cual quizá la haría más estable).

Posteriormente, se tomaron las decisiones de restringir el análisis a la frecuencia fundamental y de rebajar el procesado en tiempo real al mínimo posible, puesto que el muestreo lleva consigo una carga computacional bastante grande. El análisis de la energía no suponía un esfuerzo demasiado grande, sobre todo con el tipo de algoritmo de estima de tono que utilizamos (como veremos), pero se separaba del núcleo del programa que estábamos construyendo, más centrado en la frecuencia fundamental. En todo caso, la ampliación para considerar este parámetro no se nos imagina demasiado complicado para futuras extensiones de la aplicación.

Por otro lado, partimos de la idea de construir un reconocedor que pudiera manejarse y modificarse manualmente por el usuario, pudiendo cambiar el valor de algunos parámetros de análisis incluso en tiempo real. Para ello, cada bloque debe interrelacionarse de forma clara y rápida con el resto, por lo que optamos por crear pequeños submódulos dentro de otros mayores, y que estudiaremos más detalladamente cuando veamos la arquitectura del software en próximas secciones.

Además, era claro que el calibrado correcto de los resultados sería muy dependiente de la cantidad y calidad de las pruebas y de las evaluaciones que se realizaran. Para probar los algoritmos, se prepararon y recopilaron un conjunto de materiales sonoros, los cuales debían contener eventos prosódicos, rítmicos y temático-musicales.

Para ello, he utilizado los siguientes materiales sonoros:

- *Sequenza para Soprano*, de Luciano Berio. Es una obra en la que el compositor italiano recopila sus investigaciones acerca de las posibilidades de la voz humana. Para Berio importa más el continente que el contenido, es decir, no hay mensaje, ni letra, ni temas musicales. Sólo hay articulaciones sonoras y sordas de diversos tipos con la voz, construyendo un conjunto vocal audible complejo. La ventaja de este material es que está escrito y podemos comparar los resultados con la partitura.

- *Naat Taksim*. Consiste en un canto de alabanza turco, con letra del poeta Rumy, interpretado por un maestro de canto. En él podemos estudiar líneas melódicas complicadas, con variaciones frecuenciales muy pequeñas, y muy ricas en timbres. También es útil para estudiar las micromodulaciones de tono típicas de la música islámica, para comprobar la robustez de los algoritmos.

- *Compilación de grabaciones propias de voz e instrumentos*. Para los casos en los que los anteriores materiales no eran suficientes, grabé algunas situaciones musicales y sonoras con mi voz y con un par de instrumentos.

- Un *generador de señales sinusoidales* basado en Pure Data. Son utilizados para comprobar los resultados que arrojan los algoritmos en fase de desarrollo y prueba.

- Además, se incluyó una evaluación muy buena de los resultados gracias a la amabilidad del Dr. Francisco Javier Sánchez en prestarnos su interesantísimo programa Mapatone, en el cual se pueden cargar escalas mediante parámetros e interpretar melodías sobre un piano modificado. De esta forma, pudimos comprobar minuciosamente la salida de la aplicación, disfrutando además de la música modal que generaba automáticamente.

Con estos materiales, y hasta el momento, ha sido posible hacer una discriminación profunda de los resultados de los algoritmos.

4.2 El muestreo y el preprocesado en tiempo real.

La entrada de datos de audio bajo el sistema operativo Windows, se hace mediante la aplicación de las instrucciones y funciones que aporta su API (Application Program Interface), en el apartado Multimedia, y principalmente utilizando la librería dinámica *winmm.dll* , si bien la documentación que aportan los fabricantes es escasa en este apartado. También se puede realizar con instrucciones de más alto nivel, como es el caso del comando *soundsc*, pero con menores posibilidades y mucho peor rendimiento.

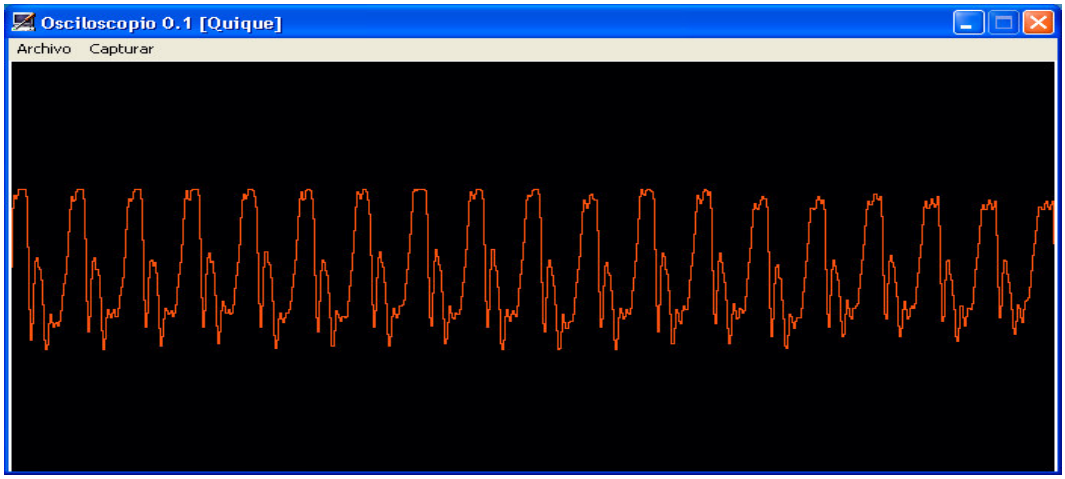
El primer objetivo concreto, era la programación de un osciloscopio en tiempo real. De esta forma sabría que los datos son correctos, y siempre nos vendría bien para estudiar las cualidades de la voz, y con los requisitos que nos impusiéramos. Por otro lado, sería el programa básico sobre el que habría que ir montando y superponiendo el resto de funciones de la aplicación.

Al final, se programaron dos versiones diferentes, una en lenguaje Visual Basic y otra en lenguaje C. La decisión de utilizar Visual Basic fue tomada porque el módulo de muestreo debía ser equivalente en ambos lenguajes, y a priori tratar datos en Visual Basic debía ser más fácil. Una vez conseguido el objetivo, pasamos los resultados a Visual C++ bajo la plataforma .NET, y programamos un osciloscopio mucho más estable, rápido, y por qué no, más vistoso. La programación fue costosa, hubo bastantes problemas con el tiempo real y con las estructuras de datos que utiliza el driver de la tarjeta de sonido, pero al final, el esfuerzo mereció la pena.

En la figura que vemos más abajo se muestra una captura del programa en funcionamiento.

El programa definitivo, fue programado en Visual Basic, por su fácil interacción con las ventanas gráficas y de posibilidades de incluir botones, barras deslizantes, etc, necesarias para controlar parámetros. Por tanto, esta versión utiliza el núcleo muestreador de Visual Basic.

En cuanto al módulo de preprocesado, no se han incluído técnicas de ningún tipo más de que cambio del formato de las muestras de entrada para hacerlo más lógico con la sintaxis de análisis del programa de estima de frecuencia fundamental. Fueron planteadas técnicas como el filtrado paso bajo de frecuencia, pero al comprobar la robustez de los resultados en casi todos los casos, se relegó esta opción al capítulo de mejoras que pueden ser incorporadas al programa en versiones posteriores.



4.3 Cómo estimar la frecuencia fundamental de la voz.

En el algoritmo de estimación de frecuencia fundamental encontramos el corazón operativo del reconocedor de modos musicales. Si somos capaces de estimar la frecuencia fundamental de un segmento de voz cantada con la suficiente precisión y seguridad, los pasos posteriores de mayor nivel, serán más fáciles de implementar.

El primer problema que se presenta cuando se afronta el estudio de la estimación de frecuencia fundamental, surge al darse cuenta de que hay gran variedad de tipos de algoritmos, basados a su vez en multitud de técnicas diferentes, y ninguno realmente es totalmente efectivo (ni totalmente documentado). En general, la eficiencia de un algoritmo de estima frecuencial no es algo absoluto: depende mucho de la base de datos con la se trabaje, de la inteligencia de decisión que se superponga y de las condiciones de calidad del material sonoro a evaluar.

Por otro lado, casi todos los centros de investigación utilizan algoritmos basados en técnicas diferentes según la aplicación objetivo sobre la que trabajen. En resumen, no es lo mismo utilizar dos algoritmos para un mismo rango de frecuencias fundamentales, o para voces de estacionariedad muy diferentes.

Algunas tesis doctorales reflejan hasta dos docenas de algoritmos diferentes, por tanto, se puede decir que todavía no hay una técnica que llegue a satisfacer a toda la comunidad de investigadores. Surge por tanto, el reto de elegir una técnica de entre todas las posibles.

Entre el conjunto de algoritmos que más páginas han dado a la literatura técnica podemos encontrar:

- Algoritmos que utilizan predicción lineal para eliminar la influencia del tracto, como el SIFT.
- Algoritmos que utilizan estimación espectral mediante herramientas derivadas de la FFT.
- Algoritmos que basados en parámetros de análisis, como el basado en Cepstrum.
- Algoritmos que analizan en el dominio temporal la forma de las ventanas, como la correlación o la autodisimilitud.
- Algoritmos basados en filtrado frecuencial por bandas, como los basados en Comb Filters.
- Algoritmos basados en Modelos de Ondas Sinusoidales Armónicas (Harmonic Sinewaves Models).

La decisión primera fue la de evaluar los resultados que arrojaban algunas de las familias para el tipo de material sonoro con el que contaba, y tomar aquél que más ventajas presentara. Para ello, simulé en unos casos los algoritmos en MATLAB y en otros los programé directamente en lenguaje C y Visual Basic.

Las evaluaciones duraron varios meses de intenso trabajo, al principio con pocos frutos, pero con paciencia. Con el tiempo se fue llegando a ciertas conclusiones importantes para la evolución del programa: de la idea original de búsqueda de un algoritmo que resolviera todo el problema por sí mismo, se pasó a un criterio y un modelo del problema diferente. Como resumen:

- Ningún algoritmo nos ofrece la frecuencia fundamental, sino que nos da ciertas medidas que podemos modelar dentro de una variable aleatoria, y de las cuales no podemos decir nada más que los resultados de operar con ciertos estimadores. El problema principal es que no podemos calcular la frecuencia fundamental, sino sólo estimarla. Como tal, los resultados son estimaciones, e incluyen cierto porcentaje de error provocado por el mismo algoritmo y por la decisión que después tomemos.
- Partiendo de la premisa anterior, lo más importante es centrarse en las técnicas pre y post estimación algorítmica para hacer robusto al modelo de resultados de las medidas. Centrarse en buscar un algoritmo demasiado

exacto no tiene sentido porque, aunque lo consigamos, tendremos errores grandes fruto de decisiones erróneas entre armónicos, o errores de decisión entre segmentos sonoros y sordos. Por tanto, hemos de encontrar un algoritmo sencillo y con cierta calidad de exactitud en los resultados, para centrarnos más en dotar de inteligencia a la decisión.

- Tras varios meses de estudios sobre los más famosos algoritmos, se tomó la decisión de abordar un algoritmo de análisis temporal, por ser arrojar buenos resultados y por ser intuitivo de programar.

En las siguientes secciones vamos a documentar los estudios sobre los algoritmos, de forma que podamos sacar conclusiones acerca de los problemas que cada uno de ellos comporta y a forma de comparación entre diversas técnicas posibles.

4.3.1 El algoritmo de estimación espectral.

El primer algoritmo a evaluar estaba basado en estimación espectral. Elegimos este primer algoritmo porque parecía el más intuitivo para empezar, quizá fruto de la continua presencia de los espectros en los planes de estudio de esta ingeniería.

De esta forma, me propuse construir una función que me permitiera controlar todos los parámetros de la FFT desde su llamada, y que además pudiera incorporar algunas mejoras importantes, como el diezmado o la interpolación.

Los parámetros que utiliza la función son:

- el fichero de entrada a analizar: en formato .wav, llamado desde el directorio de trabajo
- el número de puntos para calcular la FFT, lo cual me permite cambiar la precisión del algoritmo.
- la frecuencia máxima que pensamos que va a tener la señal. Es un parámetro necesario para, por un lado representar mejor la señal en la figura, y por otro para desechar resultados que se salen del margen de frecuencia impuesto.
- el tamaño de la ventana de análisis, lo cual me permite cambiar el número de iteraciones sobre el algoritmo, y obtener un espectro más definido en los armónicos.
- el factor de diezmado, necesario para reducir el número de operaciones, y a modo de interpolación.

El algoritmo decide cuál es la frecuencia fundamental en cada ventana, analizando espectralmente los picos y decidiendo como bueno el máximo de menor frecuencia, o el máximo global. Sin embargo, esta decisión suele ser equivocada en muchos casos. La aparición del segundo y tercer armónico con mayor energía que el fundamental, hace que la decisión sea equivocada, y que los saltos indiscriminados de una o dos octavas sean habituales.
Por tanto, se hacía necesaria la inclusión de un módulo de inteligencia adicional, que analizara los resultados por bloques, que identificara los saltos de octava y decidiera que corresponden a la nota anterior, pero con un contenido armónico acentuado en esa frecuencia.

Por otro lado, el análisis espectral tiene el defecto de que en realidad estamos discretizando el espectro continuo, y nos quedamos con unas muestras equiespaciadas. Ocurre entonces, que las frecuencias fundamentales rara vez coincidirán exactamente con las de los puntos muestreados, y el resultado será aproximado. Pero también, el resultado se ve abocado a la variación continua si la frecuencia fundamental no es perfectamente pura e invariante, saltando a menudo la decisión entre puntos de frecuencia que se encuentran consecutivos. Esto me obligó a incorporar un módulo de inteligencia que detectara estos cambios pequeños en la decisión y que, bien por ser inapreciables o por falta de precisión en el muestreo del espectro, no debía considerar. Otra forma de eliminar este problema es el uso de interpolación entre puntos del espectro, de forma que podamos sacar más puntos intermedios, aunque en la FFT no los tengamos.

Además, consideré importante calcular el módulo de la energía de la señal en cada ventana, lo cual me ayudaría a saber si había una nueva pronunciación de la nota musical, y así identificar eventos nuevos.

De esta forma, a la salida deberíamos tener un conjunto de frecuencias fundamentales que consideramos las correctas. Además, debemos agrupar resultados parecidos y resumirlos en una frecuencia representativa, hasta que aparezca una variación tal de frecuencia que se asocie a otro representante.

Sin embargo para probar el algoritmo, de poco nos sirve hablar de frecuencias en música, por lo que era necesario programar otro módulo que escribiera las notas en notación musical occidental, en la que se escribe la nota y el número de octava según el estándar MIDI.

Por último, para poder comparar diferentes ejecuciones de la función con diferentes parámetros, una vez terminada la ejecución se devuelve la resolución con la que trabajamos, así como la frecuencia máxima que el algoritmo es capaz de detectar, por si estamos fuera del margen posible.

A continuación se puede observar un ejemplo de ejecución. En el fichero de entrada, una guitarra clásica interpreta lenta y claramente las siguientes siete notas musicales:

SOL 3 – LA 3 - SI 3 – DO 4 – SI 3 – LA 3 –SOL 3

(El número que aparece tras la nota identifica la octava. Como referencia se toma LA3 *"nota la, tercera octava"* = 440 Hz)

En el contorno de frecuencia fundamental, se aprecian dos saltos de octava en las notas SOL y LA, pero el resto de decisiones son correctas. También se aprecian variaciones de energía en los puntos donde se pulsa la cuerda, y que va disminuyendo progresivamente.

esp_display2('guitarra.wav',1024*2,2000,500,4)

frecuencia_fundamental_Hz_MIDI =

2.6917 0

```
  193.7988   55.0000
  382.2144   67.0000
  398.3643   67.0000
  209.9487   56.0000
  218.0237   57.0000
  427.9724   69.0000
  212.6404   56.0000
  242.2485   59.0000
  255.7068   60.0000
  244.9402   59.0000
  239.5569   58.0000
  244.9402   59.0000
  207.2571   56.0000
  215.3320   57.0000
  427.9724   69.0000
  193.7988   55.0000
  382.2144   67.0000
```

notas_musicales =

CERO - SOL3 - SOL4 - SOL4 - SOL#3 - LA3 - LA4 - SOL#3 - SI4 - DO4 - SI4 - LA#3 - SI4 - SOL#3 - LA3 - LA4 - SOL3 - SOL4 -

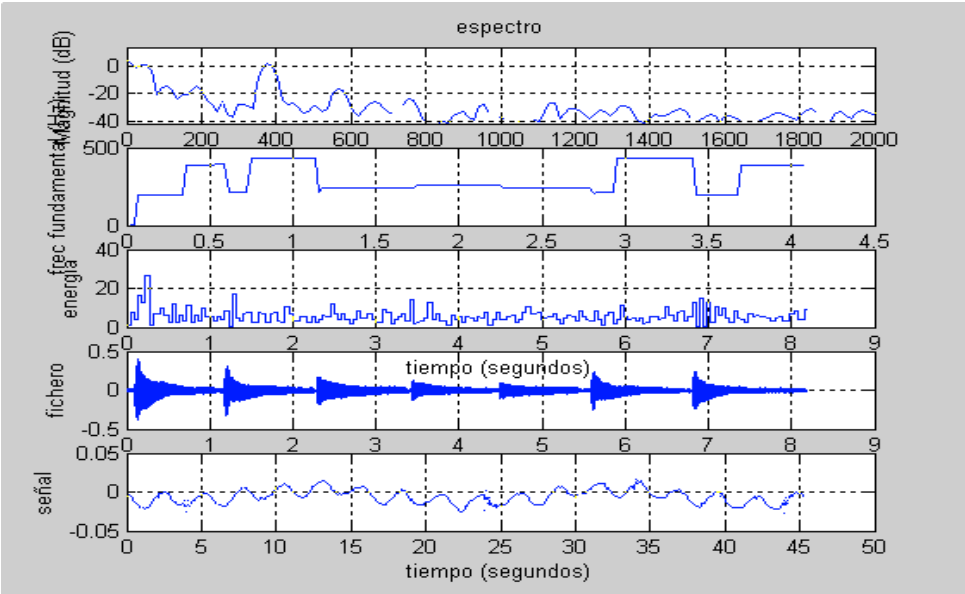
resolucion_Hz =

5.3833

frecuencia_limite =

11025

ok¡ v0.1

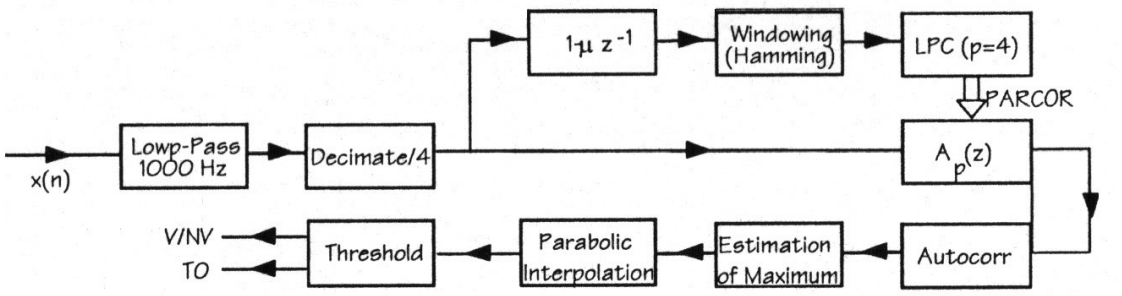


El algoritmo es una buena aproximación de partida, pero sigue adoleciendo de los problemas de la precisión de la FFT y de la estimación difícil de los máximos, por lo que me propuse programar un algoritmo basado en otra familia.

4.3.2 El algoritmo por predicción y el de análisis en dominio temporal.

La documentación abundante en el caso del algoritmo SIFT, hizo que eligiera esta técnica como objetivo. Por otro lado, el algoritmo SIFT si no se utiliza la predicción, se convierte en una técnica de estimación el dominio del tiempo mediante una correlación (si es que así terminamos encontrando la frecuencia fundamental), por lo que nos servía a la vez para estudiar ambas familias.

El algoritmo SIFT fue programado también en MATLAB, por la sencillez que tiene el hecho de que ya está implementada la autocorrelación rápida (basada en FFT), y el cálculo de los coeficientes de predicción mediante unas instrucciones del lenguaje. Por otro lado, ya teníamos programado el bucle de análisis por el algoritmo anterior. A continuación podemos ver el esquema seguido.



En nuestro caso, los parámetros de ejecución fueron:

- el tamaño de la ventana. Cogiendo más muestras en cada ventana, la correlación era mucho más grande, pero tenemos el problema de que si la voz no es demasiado estacionaria en la trama, se mezclan resultados.

- el ancho de banda del filtro paso bajo a la entrada. Un filtrado paso bajo elimina componentes armónicas que facilitan la detección, pero al ser demasiado exigentes, podemos hacer que el detector cometa errores sistemáticos.

- el número de coeficientes LPC. Un número bajo de coeficientes hace que la reconstrucción de la señal de excitación sea peor, pero con menos retardo de cálculo. Un número de coeficientes por encima de ciertos umbrales, no mejoraba sustancialmente los resultados.

- la frecuencia máxima y mínima aceptable. Para establecer ciertos criterios en la decisión, debemos discriminar aquellos resultados que no deben ser posibles,

para que si ocurre una decisión fuera de los márgenes, vuelva a procesar los resultados.

- la decisión de utilizar predicción o no. Sirve para decidir el tipo de algoritmo.

- el factor de diezmado. Se mejoraron bastante los resultados utilizando factores de diezmado hasta cierto umbral.

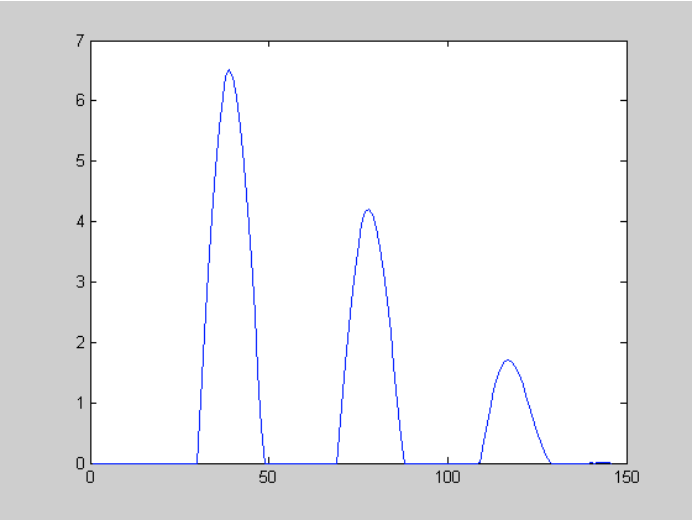
- la decisión de utilizar una interpolación de los resultados de salida. En general, los resultados de salida no son uniformes para un conjunto seguido de análisis de ventanas, aunque perceptualmente parece que la voz sí lo ha sido. Este problema es fruto de las micromodulaciones que sufren en frecuencia las notas al variar la velocidad del aire por el tracto vocal, o por variaciones inconscientes del cantante, que no acarrean información.

- el fichero en formato .wav a analizar.

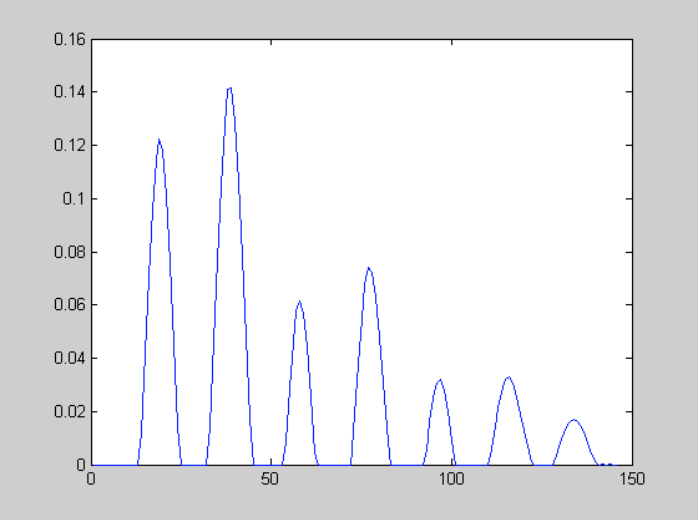
En la siguiente tabla se hace un resumen de los resultados con diferentes parámetros de entrada, considerándolos a efectos de estudio, independientes unos de otros.

	Valor	**Resultados**
Ventana	1 ms	Mala definición a baja frecuencia
	20 ms	Empieza a ser utilizable
	35 ms	Se obtienen los mejores resultados
	45 ms	Los resultados de las tramas empiezan a combinarse
Ancho de banda	< 600 Hz	Se pierde definición y los resultados empeoran
	> 600 Hz	Los resultados son más precisos
Número de coeficientes	< 3	Reconstrucción no válida en algunas tramas. Malos resultados en baja frecuencia
	3 – 10	Resultados aceptables. Correlación definida.
	> 10	Las mejoras en los resultados no son apreciables.
(fmin , fmax)	(0 ,2000)	Resultados correctos
	(0, 3000)	La energía en el origen empieza a falsear resultados
	(0, + 3000)	Los resultados no son estables
Predicción	si	Resultados aceptables
	no	Calidad final muy parecida al caso con predicción
Factor de diezmado	1	Efecto no apreciable
	2 - 4	Buena calidad para todos los ficheros analizados
	4 - 8	Calidad buena pero no garantizable en todos los casos
	> 8	La calidad empieza a rebajarse destacadamente
Interpolación	si	Resultados más intuitivos
	no	Resultados con toda la información del procesado

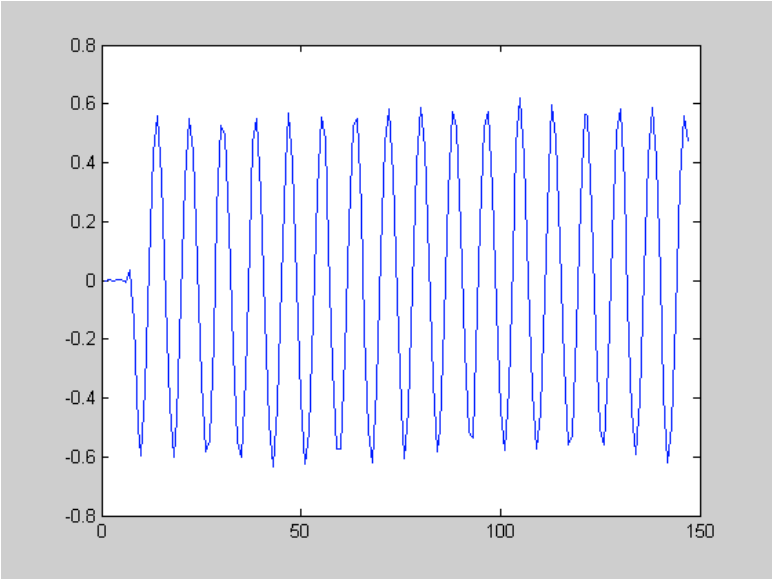
Para optimizar el algoritmo, tomamos la decisión hacer cero el valor de la autocorrelación para muestras que representaran frecuencias más altas de las esperadas, y así mejorar en el proceso de decisión.
En la siguiente figura vemos un ejemplo de la función de autocorrelación sobre un segmento sonoro. Como se observa, a baja frecuencia los resultados son buenos porque no interfiere en la decisión de máximos la energía media de la señal situada en el origen. Cuando el límite de frecuencia máximo es alto, la decisión suele hacerse mucho más difícil.



A continuación vemos un segmento de autocorrelación calculado sin predicción lineal. Vemos cómo un filtrado poco exigente, hace que los armónicos superiores tengan mayor energía que el fundamental, y que la decisión pueda ser incorrecta.



En la siguiente figura vemos la reconstrucción de la señal de excitación, fruto de la predicción lineal, sobre la que realizamos la autocorrelación. Se puede apreciar la típica forma de tren de pulsos glotales, aunque la señal no es exactamente periódica y regular: hay variaciones en amplitud y en frecuencia.



Los resultados de la evaluación de este algoritmo reflejaban cómo la decisión sonora/sorda era incorrecta en muchos casos, fruto de que los umbrales de decisión de frecuencia máxima y mínima deben estar muy ajustados para conjugar buenos resultados en la frecuencia fundamental y en la decisión sonoro/sordo. También se comprobaba cómo la aparición de ciertos armónicos mal filtrados provocaban el falseamiento de los resultados con saltos abruptos no practicables en la práctica.

Como conclusiones de este estudio de estas dos familias de algoritmos, podemos decir:

- Los resultados correctos son más precisos que en el caso de estimación espectral, lo cual es fundamental para el tipo de aplicaciones requeridas.

- La decisión de utilizar predicción lineal para calcular la autocorrelación sobre la señal de excitación no arroja resultados mucho mejores a efectos prácticos, aunque sí funciones de mucha mayor resolución, sobre todo a alta frecuencia.

- En el caso de obtener resultados correctos, lo principal es implementar un algoritmo de decisión inteligente, adecuado para la aplicación particular al que va destinado.

- El coste computacional de la utilización de la autocorrelación y de la predicción lineal es elevado, aunque afrontable con la tecnología actual. Por otro lado, la programación de las funciones de cálculo, en un lenguaje de programación de no muy alto nivel, podría ser muy complicada.

Por último, hay que reconocer que restaría por afrontar el análisis de la familia de los algoritmos basados en parámetros específicos, como pueden ser los basados en Cepstrum. Sin embargo, declinamos la opción de simular o programar otro algoritmo para este caso, pues la experiencia ya había sido constructiva, y la practicidad y los plazos temporales empezaban a afectar al plan de ejecución del

proyecto. La decisión fue entrar de lleno en un algoritmo temporal de menor coste computacional y de menor complejidad a la hora de programarlo. La solución fue el algoritmo basado en la función autodisimilitud con procesado adaptativo (AADF).

4.3.3 Estimación de tono mediante la función autodisimilitud.

Este algoritmo se basa en la aplicación de la función de autodisimilitud para el cálculo de la frecuencia fundamental, desarrollado por el Dr. Francisco Javier Sánchez en su tesis doctoral. En los siguientes apartados nos adentraremos en las bases matemáticas que soportan a la función y sus propiedades, así como la forma de construir un algoritmo como núcleo de un estimador de frecuencia fundamental para la voz.

4.3.3.1 La función de disimilitud.

La función de disimilitud es una función que evalúa el parecido entre las formas de dos señales que encontramos en dos determinadas ventanas de análisis. En realidad, sigue la misma intención que el cálculo de la correlación entre segmentos de datos. Generalizando, podemos decir que vamos a ir comparando los valores de la aplicación de una norma sobre diferentes segmentos de una señal.

La idea principal, es que una señal periódica va a ser muy parecida de periodo en periodo, y que si somos capaces de encontrar una función que evalúe ese parecido para cada periodo posible, encontraremos una forma de medir la frecuencia. Por tanto, la función, más que la frecuencia de la señal, evalúa para qué periodo la señal es más periódica.

Para entender el algoritmo es necesario explicar y desarrollar ciertos principios matemáticos, que abordamos a continuación.
El espacio matemático que vamos a trabajar es el de los números reales *discretizados,* en el segmento de datos que podemos abarcar, y en el cual podemos definir una métrica y una norma a aplicar. Su definición sería:

$$\| f \|_{p,w} = (\Sigma_{Sw} w \, | f_t |^p)^{1/p} \qquad 1 <= p < infinito$$

$$w_t > 0 \qquad \text{si } t \text{ está dentro de } Sw$$

$$w_t = 0 \qquad \text{en otro caso}$$

$$\| f \|_{p,w,x} = (\Sigma_{Sw} d_x w \, | f_t |^p)^{1/p}$$

Donde como vemos, *w* es la ventana de análisis sobre la que aplicamos la norma.

El valor de esta norma significa más o menos el valor colectivo del valor absoluto de las muestras, que está relacionada con el concepto de energía por ventana.

Por otro lado, esta norma gozará de un conjunto de propiedades, que nos sirven para sacar nuevas conclusiones, y para poder relacionar el resultado de las normas sobre diferentes segmentos de datos. Algunas de ellas son:

1. $\| 0 \| = 0$ ***la p-norma de señales nulas es nula .***

2. si $\| f \| = 0$ ***significa que*** $f = 0$ ***sólo dentro de Sw (lo que hace que las p-normas sean sólo seminormas).***

3. $\| \alpha f \| = \alpha \| f \|$ ***para todo*** α ***real y positivo***

4. $\| -f \| = \| f \|$ ***para todo*** α ***real y positivo***

5. $\| f - g \| <= \| f \| + \| g \|$ ***desigualdad de Minkowsky***

6. $\| f - g \| >= | \| f \| - \| g \| |$ ***derivada de la desigualdad de Minkowsky.***

y sobre todo

$$| \ \| f \| - \| g \| \ | \ <= \ \| f - g \| <= \ \| f \| + \| g \|$$

que vamos a utilizar para construir la función de evaluación de periodicidad.

La norma de la diferencia de dos señales (su distancia en el espacio métrico normado *S*) crece cuando sus formas de onda se diferencian. De esta forma, la función es un índice bueno de desigualdad de la forma. Pero también, está afectada por sus amplitudes:

$$\| af - ag \| = \| a (f - g) \| = a \| f - g \|$$

Para evitar este efecto indeseable podemos usar la segunda expresión de Minkowsky, que nos proporciona un factor normalizado de la diferencia, para que su cociente siempre se comprenda entre 0 y 1.

$$\boldsymbol{0 <= \quad \| f - g \| \; / \; (\| f \| + \| g \|) \quad <= 1} \quad \textbf{con}$$

$$\boldsymbol{\| f \|} \textbf{ o } \boldsymbol{\| g \|} \textbf{ no nulos}$$

Llamemos a *D* al término central de las desigualdades anteriores, comprendido entre los valores 0 y 1. El valor 0 se alcanza cuando el numerador es nulo, lo que ocurre si f = g , y ambas señales coincidan. Así el valor 0 corresponde a similitud máxima o mínimo de disimilitud.
D sólo se define si las normas de f y g no son simultáneamente 0, una situación que anularía el denominador de *D* y el propio *D* indeterminado. Si solo uno de ellos es nulo, *D* será:

$$\| f - 0 \| / (\| f \| + \| 0 \|) = \| f \| / \| f \| = 1 \;\; ,$$

y entonces eso significa que cualquier señal es muy disímil matemáticamente al nulo.

Por otro lado, ¿ la función puede alcanzar el valor 1 ?. La respuesta es que sí,
cuando g = -f, porque entonces

$$D = \| f - (-f) \| / (\| f \| + \| -f \|) = \| 2 f \| / (\| f \| + \| f \|) = 2 \| f \| / 2 \| f \| = 1$$

(las señales opuestas son las más disímiles)

Sin embargo, encontramos que la expresión de ***D*** no es demasiado buena cuando el objetivo es comparar dos señales iguales pero de diferente tamaño. Este es el caso de si g es un fragmento de f: g = a f.
En este caso

$$D = \| f - a f \| / (\| f \| + \| a f \|) = (1-a) \| f \| / (1+a) \| f \| = (1-a) / (1+a)$$

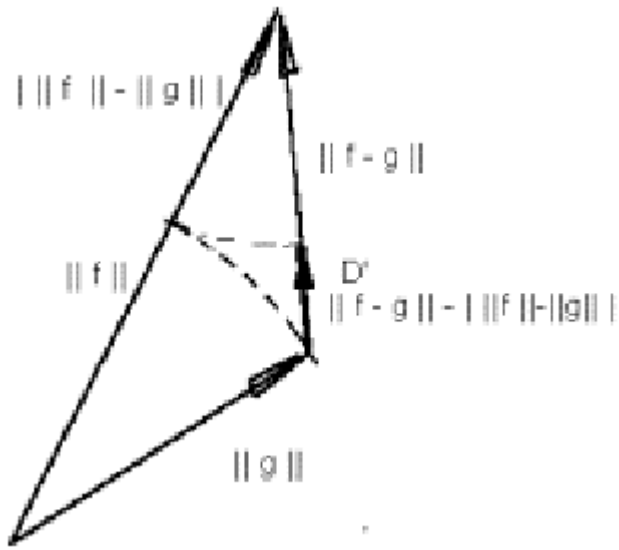
De forma que el resultado se vuelve cerca de 1 sólo cuando *a* es pequeño. Como queremos rechazar la influencia del tamaño en la desigualdad, para que el resultado sea más independiente, debemos imponer un valor bajo para el parámetro (sin llegar al extremos *a* =0, caso ya visto anteriormente para el que ***D*** =1, pero que no nos vale).

Encontramos otro elemento correctivo ideal en el primer término de desigualdades de Minkowsky: cuando restamos del numerador de ***D*** un número, aunque sin superar su valor, éste disminuye en una cantidad equivalente a la diferencia de las normas.

Así, definimos una nueva expresión ***D'***, como:

$$D'(f,g) = \quad (\; \| f - g \| - | \, \| f \| - \| g \| \, | \;) / \; (\, \| f \| + \| g \| \,)$$

Lo cual se puede representar vectorialmente:



Ahora el problema anterior se elimina: para g=af

$$D' = \; (\, \| f - a f \| - | \, \| f \| - \| a f \| \, |) \, / \, (\, \| f \| + \| a f \| \, |) \; =$$

$$= \; (\, (1-a)-(1-a) \,) \, \| f \| \, / \, (\;\; (1+a) \, \| f \| \;\;) = 0$$

lo que corresponde a una desigualdad mínima para señales homotéticas, como queríamos: sólo la forma cuenta, no el tamaño.

Sin embargo, ahora para una señal nula,

$$D' = (\, \| f - 0 \| - | \, \| f \| - \| 0 \| \, | \;\;) \, / \, (\, \| f \| + \| 0 \| \,) = 0 \, ,$$

Y los casos son similares, se igualan, es decir, algo no esperado ni deseable.

Para obtener una solución, debe hacerse un compromiso entre la inclusión del término substrayendo o su supresión: esto se hace con un coeficiente de escalado β comprendido entre 0 y 1. De esta manera, D' llega a su última expresión: la desigualdad entre dos señales se define como

$$D'(f,g) = (\| f - g \| - \beta \, | \, \| f \| - \| g \| \, | \,) \; / \; (\| f \| + \| g \|)$$

La representación del valor de **D'** como una función de a con β como parámetro se muestra eficiente con señales homotéticas de la forma f = a f,

$$D' = (\| f - a f \| - \beta \, | \, \| f \| - \| a f \|) \; / \; (\| f \| + \| a f \|) =$$

$$= ((1-a) - \beta (1-a)) \, \| f \| \; / \; ((1+a) \, \| f \|) = (1-\beta) \, (1-a) \; / \; (1+a)$$

obteniendo ya un valor igual a 0 para f = g, un valor igual a 1 para señales con signos opuestos, f = -g, y un valor igual a **(1-β)** para g = 0.
Para un valor de β = 0.5, se obtiene D' = 1/6 = 0.15, un valor bajo no lejos de 0 en el rango 0 hasta 1.
El hecho de incluir el factor β, nos obliga a imponer un umbral de decisión para la amplitud de la función. En próximos apartados nos adentraremos en las cualidades prácticas del algoritmo.

Por tanto, como resumen, podemos considerar que la función va a ir variando según las forma de la señal en cada una de las ventanas, y si ocurre que son parecidas en el tiempo la función podrá ir alcanzando mínimos. De esta forma, y si la señal es periódica, podemos ir encontrando el periodo de la señal.

4.3.3.2 La función de autodisimilitud.

La función de disimilitud adquiere el nombre de autodisimilitud cuando para su cálculo se toman dos segmentos de una misma señal desplazados en el tiempo.

Para caracterizar a la función, ahora hace falta especificar un parámetro de desplazamiento, al que vamos a llamar **τ**. De esta forma, la función se expresará así:

$$ADF_{f,t,p,w,\beta}(\tau) = \frac{\| f_{t+\tau/2} - f_{t-\tau/2} \| - \beta \, | \, \| f_{t+\tau/2} \| - \| f_{t-\tau/2} \| \, |}{\| f_{t+\tau/2} \| + \| f_{t-\tau/2} \|}$$

Tenemos una nueva función relacionada con las propiedades periódicas de f en t (alrededor de t), como deseábamos.

Si suponemos ese segmento f periódico con periodo T, cuando τ crece los segmentos se vuelven diferentes hasta que τ sea igual al periodo señalado T, y los segmentos seleccionados se vuelven más tarde iguales; el proceso continúa mientras lo haga la señal, así que podemos obtener una función periódica del retraso τ.

En todos los demás puntos, la ADF será positiva. Sólo se alcanza el valor 1 si f pasa a ser igual a -f dentro de otras ventanas posteriores. Esto pasa, por ejemplo, en señales sinusoidales, triangulares, etc, con valor medio nulo.

Si *w* (ventana de análisis de la norma) vale igual para todas las muestras en el periodo, la ADF en diferentes momentos *t* será la misma, porque para cada uno, sólo el orden de las muestras cambiaría el cálculo de las normas, no sus valores.

Para señales cuasiperiódicas, en las que nos falta parte de la información para considerarlas totalmente periódicas, puede esperarse que la ADF tendrá un carácter cuasiperiódico, porque las muestras varían despacio de periodo en periodo.

Por tanto, las señales periódicas presentarán mínimos nulos para todos los múltiplos de T, incluido 0. Para señales cuasiperiódicas, los mínimos (nulo sólo para t = 0) se esperan en múltiplos del cuasiperiodo. El más pequeño de los mínimos será el que tenga el carácter más periódico, tendrá más profundidad.
Los mínimos para los múltiplos altos de la frecuencia serán ciertamente mayores que para los múltiplos pequeños, porque los cambios de un cuasiperiodo se acumulan para cuasiperiodos lejanos.

Un problema teórico y práctico, es determinar la profundidad del valor de un mínimo para decidir que el signo procesado es ' *cuasiperiódico*' o no, una materia sin una respuesta definida. Es necesario imponer un umbral Cuando el mínimo es más pequeño que el umbral, el segmento f señalado será tomado como *cuasiperiódico* en el t instante, y si no*, aperiódico*. Es decir, primero es necesaria una discriminación entre segmentos sonoros y sordos.

4.3.3.3 Descripción de la estructura del algoritmo.

En los siguientes puntos, se describen los pasos del algoritmo básico de cálculo de la frecuencia fundamental a partir de la función autodisimilitud. Plantearemos además algunas cuestiones relacionadas con el algoritmo, como el uso de las ventanas, la posibilidad de utilizar estrategias adaptativas, y ciertos problemas que se presentan en su realización.

Para empezar, describimos los pasos que constituyen el esqueleto de funcionamiento secuencial del programa, animando a otros programadores a mejorarlo.

1. Suponemos las muestras de audio almacenadas en un vector, señal(n).

2. Seleccionamos un punto t, lo suficientemente adelantado como para permitir superponer una ventana de máximo del periodo posible (menor tono detectable).

3. Calculamos la función Autodisimilitud para este punto temporal t.

3.1 Elegimos dos nuevos puntos t1 y t2, separados por una misma distancia temporal al punto t, es decir, en $t - \tau/2$ y en $t + \tau/2$. Por tanto, el intervalo de tiempo entre estos puntos t1 y t2 será τ.

3.2 Centrados en t1 y t2 colocamos dos ventanas iguales (ver ventanas posibles más abajo) y multiplicamos cada muestra de las ventanas por el valor de la ventana en cada punto. De esta forma, seleccionamos dos vectores de señal que son los que vamos a comparar.

3.3 Calculamos la energía (utilizando un conjunto de muestras o por un método más sofisticado) de cada uno de los vectores, sus normas, y la norma de la diferencia de los dos vectores (incluso podemos calcular la diferencia de los vectores antes de aplicarles la ventana para ahorrar cierto tiempo).

3.4 Con esos valores, N1, N2 y D (sumandos del numerador y denominador respectivamente), calculamos la Autodisimilitud de los segmentos de señal centrados en t1 y t2, para un valor particular de τ.

3.5 Repetimos los pasos 3.1 a 3.4 para dos nuevos puntos t1' y t2', ahora separados por un retardo τ_1, obteniendo un nuevo valor para la autodisimilitud.

3.6 Cambiando el valor de τ obtendremos nuevos valores de la autodisimilitud, por tanto calculamos la función con la variación de τ. Es preciso recordar que como hemos obtenido todos los vectores de comparación de la misma señal, podemos decir que estamos utilizando la función Autodisimilitud, ADF.

3.7 El mínimo de esta ADF apuntará a un valor de τ para el cual la similitud entre dos fragmentos de señal separados exactamente τ es máxima. Sabemos además, que en una señal periódica los segmentos separados por una valor de τ igual al periodo serán los más parecidos, por tanto su similitud será máxima y su disimilitud será mínima.

3.8 El mínimo de la ADF apunta a un τ igual al periodo de una señal periódica, o al pseudo-periodo de una señal pseudo periódica, en un punto temporal determinado por t.

4. Para orto determinado momento t' , se repiten los pasos 2 a 3.8, obteniendo de la misma forma el pseudo periodo para este momento en la misma señal.

5. Repitiendo este proceso en intervalos de tiempo iguales, encontraremos la señal de evolución del tono que corresponde a la señal analizada.

6. Y el proceso se acaba cuando el usuario lo desee.

- **El algoritmo adaptativo : AADF**

Llamamos **AADF** o Adaptative Autodissimilarity Function en sus siglas anglosajonas, a un algoritmo para calcular la ADF únicamente sobre un conjunto menor de retardos que podemos seleccionar.

Por tanto, se calcula la ADF sobre un entorno donde esperamos que el mínimo va a ocurrir. Este entorno, se toma por supuesto teniendo en cuenta el último tono estimado, puesto que esperamos que el tono varíe suavemente, no abruptamente. En términos analíticos, significa que la variación de tono entre puntos vecinos está limitada. Por tanto, tomaremos la última estimación como el valor de τ alrededor del cual calcularemos la ADF. De esta manera, el proceso de cálculo sigue al tono, ahorrándonos de calcular la ADF para muchos valores de τ que no hacen falta.

El entorno a tomar alrededor del último tono estimado depende del intervalo de tiempo entre el que se calcula la ADF y de la variación máxima de la señal que esperamos, para lo cual necesitamos de un aporte de información externa (no será lo mismo una melodía de contrabajo que la de un pájaro).

- **La elección de la ventana.**

Además del entorno adaptativo de cálculo, existe otra importante posibilidad de adaptación en el cálculo de la ADF. Esta la encontramos en el soporte de la ventana (puntos del segmento de señal que caen donde la ventana *w* no es nula), pudiendo hacerlo proporcional al retardo $\boldsymbol{\tau}$.

La razón de aplicar esta segunda adaptación reside en un aspecto perceptivo: el tamaño de la ventana de observación de un fenómeno se adapta siempre al tamaño de ese fenómeno esperado. Por ejemplo, cuando buscamos en un mapa un nombre, adaptamos nuestra ventana de búsqueda en función del tamaño del item esperado: grande para nombres de países y pequeña para nombres de pequeñas ciudades. En música también se realiza esta adaptación: ventanas de segundos para ritmos y de décimas de segundos (o menores) para la detección de tono.

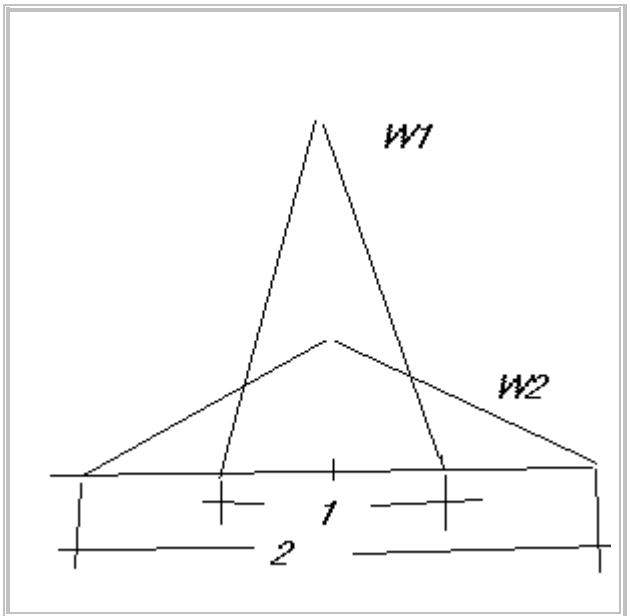
Trabajando sobre señales, cuando buscamos un fenómeno periódico, usaremos pequeñas ventanas para tonos altos y grandes ventanas para bajas frecuencias. Tendremos por tanto:

$$S_w = k \,.\, \tau$$

Además, para no modificar la energía de segmento ante la variación del soporte, tenemos que acordar un método de variación de su altura, es decir, de sus valores de promediado, para que siempre se cumpla que todas las ventanas tengan la misma área.

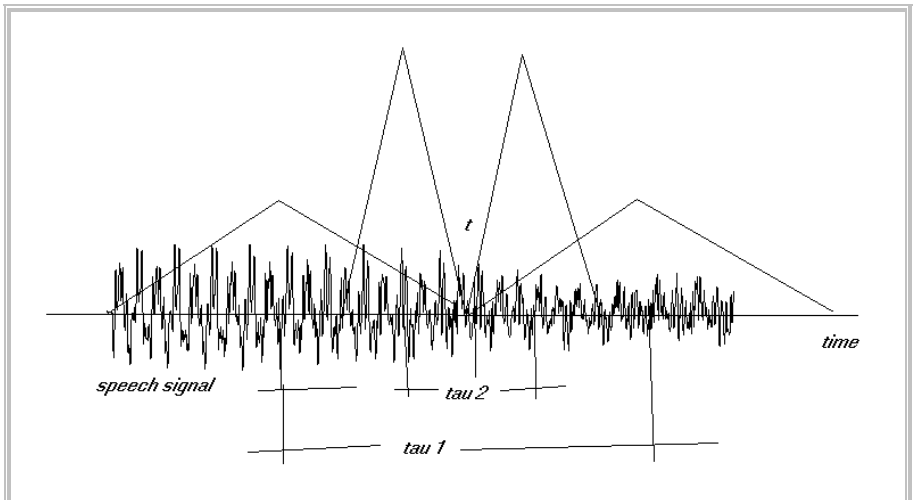
Necesitamos por tanto, que para una familia de ventanas con la misma forma, se debe cumplir la ecuación siguiente:

$$w_\alpha = 1/\alpha \, .V \, (t/\alpha)$$



Dos ventanas de la misma familia (triangular) con el mismo área, forma y soportes proporcionales a su altura.

En nuestro programa, por falta de tiempo, se acordó la elección de una constante k que toma valores alrededor de uno, lo que significa que el soporte es similar al periodo buscado T. Valores mayores provocarán un promediado de los valores del tono entre varios periodos.



Situado de ventanas para calcular la ADF en dos puntos, es decir, para dos valores del retardo τ.

4.3.3.4 Inteligencia añadida a la estimación.

A la simple estimación de la frecuencia fundamental, es necesario añadirle algunas funciones de inteligencia con el objetivo de hacer más robusto al resultado final.

Una de las misiones de esta sección es la decisión entre segmentos sonoros y sordos. Esta decisión se hace teniendo en cuenta un umbral de amplitud que imponemos a la función de autodisimilitud. Si la función no es mayor que este umbral, la decisión será de segmento sordo, y el resultado no se incluirá en la memoria de tono, ni tampoco contribuirá a la evolución del tono en su ventana correspondiente. El valor de este umbral puede ser modificado manualmente en tiempo de ejecución, según veamos que necesitamos ajustarlo a las características específicas del audio que tengamos grabado o estemos grabando.

Por otro lado, debemos discriminar entre resultados buenos y malos. El único problema de esta parte, es definir lo que es un buen resultado, o quizá lo que es uno malo. No sabemos la frecuencia que va a aparecer en el momento siguiente, y no podemos poner condiciones demasiado exigentes. Muchas soluciones al problema han sido escritas en tesis doctorales y *papers*, de muchos autores. Distinguimos entre muchas otras, las aportaciones de A. Acero con su Modelado Probabilístico a Posteriori, o el Modelado Bayesiano de Godsill y Davy del tono, incluidos en las referencias bibliográficas. En nuestro caso, después de estudiar las opciones de incluir de estas técnicas, decidimos probar primero con la inclusión de reglas lógicas que controlaran el discurso musical.

Reglas lógicas son por ejemplo, evitar saltos mayores de una quinta superior o inferior entre dos estimaciones , y controlar la variabilidad de la afinación del locutor con técnicas de promediado. Estas reglas, aunque básicas, se apreciaron como muy acertadas, puesto que en general la voz cantada no suele realizar saltos grandes, sino que progresa siempre por el camino de mínima distancia en frecuencia.

Después de probar el algoritmo con estas reglas, comprobamos que los resultados eran realmente mejores, por lo que dejamos la aplicación de los citados métodos estadísticos para posibles ampliaciones futuras del programa.

El problema de las reglas lógicas es que toda la información que se aporta es externa, no inherente al propio algoritmo, no habiendo límites para ellas, porque este modelo en realidad no puede ser expresado con claridad.

Por último, con el fin de evitar evoluciones de tono demasiado abruptas, decidimos incluir un promediado de los valores, que nos diera una idea del camino que lleva el tono en cada momento. Este promediado se hizo sólo sobre dos estimaciones, mediante un promediado de tipo media, lo que arrojó resultados adecuados a nuestros intereses. Se comprobó también la utilidad de promediados de orden superior, pero los resultados no fueron tan satisfactorios.

4.3.3.5 La programación práctica de la aplicación.

El algoritmo fue efectivamente programado en Visual Basic, aunque el núcleo de procesado es trasladable a cualquier otro lenguaje de forma fácil y directa.

Previendo los posibles problemas en la ejecución en tiempo real, una de las primeras decisiones de diseño fue la de crear controles interactivos, como botones, barras deslizantes y otros, que pudieran modificar los valores de los parámetros de la función ADF en tiempo real.

Decidimos fijar el tamaño de la ventana a un valor de

2* tau_max * alfa + exceso (en muestras)

un tamaño que tiene que ver con el retardo τ máximo que vamos a evaluar, con el parámetro $\alpha \in [0,2]$, encargado del solapamiento entre ventanas, y por una variable de exceso positiva mayor que uno, para evitar problemas con los bordes de los intervalos.

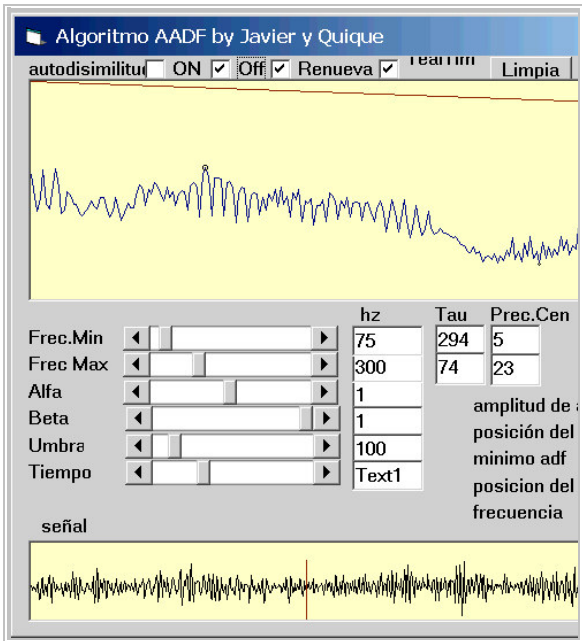
Además, como tenemos que evaluar la función para un conjunto de retardos $\boldsymbol{\tau}$, es necesario gestionar la frecuencia máxima y mínima que esperamos de la voz, lo que nos da a su vez un τ o periodo máximo y mínimo a evaluar. La elección de un margen de frecuencia adecuado es fundamental para garantizar el buen funcionamiento y su velocidad. En este caso, se han programado unas barras deslizantes para que el usuario elija el intervalo antes de comenzar la ejecución.

De la misma manera, se puede controlar el parámetro de solape $\boldsymbol{\alpha}$ y el de amplitud $\boldsymbol{\beta}$, de forma que podemos cambiarlos en ejecución, y así ver la variación en la salida.

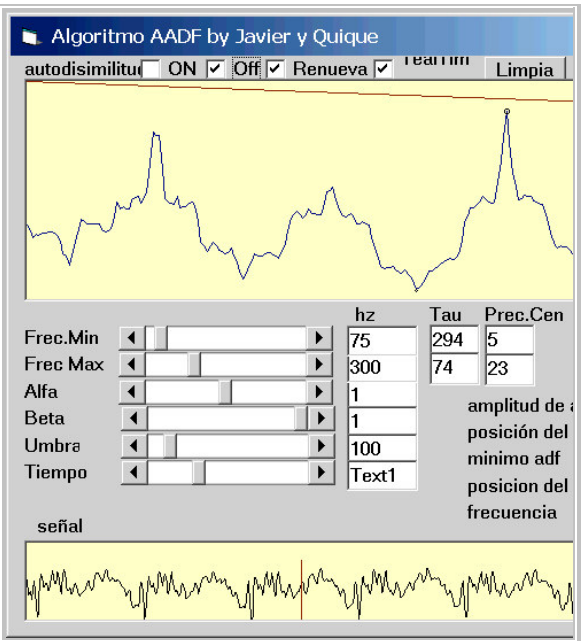
Para controlar el inicio y final de la ejecución, se programaron diferentes *checks* y botones. Además, en dos ventanas se muestra gráficamente la señal de entrada y la función de autodisimilitud. El aspecto de esta primera ventana se puede ver en la figura que podemos ver más abajo.

Como se puede observar, hay además otras pequeñas ventanas, que dan información acerca de los buffers que estamos utilizando, los valores de los máximos y mínimos de la función, y sobre todo la ventana que devuelve la frecuencia fundamental, inmersa ya en un módulo de postprocesado de los resultados, y que explicaremos un poco más tarde.

De la aplicación de este algoritmo sobre una señal vocal mostramos algunos resultados, para sonidos sordos y sonoros, los cuales también se pueden ver en la siguiente figura.



Resultado para una [s].



Resultado para una [e].

Las pruebas con este primer programa fueron muy buenas, el algoritmo era bastante exacto, y respondía en tiempo real a las variaciones de frecuencia de los archivos de audio o la señal de micrófono.

- **Problemas del algoritmo y estrategias**

Los problemas de los que adolece nuestro algoritmo, son similares a los que hemos encontrado en otros algoritmos que hemos estudiado antes:

1. Presencia de armónicos que suplantan al tono verdadero, especialmente aquellos que son enfatizados por los formantes.,

2. Presencia de sub-armónicos que desvían la decisión del tono.,

3. Presencia de ruido en la señal periódica, que reduce la fiabilidad del algoritmo.

4. Determinación del umbral de decisión sonoro-sordo. Si el umbral es demasiado bajo, se tomarán segmentos sordos como sonoros. Como los sonidos sordos pueden ser asociados con cualquier tono, fruto de su naturaleza ruidosa, puede aumentar aún más la aleatoriedad de los resultados.

5. Decisión entre resultados contiguos aleatorios. Es el problema de la discriminación de resultados, lo abordaremos en un apartado individual mostrando las estrategias propuestas.

6. Sentido de cálculo. Podemos tomar el sentido natural del tiempo, el contrario, o estrategias más complicadas para buscar resultados mejores que los obtenidos mediante

análisis continuo, sobre todo en segmentos donde la sonoridad sea más discutible, como fricativas sonoras, oclusivas sonoras o cuando la variación del tono sea muy rápida.

7. Adaptación al tipo de señal. No existe algo parecido a un algoritmo de estimación de la frecuencia fundamental universal, es decir, para todo tipo de señales sonoras. Debemos adaptar nuestro algoritmo a la señal que esperamos, en nuestro caso la señal vocal humana. Sin embargo, ni siquiera la señal vocal humana nos vale, debemos especificar más:

- Voz humana, con los subcasos de hombre, mujer y niños.
- Voz humana cantada, con los mismos subcasos y además incluyendo las tesituras: bajo, barítono, alto, tenor, contratenor, alto, contralto y soprano..
- Tonos animales, desde el más grande al más pequeño: ballenas, elefantes, toros, leones, tigres, monos, águilas, serpientes, búhos, murciélagos, etc.
- Tonos mecánicos, como rotores, motores, etc..
- Instrumentos musicales, grandes y pequeños, de cuerda, viento y percusión, de sonoridad continua y percutiva.

8. Problemas derivados de la construcción de nuestro algoritmo:

1. Elección del valor del exponente p en la definición de la norma.
2. Tamaño de la ventana y forma.
3. Valor de Alfa, que impone el soporte de la ventana.

   En nuestro caso, se toma un valor cercano a uno, por lo que el soporte de la ventana es cercano a tau.
4. Valor de Beta, que caracteriza la comparación entre segmentos retrasados
5. Filtrado de agudos
6. Incertidumbre en la calidad del tono estimado, apartado al que retrasamos su estudio a un poco más tarde.

- **Incertidumbre en la estimación del tono.**

El valor de un tono obtenido por todos los estimadores conlleva siempre un grado de incertidumbre alrededor de cierto valor central. ¿Por qué ocurre esto? Descartando errores en el cálculo o en la programación del algoritmo, podemos apuntar a ciertas fuentes de error:

1) La imposibilidad de calcular el tono exacto de una señal periódica usando una ventana de análisis limitada en el tiempo. Necesitaríamos una ventana temporal de longitud infinita para acertar el verdadero y exacto valor del tono.

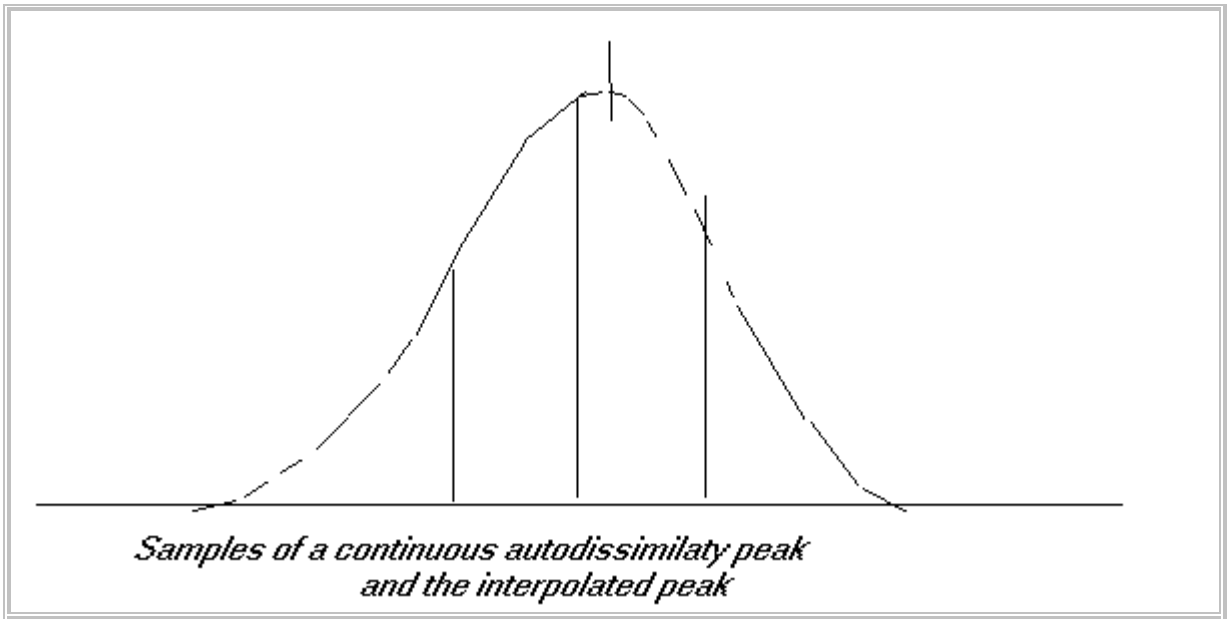
2) La posibilidad de que el punto de inicio del análisis “t”, donde empezamos a calcular el tono local, pueda caer en diferente parte del periodo de la señal para cada ventana. En cualquier caso, si desplazamos punto apunto el análisis, tomaremos para el cálculo diferentes partes del periodo, por lo que el resultado también puede ser diferente.

3) Las señales reales, como el canto o las interpretaciones con instrumentos, suelen tener habitualmente variaciones intencionadas del tono, como el vibrato. Además hay que añadir los procesados postgrabación que se añaden en la edición de la música en discos.

4) La frecuencia muestreo, puesto que limita los desplazamientos posibles o retardos para comparar dos segmentos de señal en el proceso de análisis. De hecho, el límite que nos impone es el retardo de una muestra, separada por el periodo de muestreo. Por otro lado, estamos obligados a que los retardos sean múltiplos de ese periodo de muestreo, sin posibilidad a priori de retardos intermedios.

Esta cantidad, podemos denominarla como resolución de la estimación del tono. Este límite puede ser reducido mediante técnicas de interpolación, de forma que podamos encontrar valores para la autodisimilitud entre dos valores consecutivos calculados mediante el algoritmo.

Este proceso de interpolación se ve en la figura siguiente:



Pico de autodisimilitud para la variable tau.

La verdadera expresión de la curva de la función de autodisimilitud nos es desconocida. Se pueden probar redondeos empíricos de varios tipos: sinusoidales, paraboloides, o del tipo *"sin(x)/x*". El número de puntos a tomar depende del carácter de suavizado que le queremos dar a la función. Normalmente de tres a cinco puntos suelen ser suficientes.

- **Valores de tono contiguos aleatorios.**

Una vez que hemos ejecutado el algoritmo sobre varios segmentos de la señal, debemos preguntarnos por la fiabilidad de los resultados que hemos obtenido. Como a priori podemos obtener cualquier tipo de valor de salida y no conocemos su verdadero valor,

hemos de modelar la forma de discriminar valores de tono, los cuales vamos a considerar que pueden distribuirse de forma aleatoria.

Hemos considerado dos tipos de discriminación:

A. Consideremos que la naturaleza del tono es continua, puesto que la variación tensión de los cartílagos arteoides que rigen el movimiento de las cuerdas vocales es continua. Por tanto, el tono que estimemos debe variar de forma continua en el tiempo y en una cantidad limitada, de forma que su gradiente también esté limitado.

Por tanto, una vez que estamos totalmente seguros de que hemos obtenido el verdadero tono (o el que así consideramos por conveniencia), debemos considerar que el siguiente estará incluido dentro de una ventana de valores posibles, lo cual ya nos da una cierta capacidad de discriminación de valores de tono (a la vez que es una forma básica de predicción).

De este modo, si incorporamos esta cualidad a nuestro algoritmo, podremos conseguir que sea aun más rápido, puesto que si sólo calculamos los retardos en la ventana de predicción, reducimos muchísimo el tiempo de procesado.

B. El otro acercamiento al problema genérico de elegir entre varios candidatos para el tono, es retrasar la decisión hasta que hayamos procesado posteriormente otros segmentos de la señal, y evaluar la consistencia de las familias de soluciones que se generan en cada caso.

En general, habrá resultados que rompan la característica de continuidad más que los otros, o que sean menos probables respecto a las reglas que queramos imponer. Es por tanto necesario la definición de las cualidades a esperar de los segmentos de la señal, haciéndose obligatorio el estudio previo de un conjunto grande y representativo de ejemplos de la señal, lo cual puede no ser posible a menudo.

Por tanto, el cálculo del tono se convierte ahora en el problema de la elección de candidatos a tono. Esta elección será función del tipo de evaluación que realicemos para elegir el mejor, la cual en general se limitará a la búsqueda de aquel candidato que maximice o minimice una función o un algoritmo de cálculo. Este paso generaliza nuestro problema y lo saca del ámbito del tratamiento de la voz, habiendo mucha literatura escrita acerca de posibles vías de resolución.

4.4 La aplicación de reconocimiento.

Si bien en las anteriores secciones hemos ido describiendo ciertos bloques funcionales de la aplicación, ahora es el momento de integrar todos los bloques y de explicar el funcionamiento conjunto.

En el anterior diagrama de bloques encontramos dos partes bien diferenciadas:

- En primer lugar encontramos el módulo de procesado en tiempo real con la intención de estimar la frecuencia fundamental a partir de segmentos de la señal

de voz muestreada. Sus resultados son guardados en una memoria de resultados y en un histograma de frecuencias, punto de partida del análisis del segundo módulo del programa. Este es el corazón de procesado de la aplicación. Sus características han sido explicadas anteriormente en los capítulos de muestreo y de descripción del algoritmo de estima de frecuencia fundamental.

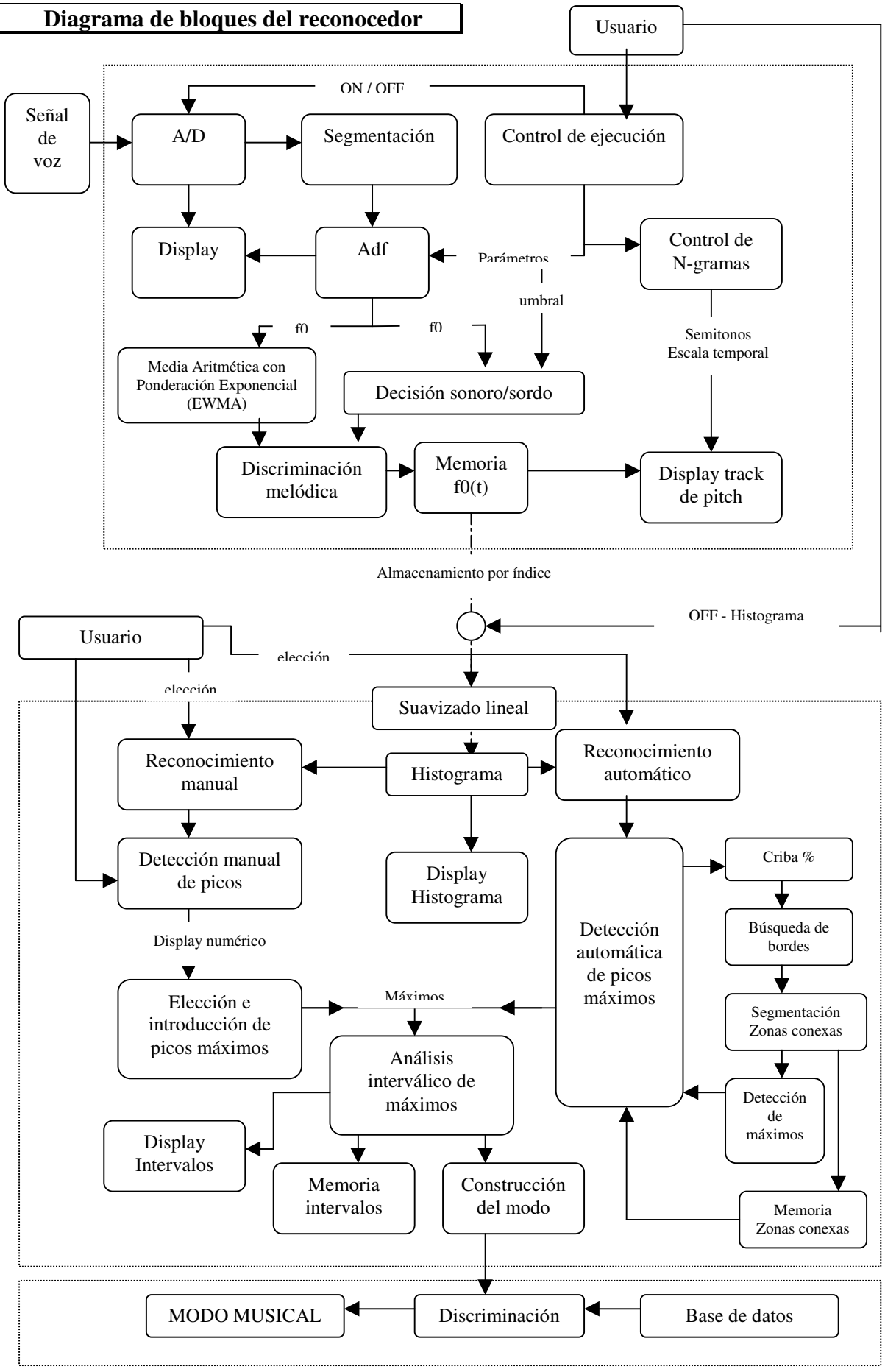
- En segundo lugar, el segundo módulo utiliza la información contenida en el histograma para sacar información de más alto nivel, consistente en encontrar las frecuencias que prevalecen en los segmentos analizados y calcular sus diferencias interválicas. Este módulo es el que nos ocupará casi todo el estudio en este capítulo, dando por sentado que se conocen ya las características del bloque de procesado.

  Este proceso se realiza mediante la interacción con el usuario, si bien éste puede participar de forma activa realizando el análisis de forma manual ayudándose del programa para elegir los picos del histograma, o bien, puede elegir detección automática, con lo que se ahorrará el proceso de selección de picos.

Por tanto, el segundo módulo utiliza la información de bajo nivel que el primer módulo le presta, de forma que pueda extraer la información de alto nivel que el usuario de la aplicación requiere. Sin embargo, para poder extraer información, antes se deben incluir los datos en un modelo adecuado del problema que queremos solucionar. Por eso, en esta sección también se explican ciertos modelos creados *ad hoc*, como el modelo de nota musical, el modelo de pico y el modelo de buena escala musical.

Una vez incluidos los datos en los modelos, se necesita de la interacción del usuario para manejar el programa, y que en ciertas partes del programa lo guíe, sobretodo en parcelas donde nos debatimos con conceptos de inteligencia artificial. Por este motivo, se ha decidido incluir una dualidad de funcionamiento manual – automática. Esta idea que surge del hecho de que la detección automática es un proceso más acercado a la inteligencia artificial que a nuestros conocimientos de procesado de señal, por lo que la fiabilidad de los resultados serán menores que haciéndolo de forma automática, lo cual por otra parte, ofrece una posibilidad de evaluación inmediata si los resultados que devuelve el programa en su versión automática no convencen al usuario.

Diagrama de bloques del reconocedor



4. 4. 1 La ventana de evolución de tono.

Una vez estimada la frecuencia fundamental, lo lógico es realimentar al usuario con el resultado, para que tenga idea de lo que está pasando. Como la estima es en tiempo real, la representación debe ser también en tiempo real.

Se optó por una representación de la evolución del tono dentro de una ventana gráfica, tomando como límite superior la frecuencia máxima y como inferior la frecuencia mínima. Para escalar los resultados, se optó por la creación de un N-grama, consistente en pintar las líneas que suponen la superación de un semitono en frecuencia, es decir, representar la evolución del tono en una ventana que recordara al ambiente de un pentagrama musical, pero en vez de utilizar cinco líneas, dividir toda la pantalla en dichas líneas. Como la evolución de los semitonos en frecuencia es logarítmica, las líneas de alta frecuencia están más cercanas que a baja frecuencia.

A continuación podemos ver un ejemplo de esta ventana.

De esta manera, el usuario puede ver la evolución del tono a lo largo del tiempo, en unas líneas que recorren los N-gramas. Adicionalmente, se representa en una ventana paralela la evolución del tono promediado temporalmente, para evitar la influencia de la variabilidad periódica de diferentes segmentos de mismo tono, o cambios bruscos en el tono.

4.4.2 El histograma y su procesado.

El histograma es un recurso muy utilizado en procesado de señal, en el que se almacenan el número de apariciones de un determinado valor de una magnitud, dando así una idea aproximada del espectro de apariciones en un determinado intervalo de valores. Si representamos este histograma en dos dimensiones (apariciones, valores), podemos encontrar valores significativos de nuestro análisis, como la aparición de picos, de valles, de regiones vacías, etc.

En nuestro caso, el histograma guarda las apariciones de las frecuencias fundamentales dentro del intervalo (fmin, fmax) que hayamos establecido. Del análisis de este histograma podremos sacar conclusiones como estas:

- regiones frecuenciales: las notas musicales serán aproximadamente iguales en diferentes apariciones en el tiempo, pero no tiene por qué ser siempre iguales fruto de los efectos de la variabilidad temporal del locutor. Por tanto, se formarán regiones de picos alrededor de la frecuencia central que más aparece.
- frecuencias fundamentales que más aparecen
- resolución de cada una de las zonas frecuenciales alrededor de una frecuencia.
- frecuencias que no se utilizan
- distancia en frecuencia entre los picos estimados.

Una vez que se cuenta con este histograma, se pasa a realizar la detección de las notas musicales, seleccionando para ello los picos más significativos. Esta tarea puede realizarse en nuestro programa de dos formas: automática o manual. Las características y diferencias en ambos casos son explicadas más profundamente un poco más tarde cuando relatemos el capítulo de decisión de picos.

Sin embargo, los resultados que el histograma nos devuelve, suelen ser en general difícil de interpretar por varios motivos.

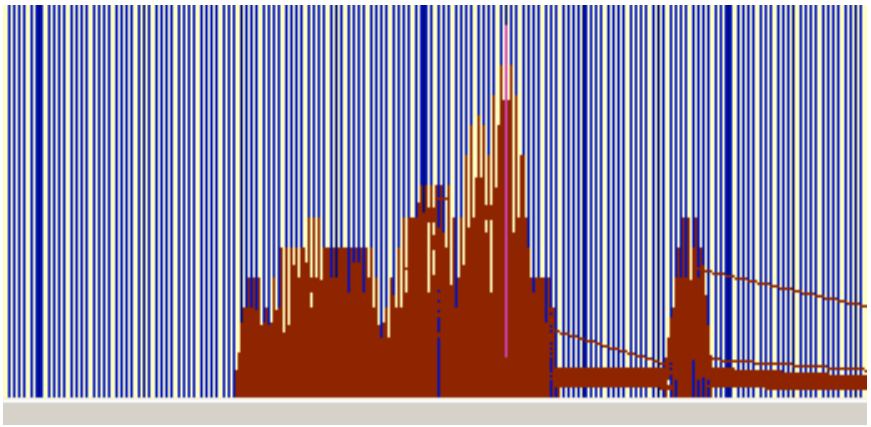
- Concentración de picos en regiones pequeñas de frecuencia, lo que obliga a representaciones muy ajustadas.

- Dispersión de valores en torno a uno o varios que parecen centrales, con dificultad para elegir un buen representante.

- Aparición de representantes de notas que no están separados por valles profundos, sino con un fondo de escala de valores difuminados por el histograma.

- Valores no representativos con concentraciones aisladas que se pueden confundir con otros que sí que conforman información.

Podemos un ejemplo de histograma en la siguiente figura.

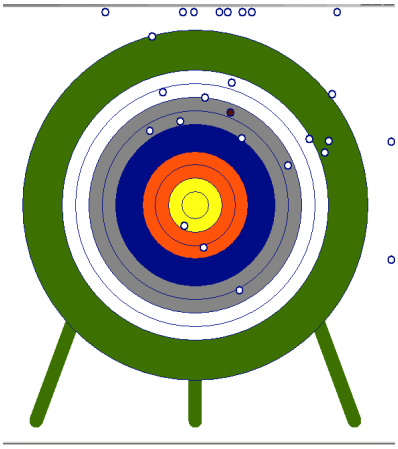


Por todo esto, se hacía necesario la creación de un modelo nuevo para poder entender los resultados que nos ofrece el histograma, y así poder sacar la información, que aunque ahí está, no sabemos discriminarla. De esta forma, introducimos unos modelos para la nota musical, para el pico y de la buena escala musical.

- **Definición de nota musical.**

Una nota es un concepto cuya simplicidad encubre una enorme ambigüedad. Suele considerársela un elemento de una escala musical, pero al medir su frecuencia, vemos que más que un valor de frecuencia es un conjunto de ellos, una nube estadística con una concentración en un punto si hay suerte. Esta suerte proviene bien de que el instrumento que la emite es de afinación fija, de que el ejecutante es muy preciso, o bien si fallan los dos casos anteriores, de que el fragmento analizado era muy largo con lo que gracias a la ley de los grandes números, se van creando distribuciones estadísticas de aspecto aproximado al normal con su media y desviación típica.

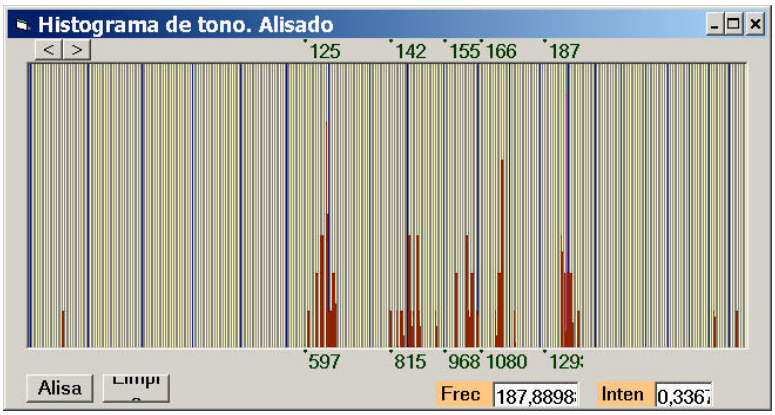
Si no ocurren esas deseables concentraciones, habría que generarla de forma artificial, intentando encontrar la nota a la que el intérprete quiso acercarse a través de los intentos no muy certeros para lograrlo. Como analogía podemos decir que se trataría de alguien que quiere meter un punzón en un agujero y hace varios intentos alrededor de lo que busca; o de un arquero que lanza flechas a un blanco al amarillo de la diana y va creando una nube de flechas alrededor.

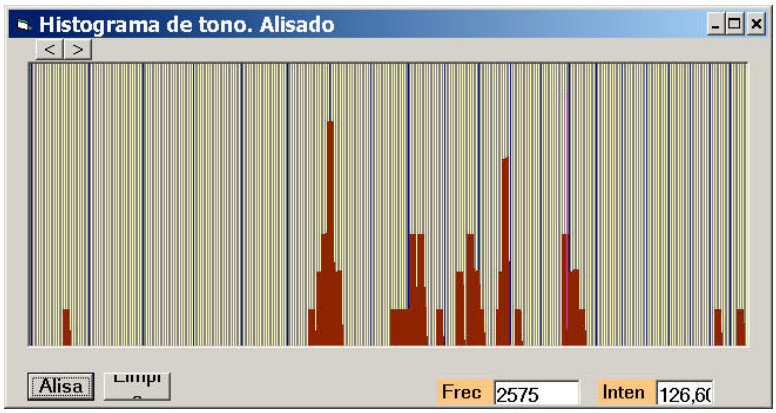


Por tanto, hay que encontrar la media de una distribución normal de desviación típica igual a la desviación típica práctica encontrada empíricamente.

Un método más fácil, es alisar por las buenas el histograma de frecuencias (es decir convolucionar el histograma con una función o ventana que suavice los picos secundarios) y si hay suerte, lograr una cima que adoptamos como frecuencia de la nota.

Si alisamos más, y esto ya es una elección y una responsabilidad, encontramos un gráfico más bonito. Pero ¿más verdadero?





Nótese que los picos que creemos encontrar en una región son filtrados que realizan nuestros ojos, o nuestra percepción visual, aplicando una ventana visual sobre el histograma y obteniendo así una región que toma como pico. Todo, pues, son operaciones con buen funcionamiento empírico pero de difícil justificación teórica.

Podemos refinar la situación del centro de una nube de tonos que es lo que llamaremos el tono de la nota, atribuyendo a esa nota la frecuencia que minimiza la varianza de todas las distancias de la nube de tonos a ella, es decir, se calcularía el centro de gravedad, lo que presupone elevar al cuadrado esas frecuencias.

Otra posibilidad sería hallar únicamente las distancias en valor absoluto, lo cual variará el resultado en función de la banda de frecuencia de la nota. Como solución simple, se puede tomar el mayor valor de la nube o el punto medio si hay varios máximos iguales.

Si tenemos varios máximos locales separados por valles en el ámbito cercano de una nota, hay ambigüedad y diremos que puede tratarse del resultado de pocos valores de estudio o de dos notas en función diferente, bien por modulación, o bien por ejecución inexacta (una nota que se acerca a la que va).

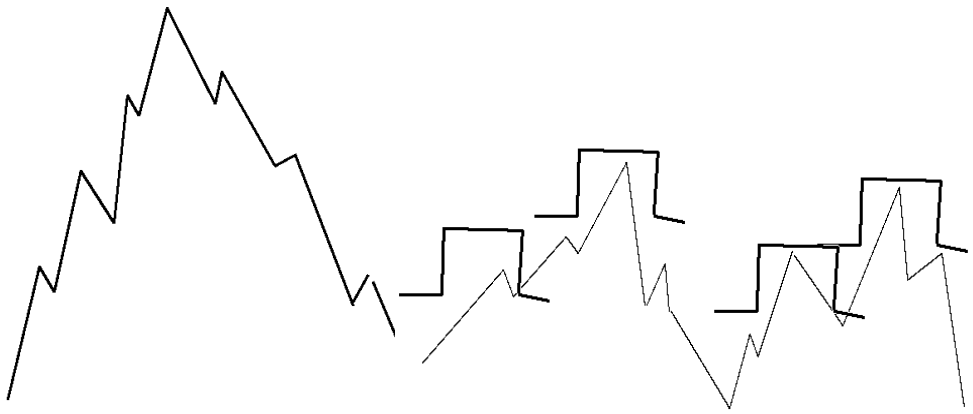
Los resultados serán además muy dependientes de la capacidad que tengamos de discernir picos con información de los que sólo añaden ambigüedad. Será por tanto necesario recurrir también a una teoría de evaluación de la presencia de picos, a la cual recurriremos en la próxima sección.

- **Teoría del pico.**

Dentro de la llamada inteligencia artificial, surge continuamente la necesidad de definir concretamente y a la vez algorítmicamente una multitud de conceptos que parecen muy claros en el lenguaje y la percepción corrientes. Esta claridad aparente, que tiene que ver con la compleción de datos sensibles por el aparato perceptivo descrito por la *Gestalt*, se desvanece cuando queremos que un ordenador sitúe, vea, oiga esas claridades.

Por ejemplo, el concepto de pico, pico en una cordillera, es de lo más elusivo pese a su aparente simplicidad. El problema reside en que siempre hay un tamaño implícito, previo, de esos picos, lo cual limita su número

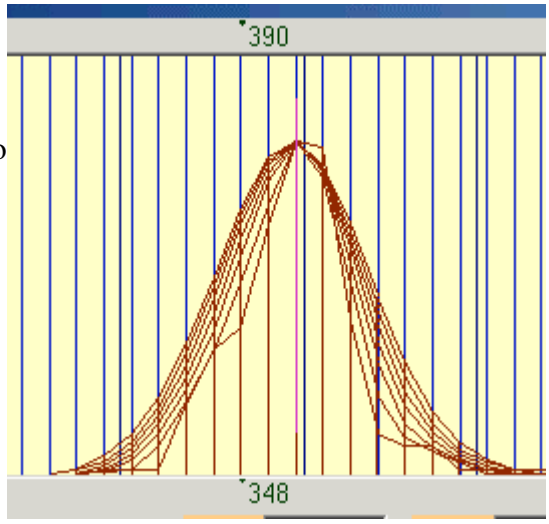
El concepto de pico en una sierra o conjunto de ellos, implica un alejamiento del contemplador, lo cual supone un alisamiento de su contorno primeramente. Además, hay pico sensible si y sólo si hay falda alrededor, una falda suficientemente larga.



Por lo tanto para encontrar los picos en una sierra o en una función, se debe disponer de una plantilla, que es como un sombrero o como un apagador de velas, es decir, un corto cilindro hueco con tapa superior que se coloca sobre el candidato a pico y si algún punto toca la tapa, entonces tenemos un pico. Ya se ve que la altura del cilindro y su diámetro determinan cuáles serán picos y cuáles no. En la figura adjunta hay, o un pico grande, o varios pequeños.

Perceptivamente el número de entes que se capturan de un vistazo o de dos, suele ser de uno a diez, lo cual determina el número de picos que esperamos ver a priori y por tanto, las dimensiones del sombrero de test.

Por ejemplo en la figura adjunta sólo hay tres picos para el sombrero elegido; pero un sombrero mucho más pequeño encontraría muchos más.



Toda esta disquisición surgió cuando intentamos encontrar en un histograma de frecuencias los puntos preponderantes, como las notas de una escala, donde la necesaria separación entre sus notas del sombrero depende del tipo de música que consideramos. Un sombrero de anchura de un semitono para una escala occidental o de un cuarto de tono, o incluso menos para una turca.

La presencia de muchos picos pequeños puede atenuarse alisando o mediante una convolución con ventana en el dominio del pico, de la que ya hemos hablado. Por ejemplo, sustituyendo un valor por la suma de sus contiguos. Este procedimiento conduce a acentuar el pico grande, como muestra el proceso de alisado repetido en un pico de tono, en la figura adjunta: el contorno quebrado se alisa convergiendo hacia una campana de Gauss.

- **Teoría de la buena escala.**

A efectos de justificar la estimación automática de una escala, pero sobre todo a efectos de comprender qué es una escala musical en su uso y percepción, es preciso elaborar una corta teoría de la buena escala.

Decimos buena escala, porque no cualquier conjunto de notas o frecuencias al azar son musicalmente útiles. Nada impide su definición y uso por una máquina, pero los oyentes humanos ya, no encontrarían interés ninguno en esos tonos incorrelados, a la vez que ningún humano podrá cantarlos o interpretarlos con instrumentos de afinación variable. Definamos por tanto, a partir de la integración de muchas escalas en muchas culturas, el concepto de una buena escala.

1) Una escala musical es un conjunto de notas (producidas por frecuencias) relacionadas entre sí por dos tipos de distancia o intervalo

El primer tipo de intervalo es el habitualmente conocido como tal, representable mediante un par de frecuencias. Cada una de estas distancias está sometida a unas reglas empíricas cuando un conjunto de tonos han de formar una buena escala musical. Estas reglas son aproximadamente:

- En una octava hay un número pequeño de notas o tonos, entre cinco y siete.

- Los intervalos entre cada dos notas consecutivas (aquellas que no tienen una frecuencia intermedia entre ellas) no son ni muy grandes ni muy pequeños, valiendo desde un semitono a tres, usualmente.

- Hay pocos tipos de intervalos contiguos, preferiblemente sólo dos, un caso típico es la escala mayor occidental, con sólo tonos y semitonos.

El segundo tipo de intervalo es el de la sonancia mutua de ambas notas. En este tipo de intervalo también admite una gradación de mayor a menor sonancia, relacionada con las teorías conocidas desde antiguo de que cuanto más simple es la relación, más cercano y consonante resulta.

Las características fundamentales que afectan al segundo tipo, el de la sonancia, son:

- Todos los tonos de la escala o grado forman entre sí intervalos tan consonantes como sea posible. Esto es muy posible para intervalos grandes, pero los pequeños necesariamente han de ser los menos, porque sus razones son menos sencillas. Se produce así una estructura arborescente donde cada gran intervalo comprende otros dentro de él, tendiendo siempre a la consonancia. Así la escala mayor natural podemos considerarla como una octava dividida en quinta inferior y cuarta superior, la quinta inferior dividida en tercera menor superior y una segunda inferior, la cuarta superior dividida bien en segunda y tercera menor, o bien en tercera y segunda. En el tercera menor inferior podemos considerarla dividida en dos tonos mayor el primero y menor el segundo, y así sucesivamente.

Esta estructura arborescente genera espacios melódicos hasta cierto punto estancos en la que ocurren cosas dentro de uno hasta que se pasa a otro. Todo esto es más bien válido en la música modal tradicional, porque la experimentación occidental ha intentado destruir este edificio constantemente, consiguiéndolo a medias.

Así que al estimar los intervalos de una escala conocida o desconocida hay que esperar siempre unos intervalos entre las notas que no sean menores que un semitono (semitono temperado, doceava parte de una octava, cuya razón de frecuencias es la raíz doceava de dos que vale 1.059 aproximadamente) y que tengan entre sí consonancias.

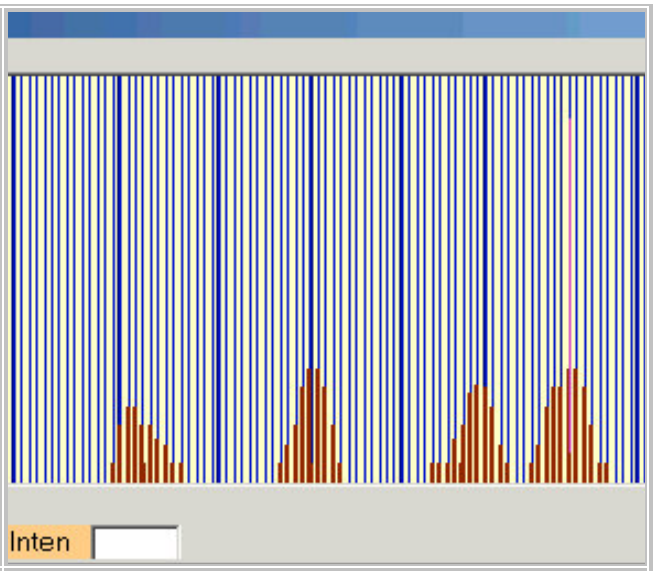
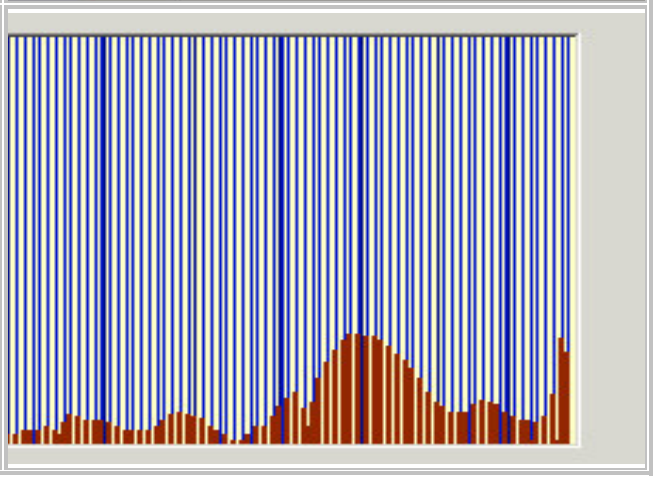
Si se encontraran unas notas cuyo intervalo es bastante menor que un semitono, hay que pensar que son dos casos de una nota o grado en diferente escala, bien por una modulación, bien por la tendencia bastante general de subir la nota ascendente y de bajar la nota descendente. O bien que el intérprete ha desafinado a lo largo de la ejecución de la pieza.

Al hablar de semitono temperado no queremos decir que haya que esperar semitonos temperados como intervalos de escala, estos pueden ser variables. Usamos este semitono como forma de medir, como unidad de tamaño de intervalo, no como intervalo real, ya que ni siquiera es consonante porque su razón es demasiado complicada

.

Estas consideraciones pueden ser útiles a la hora de enfrentarse con un histograma de frecuencias y tener que decidir a qué escala corresponde.

Algunos ejemplos de lo que podemos esperar se pueden ver en las siguientes figuras (y de cuyo significado hablaremos enseguida). Primero, tocamos en un kamanché (instrumento de cuerda frotada) un tetracordio tipo Rast. El histograma muestra que no está mal entonado respecto a la teoría. Después, una tonada interpretada por *Mairena*, melismática e imprecisa tonalmente. Sin embargo, surge la escala gitana.

INTERVALOS 192 379 382 499 186 189 307 2 120 117 KAMANCHÉ	
INTERVALOS 116 336 499 -146 220 382 -263 162 -483 -646 LA ESCALA GITANA	

- **Suavizado del histograma.**

Uno de los principales problemas del histograma es que a menudo, sacar información de sus picos suele ser más un fenómeno perceptivo del que elige los picos, que de la información aparte que se pueda incorporar.

En un histograma con abundancia de picos abruptos muy juntos, será difícil elegir un representante con total fiabilidad. Por eso, se recurren a técnicas de suavizado del histograma.

El problema de suavizar un histograma se ha planteado en mucha literatura desde antiguo. Para hacernos una idea, tan sólo hay que echar un vistazo al capítulo que le dedica el *Numerical Recipes*: podemos comparar métodos actuales y los clásicos de los años sesenta. En general, se han utilizado métodos lineales en los casos más fáciles y métodos no lineales para casos más complicados. Entre los lineales están por ejemplo la suma de gaussianas, y entre los no lineales está por ejemplo el famoso algoritmo *AMOEBA*.

En nuestro caso, y para no complicar aún más el proyecto con evaluaciones diferentes de suavizados, hemos elegido un método lineal, que va tomando la media de los valores del histograma en un entorno, utilizando diferentes pesos para cada punto del entorno. El resultado suele ser perceptivamente mucho mejor, aunque hemos de preguntarnos por la exactitud de los puntos que se extraen. Algunas consideraciones acerca de este tipo de suavizado son:

- El número de iteraciones suavizado.

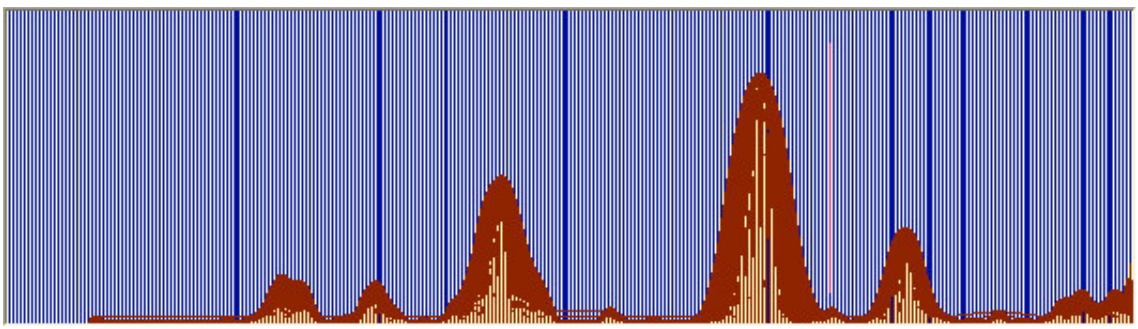
  Con el número de veces que realicemos el proceso de suavizado, podemos controlar la forma de los lóbulos que salen en el histograma. Se puede decir que cuantas más iteraciones, más suave será el lóbulo, y más alto, pero por contra podemos perder demasiada información. Por tanto, el número de iteraciones máximo debe ser tal que tras una nueva iteración los resultados no cambien en un porcentaje del resultado de la iteración anterior. El porcentaje debe ser calculado experimentalmente generalmente.

- El número de puntos que se toman para la media.

- El cálculo de los pesos aplicados a cada valor que interviene en la media.

En general, esta forma de suavizar, si acaso un poco grosera, sirve para darnos una idea de lo que está pasando, dejando para futuras ampliaciones la inclusión de un método un poco más arriesgado.

En la siguiente figura vemos la típica forma en la que queda el histograma tras su suavizado. Como se puede ver, se aprecian claramente los máximos, aunque la desviación típica de las distribuciones es un poco grande.



4.4.3 Reconocimiento en modo manual y en modo automático.

Como dijimos anteriormente, la tarea de encontrar el modo a partir de los picos del histograma, se puede hacer con dos filosofías diferentes de participación del usuario: manual o automática.

1) En la versión manual, el usuario va a encontrar visualmente los picos y a introducir su valor al programa para que realice el análisis interválico.

2) En la versión automática, un algoritmo es el encargado de encontrar los máximos y de guardarlos para su posterior análisis.

Como se puede suponer, la búsqueda de picos en un histograma no es una tarea que tenga una única solución, ni una solución válida en todos los casos. Nos adentramos por tanto en campos cercanos a la inteligencia artificial, puesto que debe ser la máquina quien haga una tarea inteligente, discriminar de forma independiente los picos que valen y los que no. En nuestro caso, la solución se basa en la búsqueda iterativa de bordes en el histograma para encontrar regiones, para después restringir la búsqueda de máximos a regiones más pequeñas donde hay más probabilidad de acertar con el pico acertado. Nos adentraremos más particularmente en el proceso automático en las próximas secciones.

- Reconocimiento manual.

Como explicábamos anteriormente, en este momento la misión es encontrar los máximos significativos en el histograma, los cuales no tienen por qué tener amplitudes comparables: si la amplitud un pico es la cuarta parte de la de otro, y sin embargo está rodeado por zonas más bajas delimitando una zona frecuencial, querrá decir que es un buen candidato a ser una nota de la escala del modo musical, y por eso no podemos desechar cualquier pico por su amplitud.

Por tanto, lo mejor es que el usuario experimentado, y con cierta intuición de los resultados de lo que pretende encontrar, recorra visualmente el histograma y pueda decidir con qué picos de frecuencia se queda.

Esta tarea ha sido pensada en el programa de la forma siguiente:

- El usuario termina su análisis en tiempo real de la frecuencia fundamental y pulsa el botón de *Histograma*, con lo que aparece una nueva ventana, con varias ventanas gráficas, numéricas y varios botones.

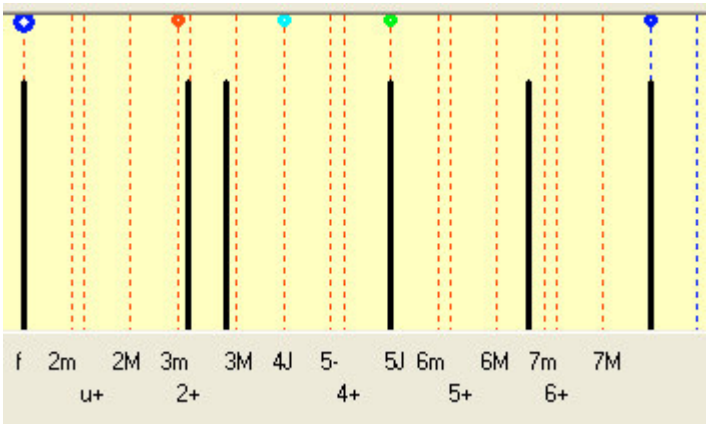
- En una de las ventanas gráficas se muestra el histograma. En la misma ventana se le ofrece la posibilidad de alisar el histograma, y si así lo desea, va obteniendo las variaciones en la forma de los lóbulos y los picos.

- Mediante un click en cualquier posición de la ventana del histograma se obtiene la posición de frecuencia, y se muestra al usuario en una pequeña ventana. Si el usuario lo desea, almacena los valores y se pasan para buscar las relaciones frecuenciales.

Un punto fundamental para continuar el análisis es la decisión del pico fundamental. Como explicábamos en las primeras secciones teóricas de esta memoria, cada escala musical tiene una nota fundamental, la cual aparece a menudo, y sobre la que se establecen las relaciones interválicas. Por tanto, una vez que tenemos los picos, es necesario saber cual es la nota o pico fundamental para poder relacionar las distancias con esa nota. En general, la condición a poner es que la fundamental suele ser la nota que más apariciones tiene, pero no tiene por qué convertirse en norma. También aparecen muy a menudo la dominante y subdominante así como el séptimo grado.

Una vez que tenemos la fundamental y los picos, un submódulo se encarga de dividir mutuamente los valores de esos picos por la fundamental, encontrando unas relaciones absolutas. Esas relaciones son los intervalos entre las notas. Una forma más adecuada de ver esos intervalos es en unidades *cents* ; puesto que una octava son 1200 cents, cada semitono son 100 *cents*, y eso nos da una forma de comparar distancias mucho más exacta. Las relaciones absolutas y en *cents* se muestran en una ventana numérica, donde podemos ver exactamente sus valores.

Como la representación numérica es poco intuitiva a la hora de una comparación rápida, la solución adoptada además fue la representación visual en escala interválica. Sobre una ventana rectangular, dividimos su ancho entre las notas del sistema pitagórico, y después superponemos las notas que el usuario ha introducido por teclado, dándonos idea de la naturaleza de los intervalos. Esta ventana se ve a continuación.



Una vez que hemos dado por buenos los resultados, podemos construir el modo a partir de las notas, almacenándolo y comparándolo con otros que hayamos introducido anteriormente, realizando así la tarea musicológica de forma completa.

- Reconocimiento automático.

Una vez vistos los problemas de la decisión en la versión manual, afrontamos la tarea de crear un algoritmo de decisión automática.

Este algoritmo debe hacer en esencia las mismas tareas que hacía el usuario anteriormente, pero sin ayuda, sólo con la información residente en el histograma. Por

tanto, se independiza al usuario del análisis, esperando que los resultados sean similares a los que se consiguen con la versión manual.

En primer lugar, se debe afrontar la tarea de encontrar los máximos en le histograma. En general, el histograma fruto de un análisis en el tiempo de la frecuencia fundamental sobre segmentos de voz cantada, suele ser un conjunto de picos de variadas amplitudes, que se agrupan en regiones de amplitud similar, y que se diferencian de otras regiones por valles de apariciones. Estos valles, pueden tener amplitudes similares a las amplitudes de algunas de las regiones, y no podemos afrontar la búsqueda de picos desde técnicas básicas, sino que debemos dotar al algoritmo de una inteligencia adicional.

Los pasos seguidos para la búsqueda de picos es la siguiente:

- Sobre el histograma, ponemos a cero los valores que no superen un tanto por ciento el valor del máximo del histograma. Con esta técnica, eliminamos valores que no tienen importancia, por ser de transición entre regiones, y a la vez acentuamos los valles entre regiones. Se han utilizado cribas del 10% y del 15%, las cuales se mostraron eficaces para la mayoría.

- Sobre el histograma cribado, se realiza una búsqueda de bordes. Definimos borde como el inicio de una región en el histograma, la cual se caracteriza por la acumulación de valores altos a continuación del borde, cuando anteriormente no encontrábamos esa acumulación. La búsqueda de bordes se realiza de la frecuencia máxima a la mínima, y de la mínima a la máxima, encontrando los bordes que delimitan las regiones por encima y por debajo. Si la criba anterior fue adecuada, los bordes suelen acertar con las regiones. Si no fue adecuada, se eliminan algunos bordes intermedios, los cuales deberían delimitar nuevas regiones. Este problema es tenido en cuenta posteriormente, y solucionado.

- Con la información de bordes, obtenemos las zonas conexas del histograma. Dentro de esas regiones, vamos a realizar un nuevo análisis iterativo de criba y de bordes, cada vez más exigente, hasta que no encontremos nuevos resultados. De esta forma, podemos encontrar nuevas regiones inmersas en zonas conexas anteriormente clasificadas.

- En cada una de las zonas conexas que hemos encontrado, quizá de unos pocos valores de frecuencia de ancho, buscamos el máximo, que almacenamos como pico candidato a nota de la escala.

La utilización de este algoritmo se mostró adecuada en muchos ejemplos musicales, sobre todo si no eran muy extensos temporalmente, hasta tres o cuatro minutos. Piezas musicales muy extensas, empiezan a mostrar histogramas muy difíciles de comprender, incluso en la versión manual.

Una vez encontrados estos candidatos, mandamos de forma automáticamente sus valores en frecuencia al análisis interválico, realizando el mismo proceso que en el caso manual. De la misma forma, se muestran visualmente los intervalos.

En este caso, tenemos el problema de la detección de la nota fundamental de la escala. Evaluaciones prácticas sobre diferentes ejemplos musicales, muestran que nada se

puede decir acerca de la predicción de la amplitud del pico correspondiente a esta nota, por lo que la decisión de elegir el pico de mayor amplitud se mostró errónea en muchos casos. La decisión tomada fue una técnica mixta manual-automática, donde la búsqueda de picos es automática, pero la introducción de la nota fundamental es manual.

4.5 Evaluación

La evaluación de la aplicación consistió en la ejecución controlada de los diferentes módulos de los que consta con unos ficheros de sonido elegidos especialmente por su representatividad desde el punto de vista del objetivo de la aplicación.

4.5.1 Documentos de prueba.

Sobre el conjunto de posibilidades que podríamos haber escogido, seleccionamos aquéllas que se acercaran más a la música objetivo con la que la aplicación tuviera que pelear.

En fases anteriores de diseño, este objetivo fue delimitado a música modal cantada de todo tipo, interpretada sobre cualquier escala modal. Por tanto, se nos abría una gran familia de músicas tradicionales, ¿cómo elegir la adecuada? Decidimos por un lado incorporar música viva diferente, y por otro lado generarnos un conjunto de materiales de evaluación artificiales, que fueran totalmente objetivos. De esta manera, con la música viva realizaríamos una evaluación real y con la generada podríamos ajustar ciertos parámetros.

En especial, se evaluó sobre los siguientes documentos sonoros:

- Un *generador de señales sinusoidales* basado en Pure Data. Se utiliza para comprobar los resultados que arrojan los algoritmos en fase de desarrollo y prueba. Al no introducir armónicos a la señal, si la decisión es correcta, es una buena forma de apreciar la exactitud del algoritmo en ausencia de fallos debidos a otras fuentes de error.

- *Naat Taksim*. Consiste en un canto de alabanza turco, interpretado por un maestro de canto de Konya. En él podemos estudiar líneas melódicas complicadas, con variaciones frecuenciales muy pequeñas, y muy ricas en timbres. También es útil para estudiar las micromodulaciones de tono típicas de la música islámica. Por tanto, este documento se utilizará para comprobar la robustez de la aplicación, puesto que incluye una gran complejidad en todos los campos.

- *Algunos cantes flamencos.* Dada la proximidad geográfica y cultural de esta música modal, parecida adecuado acercarse a ella y tomarla como parte de la evaluación.

- Además, se incluyó una evaluación muy buena de los resultados gracias a la amabilidad del Dr. Francisco Javier Sánchez, el cual no dudó en prestarnos su interesantísimo programa *Mapatone*, en el cual se pueden cargar escalas mediante parámetros e interpretar melodías sobre un piano modificado. De esta forma, pudimos comprobar minuciosamente la salida de la aplicación, disfrutando además de la música modal que generaba automáticamente.

4.5.2 El proceso de evaluación.

El proceso de evaluación se dividió de la siguiente manera:

1) Evaluación de la exactitud de la estima de la frecuencia fundamental mediante el algoritmo de autodisimilitud mediante tonos puros.

   La prueba se realizó conectando acústicamente al algoritmo a un generador sinusoidal cargado antes en el ordenador. Se evaluó la estabilidad en régimen permanente y la exactitud en régimen transitorio. De esta forma podemos observar objetivamente la precisión del algoritmo en su estima.

2) Evaluación de la exactitud en la decisión automática de las notas sobre el histograma mediante música modal generada con *Mapatone*. Mediante este camino, podemos evaluar la calidad en la decisión automática de forma objetiva.

3) Evaluación de la exactitud en la estima de la frecuencia fundamental y en la decisión sobre el histograma en música viva. Se combinan ambos problemas y el grado se subjetividad del usuario que lo prueba, puesto que no se tiene la solución objetiva para poder comparar.

En los siguientes apartados iremos describiendo cómo fue el proceso de evaluación y sus resultados, para dar así muestra de la calidad de la aplicación.

4.5.2.1 Evaluación mediante tonos puros.

La evaluación se realizó probando la estabilidad en la estima de la frecuencia fundamental en régimen permanente y en régimen dinámico.

Por un lado, en régimen permanente, la prueba consiste en elegir una frecuencia en el generador en el margen de frecuencias que hayamos elegido, dejar que el algoritmo se estabilice, y ver con qué precisión se estima en la aplicación.

Este proceso se repite incrementando la frecuencia del generador en pequeños saltos de frecuencia, registrando la variación entre la frecuencia introducida y la de la aplicación.

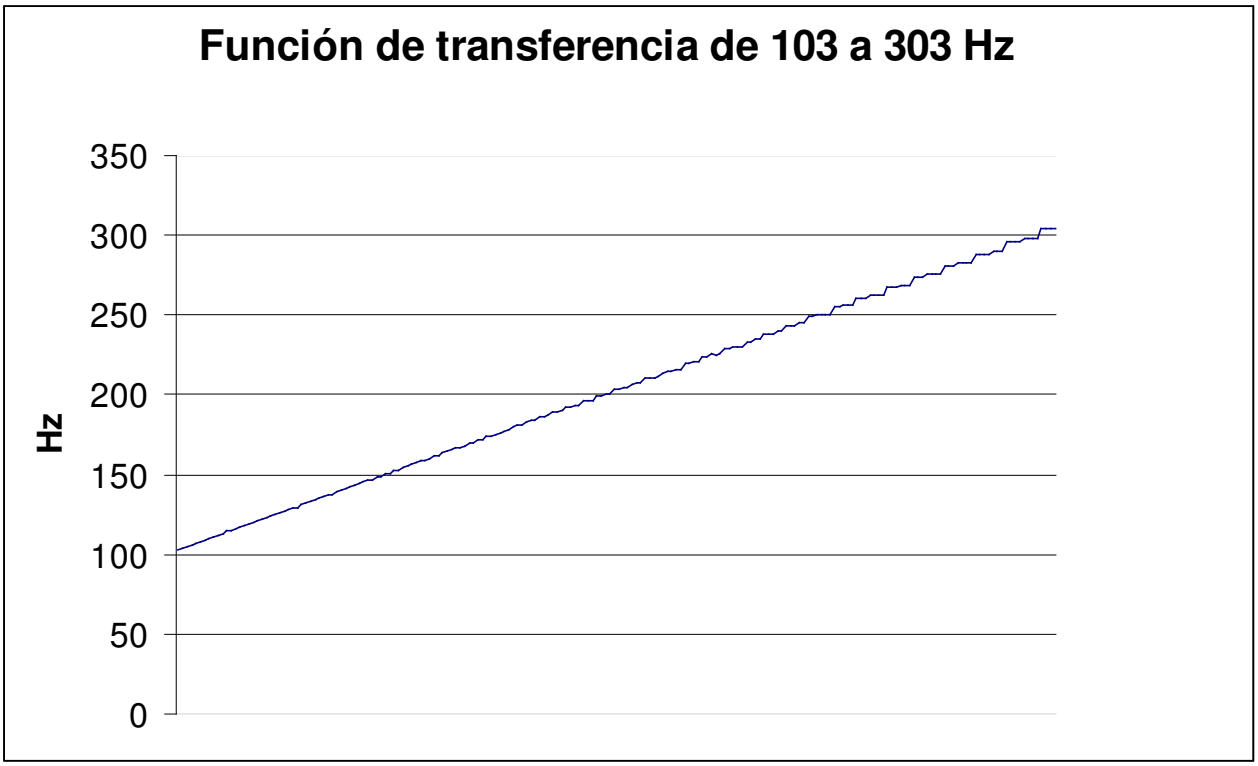
En nuestro caso, hemos decidido que el incremento de frecuencia sea 1 Hz, de forma que podamos ir viendo la evolución de la estimación se forma continua. Esta evolución la podemos observar en la gráfica siguiente, en la que se ha calculado la función de transferencia en el margen más utilizado, de 103 a 303 Hz, y con una frecuencia de muestreo de 44100 Hz.

Como se puede ver, la función de transferencia es a nivel global, muy lineal, lo que quiere decir que la respuesta del algoritmo es buena. Sobretodo en las frecuencias más bajas, la precisión es muy alta, con errores que en ningún caso superan el 0.3 % en frecuencia, lo cual en este margen supone 0,4 Hz aproximadamente.

Por otro lado, vemos cómo a partir de 200 Hz la estima va siendo menos lineal en detalle. Esto es debido a que un conjunto de frecuencias es estimado como una sola, que las representa a partir de ese momento.

Sin embargo, ese representante sólo agrupa a un conjunto pequeño de 4 o 5 frecuencias como máximo, es decir, cometemos un error de 5 Hz como máximo, que corresponde a un error de un 2 % en frecuencia.

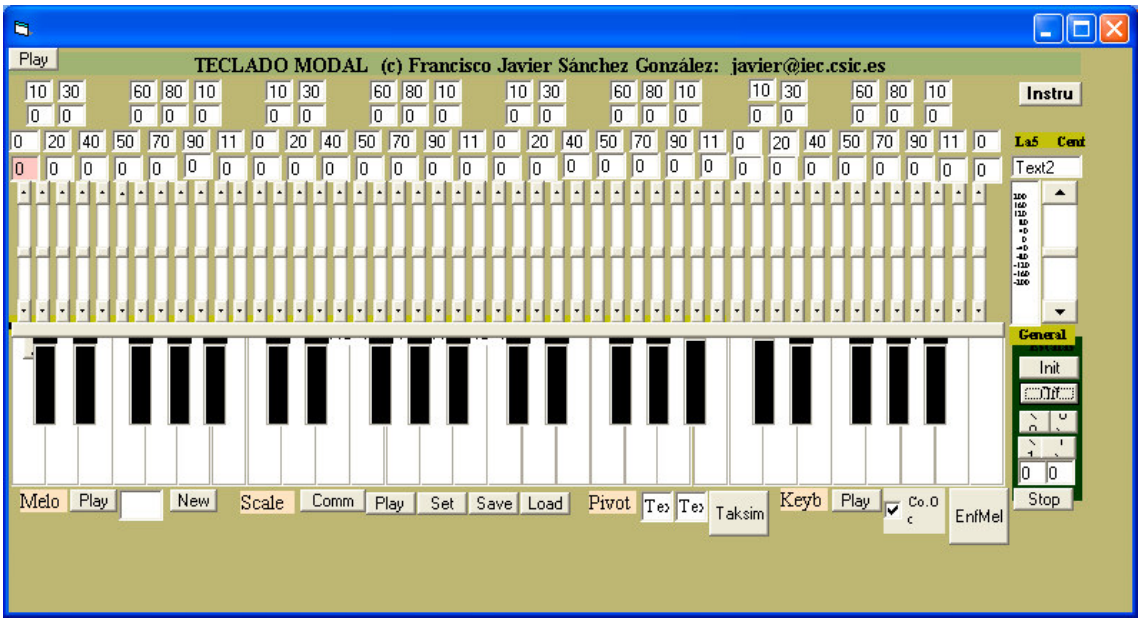
.



Por tanto, como conclusión, podemos decir que en las frecuencias más bajas la precisión está dominada por los valores posibles que nos ofrece el muestreo. A frecuencias más altas, se añade al anterior de precisión los errores de procesado, quizá fruto de la capacidad limitada de cálculo del ordenador o de la forma de gestionar los buffers de muestreo. Sin embargo, podemos decir que un error de precisión de un 2 % como máximo es un margen aceptable, incluso bueno comparándolo con los resultados de otros tipos de algoritmos.

4.5.2.2 Evaluación mediante música generada por *Mapatone*.

Mapatone es una aplicación programada por el Dr. Francisco Javier Sánchez González en el seno del CSIC en Madrid. Se trata de una aplicación que genera música de forma automática a partir de un conjunto grande de parámetros. El apartado más importante de este programa para nosotros es el *teclado modal*, que sirve para interpretar música en un teclado afinado con la escala que se desee. Una captura de este programa se puede ver a continuación.



Por tanto, podemos generar música con las escalas de los modos que queramos, con mucha precisión, y que se escuche a partir de notas MIDI. Después analizaremos con nuestra aplicación en tiempo real el audio que nos ofrece, calcularemos la decisión sobre el histograma y compararemos los resultados con los originales de *Mapatone*.

En particular, generamos varias escalas, unas conocidas ya anteriormente en esta memoria, como la *Natural*, la *Pitagórica* y la *Temperada*, u otras menos conocidas, como la *Rast*, la *Huzzam* y la *Susinak*, cada una con sus características.

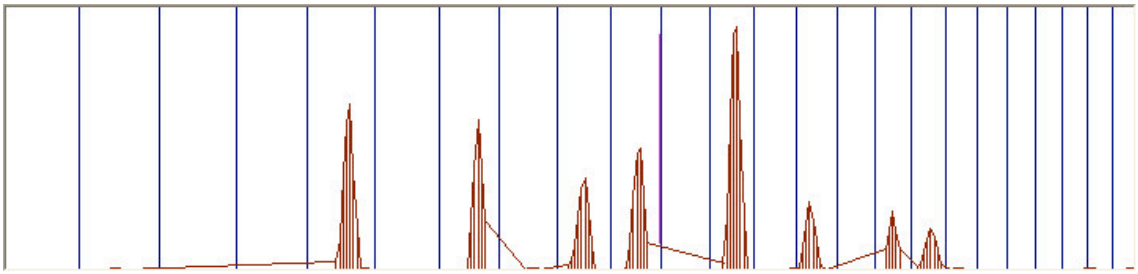
El proceso de evaluación será dejar sonar unos fragmentos generados con estas escalas y que nuestra aplicación las analice. Calcularemos los histogramas, y aplicaremos los algoritmos de decisión automática y manual, y estos resultados serán los que compararemos con los originales.

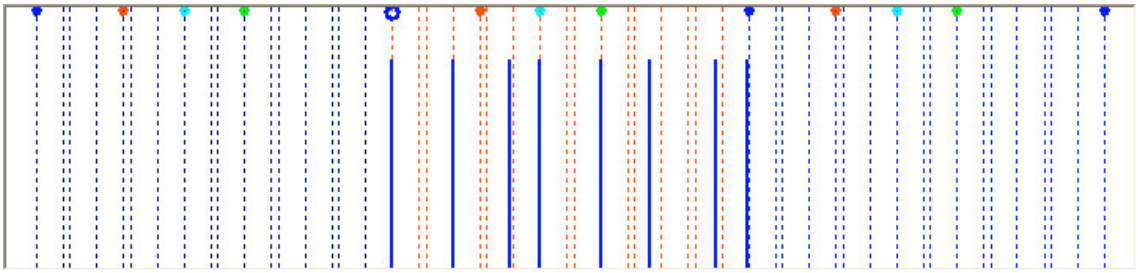
Para poder organizar la descripción de esta evaluación, primero describiremos la calidad en la determinación de los histogramas, y después analizaremos los resultados de las decisiones tomadas en la versión automática y en la manual.

4.5.2.2.1 Precisión en la generación de histogramas

Como hemos dicho, los histogramas se generan a partir del análisis de fragmentos de música en formato MIDI. El problema de este formato es que la música que se crea es demasiado perfecta a efectos prácticos: no hay desviación de frecuencia en la interpretación de las notas, y el timbre es siempre el mismo. Esta característica hace que el análisis sea mucho más fácil y obtendremos mejores resultados, pero perdemos la capacidad de calcular la robustez de la aplicación a música interpretada realmente. Este problema lo solucionaremos en siguientes apartados, cuando describamos la evaluación sobre música viva.

Para empezar, vamos a estudiar la precisión con la que se genera el histograma en el caso de la escala *Natural*. En las siguientes figuras, primero vemos el histograma, habiendo sido un poco suavizado previamente. Más abajo vemos la tabla de intervalos, habiendo hecho una selección manual, en formato de distancias en *cents* entre notas. Por último, observamos un esquema visual comparativo con la escala pitagórica acerca de la posición de las notas en la escala.





La escala natural está afinada en *cents* como se puede observar en el recuadro amarillo de la segunda figura, donde las filas son las notas de la escala, la primera columna es la distancia entre cada nota y la primera, y la segunda columna es la distancia de una nota a su anterior. Comparando con la obtenida con nuestra aplicación, vemos que:

- Hay 2 *cents* de diferencia en el Re (segunda nota).
- Hay 2 *cents* de diferencia en el Mi.

- Hay 2 *cents* de diferencia en el Fa.
- Hay 1 *cent* de diferencia en el Sol.
- Hay 2 *cents* de diferencia en el La.
- Hay 2 *cents* de diferencia en el Si.

Por tanto, el error máximo que estamos cometiendo es de 2 *cents*, cantidad ínfima. Este error, como MIDI no comete errores, es fruto de la selección manual con el ratón de los picos y del suavizado de los picos. Sin embargo, este error se puede minimizar si evitamos el uso de suavizado para música MIDI, o apuntando mejor en la selección manual de los picos. Se puede llegar incluso a no cometer errores de este tipo en el uso de la aplicación sin tener por ellos que haber atesorado demasiada experiencia.

4.5.2.2.2 Precisión en la decisión manual.

Como hemos visto en el apartado anterior, la selección manual tiene el problema de que puede ser algo aproximada si no se apunta de forma precisa en el histograma la posición del pico. De esta forma, aunque la precisión en la generación del histograma haya sido perfecta, estaremos añadiendo cierto porcentaje de error por una selección no muy ajustada.

Este error tiende a disminuir y a incluso desaparecer con la experiencia en el uso de la aplicación, pues después de ciertas iteraciones, se aprende a buscar mejor el punto preciso del pico.

En próximas versiones del software, está prevista la incorporación de un zoom, para ayudar al usuario en el proceso de selección de picos en el eje de frecuencias.

4.5.2.2.2 Precisión en la decisión automática.

La decisión automática consistirá en la búsqueda automática de picos en el histograma, y su visionado en el diagrama de barras. El proceso es totalmente automático. Como hemos visto anteriormente, el programa:

- hace un análisis del histograma,
- se criban los valores con un umbral, de forma que acentuamos los valles entre picos,
- se buscan bordes en el histograma cribado,
- se crean regiones agrupando los bordes,
- se hace una búsqueda de picos por regiones, pudiendo realizar un proceso concurrente al descrito.

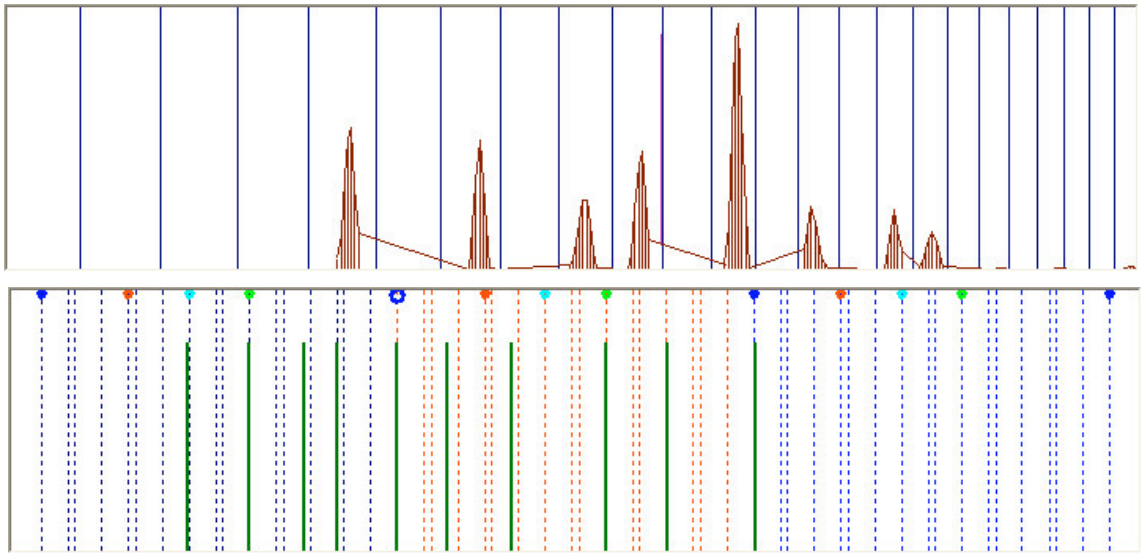
Este análisis por regiones, tiene el problema de que si tenemos picos muy poco superados, es probable que se asocie una misma región. Al utilizar un umbral de criba, también es posible que suprimamos la presencia de algunos picos que sí que tienen información.

Además, tenemos el problema de que para ordenar las notas en el diagrama de barras, necesitamos etiquetar a un pico como pico fundamental, para hallar dividiendo por él los intervalos. Hemos decidido seleccionar el pico con mayor amplitud del histograma, pero en general no tiene por qué cumplirse esta condición. En este caso, los intervalos pueden detectarse correctamente, pero saldrán desordenados en el diagrama de barras.

En todo caso, tendremos que contar con la colaboración del usuario, el cual tendrá que interpretar los resultados y buscar la información que necesite.

La evaluación se realizó sobre las escalas *Natural*, *Pitagórica*, *Susinak* y *Rast*. Los resultados que obtuvimos son las siguientes:

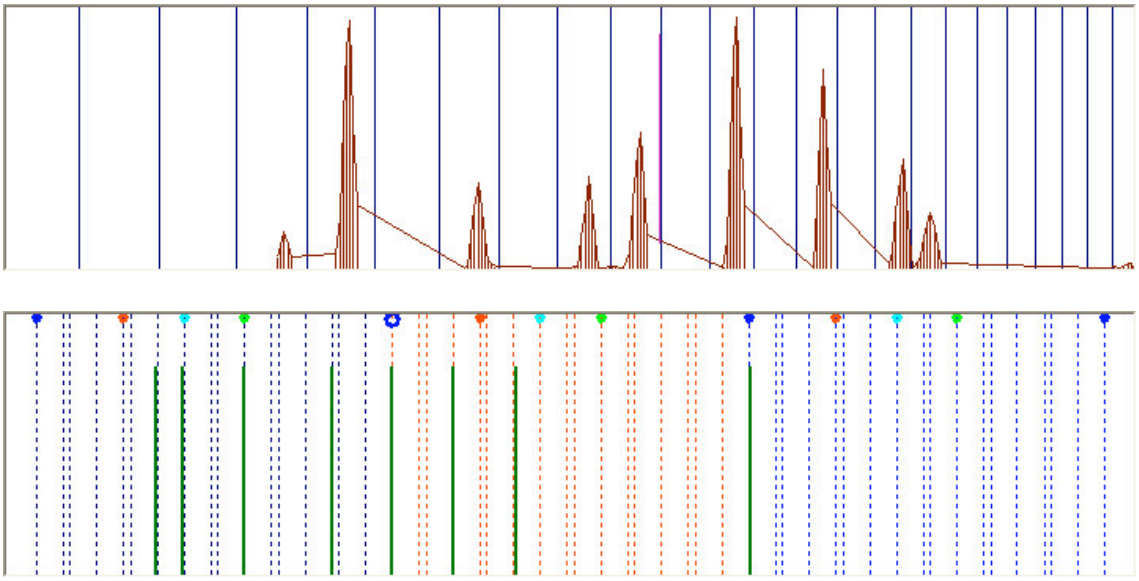
Natural:



Se observa cómo es la quinta nota la predominante en este segmento de música. Como hemos generado nosotros, sabemos que es la quinta nota de la escala, la dominante. Por tanto, se tomará esta nota como la fundamental, aunque sea incorrecto. Vemos cómo hay dos tipos de tonos y de semitonos, como ocurre en la escala *Natural*. También vemos cómo se ha omitido la octava nota porque es confundida por la séptima, por estar muy cerca y caer en su región.

Por tanto, la decisión automática en este caso es buena, aunque hay que contar con la participación del usuario para interpretar los resultados.

Pitagórica:

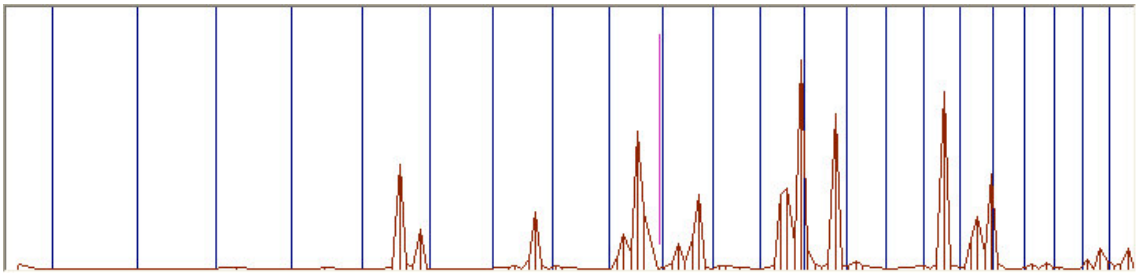


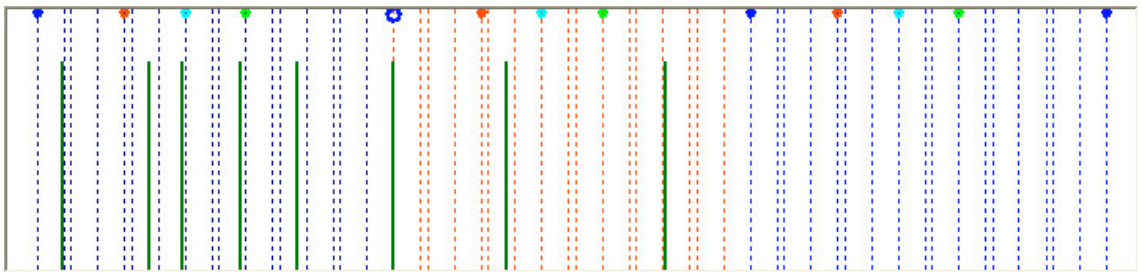
La escala pitagórica es la misma con la que hemos generado la escala del fondo del diagrama de barras. Vemos que la generación del histograma y la selección de picos ha sido bastante buena, pues coincide exactamente las barras fruto de la decisión con la escala del fondo.

En este caso se ha vuelto a decidir la fundamental como la quinta de la escala, y se ha omitido la tercera de la escala, dejando ese hueco en el diagrama de barras entre los círculos verde y azul en el tramo en azul (izquierda).

Por tanto, hemos perdido algo de precisión en la decisión automática, aunque todavía tenemos un acierto importante. Complementando la información con la selección manual, terminaríamos con encontrar toda la información de la escala.

Susinak:





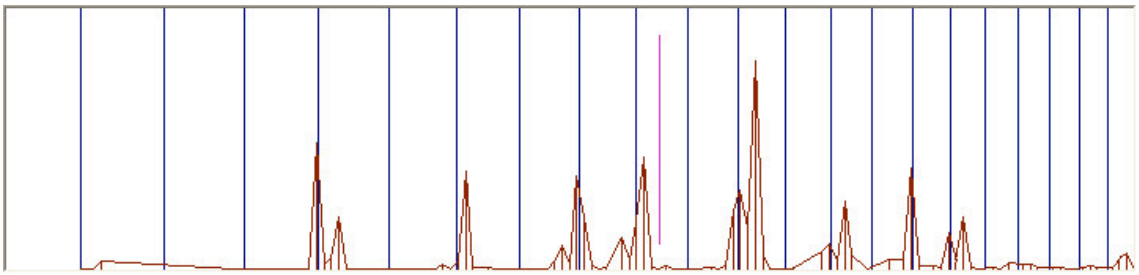
En la interpretación de la escala *Susinak*, se utilizó un timbre específico, con bastante vibrato. De esta forma, cada nota es además modulada lentamente en frecuencia (como máximo 10 Hz).

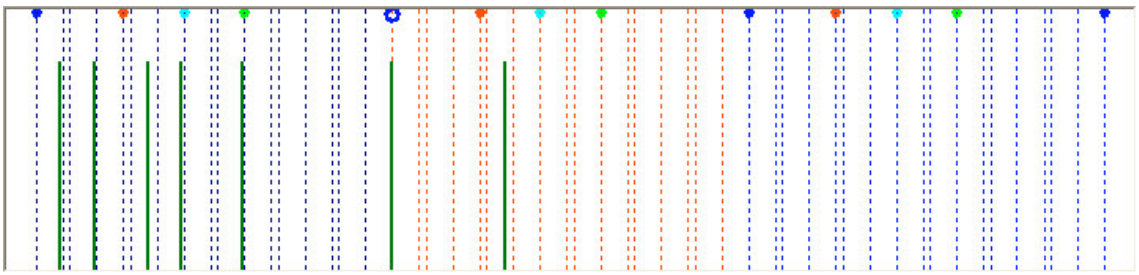
El efecto del vibrato se puede observar en el histograma, donde además de la nota fundamental podemos ver otros picos menores adyacentes, fruto de esa variación de frecuencia. Por tanto, hemos descubierto el efecto del vibrato en el histograma, lo cual nos va a ayudar bastante para sacar información en los casos de música viva.

Analizando las barras vemos cómo de nuevo es la quinta la que tiene mayor amplitud, por lo que la decisión está desplazada. Se omiten en este caso la cuarta nota y la sexta, por estar muy cerca de la tercera y de la quinta, respectivamente.

Por tanto, tenemos todavía algunos errores en la decisión automática, y necesitaríamos complementar con la selección manual.

Rast:





La escala *Rast* es otra escala popular árabe. En el histograma podemos ver cómo aparece de nuevo el mismo proceso de vibrato, con pequeños picos adyacentes a los

principales. Sin embargo, los picos con información son suficientemente altos como para no confundirse.

En este caso perdemos la sexta, por confundirse con la séptima. El problema en este caso son los valores de fondo no despreciables, que hacen que las regiones no se dividan correctamente. Sin embargo, una decisión incorrecta a la octava inferior de la sexta, aparece en el extremo izquierdo por lo que sí que la podemos encontrar. Este es otro aspecto importante del algoritmo: si en el margen de frecuencias que estamos analizando se produce una decisión incorrecta de octava, la nota aparecerá desplazada, pero en el diagrama de barras aparecerá en su lugar, en este caso, en el lugar de la sexta, aunque desplazada una octava. Esto, aunque hace que haya que confiar más en el usuario, provoca que el algoritmo sea un poco más robusto.

Por algún motivo, también la tercera y la cuarta han sido omitidas.

La decisión automática sufre del problema de que los resultados dependen de la complejidad del fragmento de música analizado. Por tanto, debemos incorporar algún tipo de información externa para adecuar el análisis al tipo de histograma que hayamos calculado y a la información que queramos encontrar.

Una buena forma de controlar el análisis es proponer un umbral de criba de valores variable. El umbral de criba funciona de la siguiente manera: se impone un umbral como un tanto por ciento del valor de amplitud del pico más alto. Los valores del histograma que no superen este umbral son puestos a cero, y ya no participan en el análisis. La consecuencia de este análisis es que hacemos más anchos los valles entre picos, ayudando al proceso de obtención de bordes a encontrar mejores soluciones. Imponiendo un umbral variable, podemos elegir el tipo de información que queremos obtener en cierta manera, puesto que si sólo queremos quedarnos con los picos más altos, impondremos un umbral grande, y si queremos sacar información de picos pequeños, lo rebajaremos hasta que queramos.

La incorporación de este umbral ha hecho que la decisión automática sea mucho más robusta, a la vez que la hace más versátil, puesto que es ahora el usuario quien puede incorporar ahora algo de información y obtener realimentación, lo cual siempre es más amigable.

4.5.2.3 Evaluación sobre música viva.

Llamamos música viva a la interpretada por humanos, frente a la música generada por MIDI, aunque para algunos esta denominación pueda llevar a cierta discusión.

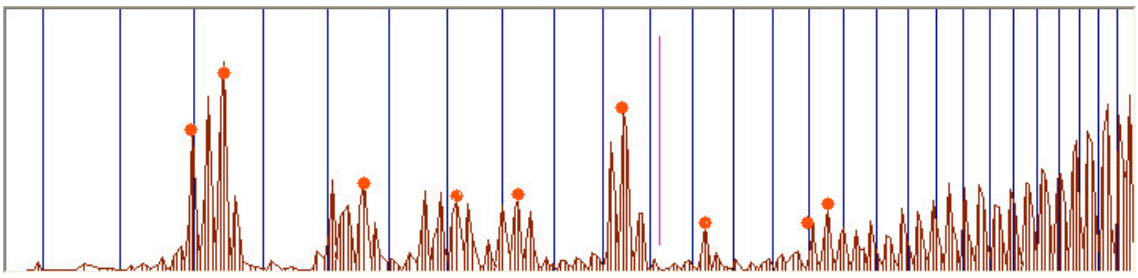
Hemos elegido dos fragmentos de música: el *taksim* turco y la *toná* flamenca. Pensamos que esta selección de materiales es adecuada porque representa muy bien el tipo de dificultades de la música a la que la aplicación va orientada.

Sin embargo, hasta que no analizamos la música, no somos capaces de hacernos idea de su complejidad, a la vez que nos damos cuenta de eventos musicales que pueden escapársenos en la escucha. Esta complejidad de los parámetros que rigen la música en

general, hace que la evaluación no pueda hacerse se una forma demasiado formal, por lo que recorreremos dos ejemplos para dar idea de las posibilidades de la aplicación.

Para empezar, introducimos el histograma del *taksim* y su diagrama de barras y más tarde analizaremos la *toná*.

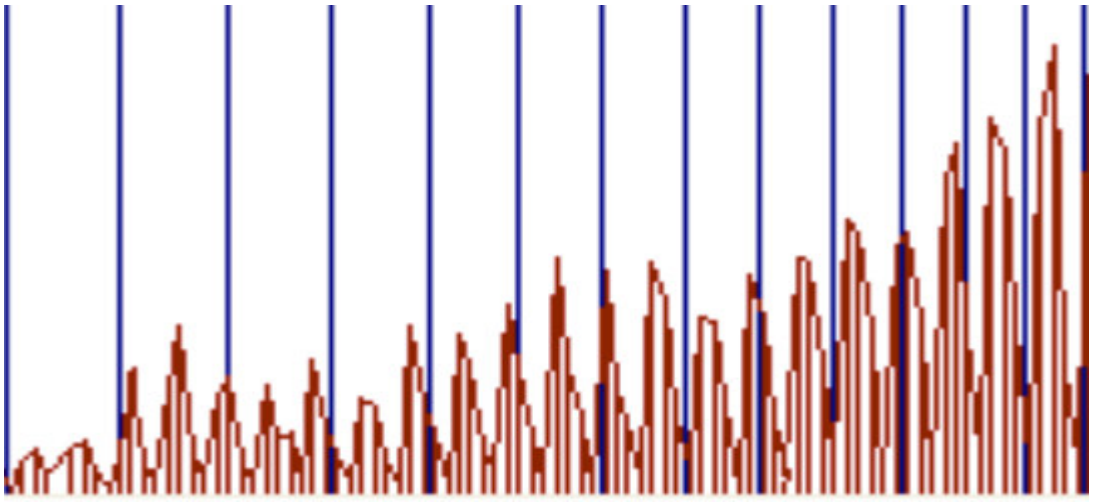
Del proceso de análisis de la música turca obtenemos el siguiente histograma:



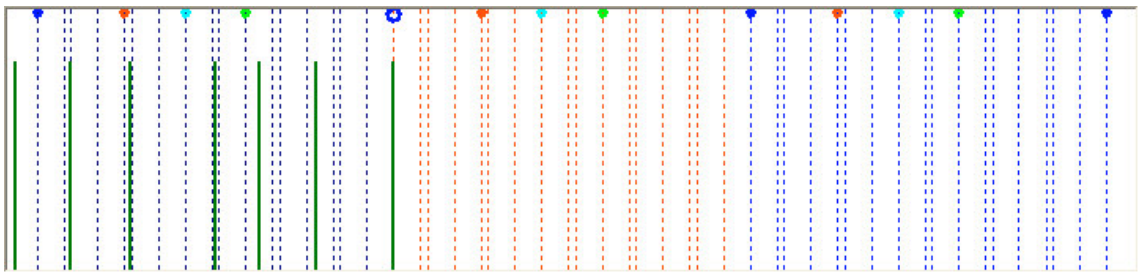
En este caso descubrimos el secreto del histograma escuchando la música, y descubrimos los secretos de la música mirando el histograma. Vemos cómo a baja frecuencia se observan cinco núcleos importantes de notas. Cada núcleo está formado por un conjunto de varios picos muy bien definidos. Como estos picos están tan bien definidos, debemos considerar que comportan información, y no podemos recurrir a técnicas de suavizado, pues estropearíamos el análisis.

En este histograma vemos la capacidad del intérprete por recurrir a notas mucho más pequeñas de medio semitono, algo difícil de abordar por músicos occidentales con tanta precisión. Acercándonos a la tarea que podría realizar un musicólogo, hemos analizado el conjunto. Las notas principales están señaladas con círculos rojos, y consideramos el resto de picos como fruto de micromodulaciones y del efecto del vibrato controlado.

En alta frecuencia vemos un conjunto mucho más grande de picos sin separaciones, con nuevas modulaciones por medios semitonos, que es fruto de la interpretación del cantante. Este proceso de recorrido de improvisación de la interpretación se puede ver en la figura siguiente.

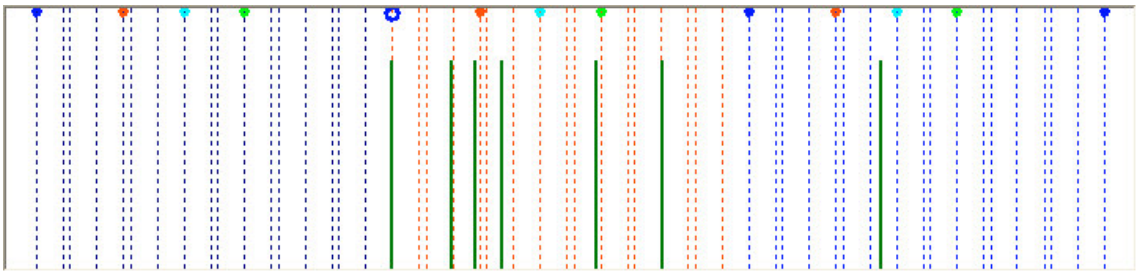


Hemos comprobado por tanto, la robustez del algoritmo a construir histogramas representativos sin pérdida de información, pero ¿cómo hace el algoritmo para sacar las notas con información de la maraña de picos? Comprobemos la decisión que toma la aplicación.



Lo que comprobamos es que la aparición de una alta amplitud en el extremo derecho del histograma, hace que se tome una decisión incorrecta de fundamental cuando esta era mucho más fácil. Por tanto, las barras de las notas salen desplazadas hacia abajo. Sin embargo podemos ver la distancia de quinta entre la nota fundamental y su dominante, las cuales aparecen claramente (la fundamental entre cuarta y quinta y la quinta como segunda), y las demás notas que reconocíamos en el histograma recorriendo circularmente el diagrama.

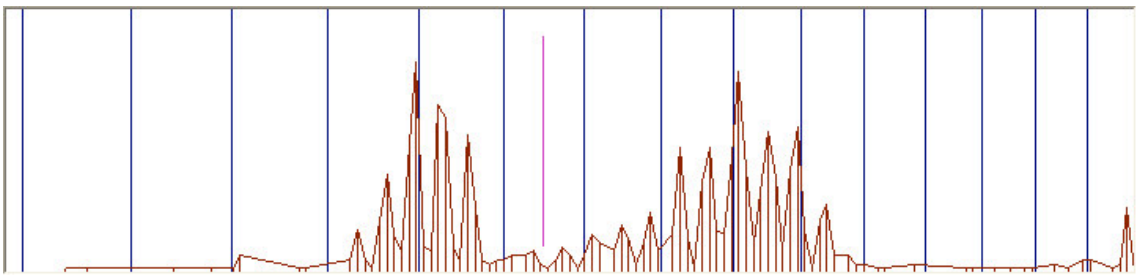
La aglomeración en el extremo derecho del histograma se debe más a la acumulación de estimaciones de sílabas sordas que a verdaderos valores de frecuencia. Por este motivo, se tomó la decisión de ganar en robustez poniendo a cero los valores del extremo del histograma, y así evitar este tipo de problemas. Una vez realizado este cambio, encontramos el siguiente diagrama de barras:



Encontrando claramente ahora la fundamental verdadera, de forma que podemos ver ahora claramente la segunda, quinta y la sexta. La decisión acerca de la tercera es más arriesgada, teniendo que recurrir a decisión manual para terminar la selección.

Por tanto, como conclusión hasta ahora, podemos decir que la decisión automática, al no ser tan robusta como la manual, necesita de la colaboración del usuario. Aunque la información puede aparecer desordenada, está presente en los resultados, por lo que tenemos que dar por sentado un poco de interpretación por parte del lado humano.

Por otro lado, tenemos el ejemplo de la toná. En la siguiente figura, vemos el histograma generado a partir de su análisis.

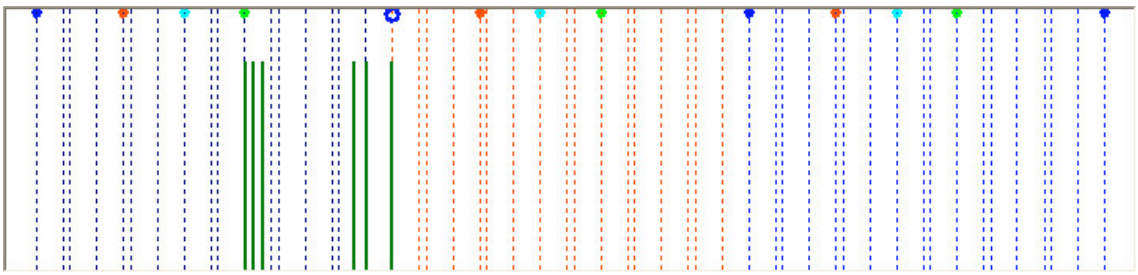


Para comprender el histograma, de nuevo debemos recurrir a escuchar el fragmento. En el proceso de audición, descubrimos cómo el cantante se apoya en una nota principal y otra a distancia aproximada de tercera mayor, basculando continuamente de una a otra, como creando tensión y distensión. El paso entre ambas notas se hace mediante rápidos golpes de glotis. La nota intermedia, escuchándola, no parece una segunda estándar, sino que parece estar más cerca de la tercera cuando asciende a ella y estar más cerca de la fundamental cuando baja. Este presupuesto, que ha sido teorizado por algún estudioso, se ve claramente en el histograma del ejemplo.

Además por encima de la tercera, encontramos un pico que representa a la cuarta. Por debajo de la fundamental, encontramos otro pico que representa a la nota que actúa como sensible de la fundamental, con un pico muy pequeño en amplitud (es el pico del extremo izquierdo).

Las notas que están rodeando inmediatamente a la fundamental y a la tónica nos recuerdan mucho a la característica del vibrato que ya habíamos estudiado con los ejemplos MIDI. Por tanto, como tal los caracterizaremos, y no les asignaremos mayor información.

La decisión automática del algoritmo de selección de picos fue la siguiente:



Por tanto, vemos cómo la selección de la fundamental a sido incorrecta, pero la búsqueda de los picos ha sido muy buena, sacando las notas intermedias que esperábamos.

4.5.2.3 Conclusiones de la evaluación.

De todo este proceso de evaluación general que hemos abordado, podemos concluir que:

- La selección de la fundamental por su amplitud, no garantiza su correcta selección automática. Es necesario un aporte de información externa para solucionarlo.
- La selección de los picos del histograma suele ser buena. La posibilidad de cambiar el umbral de criba, hace que el usuario pueda encontrar el tipo de información que quiera, ganando además en calidad en análisis de bordes.
- El suavizado lineal es una herramienta muy adecuada cuando se obtienen histogramas sencillos. Cuando el histograma es complicado, es preferible recurrir al análisis sin suavizado y después compararlo con el resultado de la selección con suavizado.
- En el proceso de selección automático, los resultados pueden ser poco robustos. La correcta interpretación de los resultados por parte del usuario es fundamental para la obtención de la información adecuada.
- El proceso de selección manual se ha mostrado como de alta calidad y de ayuda. Es el perfecto complemento para el proceso de selección automático.

4.10 Conclusiones a la aplicación de reconocimiento automático de modos musicales

Hemos visto en este apartado la arquitectura software de una aplicación dedicada al reconocimiento automático de modos musicales en voz cantada. En dicha arquitectura encontrábamos dos módulos principales que realizan dos tareas diferentes: la extracción de la frecuencia fundamental de la voz y el procesado de alto nivel de un histograma generado a partir de los resultados del primer módulo.

Se ha mostrado como eficientes las decisiones de hacer más rápido el procesado del primer módulo para realizar sus operaciones en tiempo real, y de interaccionar con el usuario en tiempo diferido para sacar los resultados musicológicos. De la misma forma, se ha apreciado como adecuado el ofrecer la posibilidad de un reconocimiento dual automático y manual, que hace más intuitivo el programa, y más robusto.

El módulo en tiempo real en realidad un estimador prosódico, particularizado a frecuencia fundamental, muy eficiente, y con amplias posibilidades de utilización en otros campos de la investigación de la voz hablada y cantada. Aunque este procesado en tiempo real es computacionalmente muy duro, el avance futuro en la rapidez de los procesadores dará alas a nuevos desarrollos de este tipo de algoritmos en tiempo real.

En el módulo de procesado de alto nivel del histograma residen los principales problemas de la aplicación. Estos problemas no son tanto derivados de las características del diseño del procesado como de la interpretación que la inteligencia limitada de la computadora realiza. Por tanto, el mayor desarrollo futuro de esta aplicación es sin duda la extensión de su inteligencia de búsqueda de picos, de decisión y de discriminación de resultados. Nuestras esperanzas de que estos problemas genéricos se solucionen son grandes, pero quizá no a tan corto plazo como desearíamos si en realidad ello es posible.

En cuanto al interfaz de interacción con el usuario, pensamos que sería adecuada una revisión desde el punto de vista de los profesionales a los que va dedicada la aplicación, para adecuarse más sus necesidades y recoger sus inquietudes, lo cual seguro que aportaría nuevos módulos de procesado a la aplicación.

5. Conclusiones finales

Desde el momento en que empezamos a discutir el tema de este Proyecto Fin de Carrera ha pasado ya un año de trabajo continuado, en el que hemos solucionado problemas de indefiniciones en los caminos a tomar, de inexistencia de documentación útil, es decir, de algoritmos de uso público que nos ayudaran en el desarrollo. Si bien estas dificultades a veces se mostraron como fronteras infranqueables, la superación de las mismas fueron un acicate al esfuerzo, razón por la que ahora, bastantes meses después, duele reconocer que estas serán las últimas líneas que escribiremos.

Los objetivos que nos propusimos al definir este Proyecto Fin de Carrera fueron los siguientes: la programación de una aplicación de procesado de la voz, la consecución de un algoritmo de estimación de frecuencia fundamental, y la integración aplicación y algoritmo en un desarrollo en tiempo real para caracterización de modos musicales.

Si bien estos objetivos han sido claramente conseguidos, como hemos comprobado en la presente memoria, el estudio de una amplia bibliografía ha hecho que adicionalmente hayamos conseguido mayores conocimientos:

- El estudio introductorio de numerosos métodos matemáticos para reconocimiento de habla. Entre ellos encontramos los Modelos Ocultos de Markov, las Redes Neuronales, los Modelos Bayesianos de decisión o el Modelado probabilístico por Gaussianas.

- El conocimiento de un amplio repertorio de posibilidades para la estimación prosódica de la voz, incluso aquellas que son de más reciente publicación.

- El conocimiento profundo de herramientas de programación, sobre todo para el problema del tiempo real.

- El conocimiento profundo de las características articulatorias de la voz humana, así como de los modelos principales que existen para explicar los fenómenos de percepción de la frecuencia.

- El aportar un nuevo conocimiento más científico a la síntesis de escalas musicales, acercando el mundo de la ingeniería y de la música, no tan alejado como antiguamente parecía.

Hemos superado ampliamente, por tanto, los objetivos que nos planteamos al inicio. Fruto de los límites temporales que nos impusimos, existen bordes o vértices mejorables en el desarrollo de la aplicación, puntos que de solucionarse, la consolidarían como una aplicación de alta utilidad y calidad.

En todo caso, se anima al lector futuro de esta memoria a revisar los contenidos y los desarrollos de la actual memoria, y a ponerse en contacto con el autor para transmitirle sus dudas y también sus inquietudes.

ANEXOS

ANEXO 1

Tabla comparativa de los principales intervalos en las escalas temperada, justa y pitagórica:

		entonación pitagórica			entonación justa			temperamento igual	
intervalo	nota	origen numérico	relación interválica	cents	origen numérico	relación interválica	cents	relación interválica	cents
unísono	DO	1:1	1,000	0,0	1:1	1,000	0,0	1,000	0
Segunda menor	REb	28:35	1,053	90,2	16:15	1,067	111,7	1,059	100
Unísono aumentado	DO#	37:211	1,068	113,7	16:15	1,067	111,7	1,059	100
Segunda mayor	RE	32:23	1,125	203,9	10:9 9:8	1,111 1,125	182,4 203,9	1,122	200
Tercera menor	MIb	25:33	1,186	294,1	6:5	1,200	315,6	1,189	300
Segunda aumentada	RE#	39:214	1,201	317,6	6:5	1,200	315,6	1,189	300
Tercera mayor	MI	34:26	1,265	407,8	5:4	1,250	386,3	1,260	400
cuarta justa	FA	22:3	1,333	498,1	4:3	1,333	498,1	1,332	500
Quinta disminuida	SOLb	210:36	1,407	588,3	45:32	1,406	590,2	1,414	600
Cuarta aumentada	FA#	36:29	1,424	611,7	64:45	1,422	609,8	1,414	600
quinta justa	SOL	3:2	1,500	702,0	3:2	1,500	702,0	1,498	700
sexta menor	LAb	27:34	1,580	792,2	8:5	1,600	813,7	1,587	800
Quinta aumentada	SOL#	38:212	1,602	815,6	8:5	1,600	813,7	1,587	800
sexta mayor	LA	33:24	1,688	905,0	5:3	1,667	884,4	1,682	900
Séptima menor	SIb	24:32	1,788	996,1	16:9 7:4	1,777 1,750	996,1 968,8	1,782	1000
Sexta aumentada	LA#	310:215	1,802	1019,1	9:5	1,800	1017,6	1,782	1000
Séptima mayor	SI	35:27	1,900	1109,8	15:8	1,875	1088,3	1,888	1100
octava	DO	2:1	2,000	1200,0	2:1	2,000	1200,0	2,000	1200

Referencias Bibliográficas.

A. Benade, *Fundamentals of musical acoustics*, Oxford University Press, 1976 .

P. Boomlister & W. Creel, *The long pattern hypothesis in harmony and hearing*. In: Journal of Music Theory 5, pp. 2-31. 1961.

A. de Cheveigné, *Pitch Perception Models* , Draft to Appear in Plack , C. and Oxenhann a. (pitch) New York: Springer Verlag 2004.

A. de Cheveigné, *YIN, a fundamental frequency estimator for speech and music*, Acoustical Society of America 2002.

Chunyan Li , Vladimir Cuperman and Allen Gersho*, Robust Closed Loop Pitch Estimation*, 1992.

R. Erickson, *Sound Structure in Music*. Berkeley: University of California (capitulo 2), 1975.

S. Furui, *Digital Speech ProcessingSynthesis and Recognition (Second Edition, Revised and Expanded)* Marcel Dekker, Inc. New York, 2001.

D. Gerhard*, Pitch Extraction and Fundamental Frequency*: History and Current Techniques, Technical Report TR-CS 2003-2006, November 2003.

S. Godsill & M. Davy , *Bayesian Harmonic Models for Musical Pitch Estimation and Analysis*, 2002.

J. Goslin, *Pitch Detection using AM Analysis Techniques*, 1996.

H. Helmholtz, *The sensations of tone*, traducción de la edición alemana de 1877, Dover, 1954.

W. J. Hess, *Pitch and Voicing Determination*.

X. Huang, A. Acero and H-W Hon,, Spoken Language Processing: A Guide to theory, algorithm, and system development Prentice Hall, New Jersey, 2001.

A. Jefremov & W. Bastiaan Kleijn , *Spline Based Continous-time Pitch Estimation.*

J. C. Junqua and J. P. Haton , *Robustness in Automatic Speech Recognition, Fundamentals ans Applications*, Kluwer Academic Publishers, 1996.

H. Kawahara, *Comparative Evaluation of F0 Estimation Algorithms*, EuroSpeech 2001, Scandinavian.

J.D. Markel and A.H Gray Jr., *LinearPrediction of Speech*, Springer- Verlag, New York, 1976.

G. A. Miller, *The magical number seven, plus or minus two: some limits on our capacity for processing information*. In: Psychological Review 63, pp. 81-96. 1956.

H Moran, & C. C. Pratt, *Variability of judgements of musical intervals*. In: Journal of Experimental Psychology 9, pp. 492-500. 1926.

Morton, *Vocal tones in traditional Thai music*. In: Selected reports in ethnomusicology vol. 2, pp. 88-99 . 1974.

F.J. Owens. *Signal Processing of Speech.* Mc. New Electronics.

O`Shaughnessy, *SpeechCommunication. Human and machine*. Addisson-Wesley 1987.

R Plomp. & W. Levelt, *Tonal consonance and critical band-with*. In: Journal of the Acoustical Society of America 35, pp. 548-560. 1965.

R. Plomp, y H. J. M. Steeneken, *Effect of Phase on the Timbre of Complex Tones*. J. Acoust. Soc. Am. 46:409. 1969.

T. F. Quatieri. Speech Signal Processing, Principles and Practice. Prentice Hall Processing Series. 2001

L.R. Rabiner and R.W. Schafer, *DigitalProcessing of Speech Signals*, Prentice-Hall, 1978.

R. A. Rasch y R. Plomp *"The Perception of Musical Tones"* . "The Psychology of Music", Diana Deutsch, Academic Press. 1982.

J. Roederer , *Introduction to the physics and psychophysics of music.* Springler-Verlag. 1973.

J. Sánchez González, *Análisis de intervalos y escalas musicales*, Revista de Musicología, Volumen XII nº 1, 1989.

SANCHEZ, F. J. *Tratamiento de señales cuasiperiódicas. Aplicación a la estimación del tono fundamental.* Tesis Doctoral. Escuela Técnica Superior de Ingenieros Industriales, Madrid, 1982.

SANCHEZ, F. J. "Dissimilarity and Aperiodicity functions; temporal processing of quasiperiodic signals", 9th International Congress of Acoustics, Madrid, p.859, Ju1.1977.

SANCHEZ, F. J. "Application of dissimilarity and aperiodicity functions to fundamental frequency measurement of speech and voiced-unvoiced decision", 9th International Congress of Acoustics, p.523, Madrid, 1977.

R. Shepard, *Circularity in judgements of relative pitch*. In: Journal of the Acoustical Society of America 36, pp. 2346-2353. 1964.

M. Slaney & R. F. Lyon, *A Perceptual Pitch Detector,*

Apple Computer Inc, International Conference on Acustic Speech and Signal Processing, (p 357-360 vol 1), 1990.

I. Tasaki, *Nerve Impulses in Individual Auditory Nerve Fibr*. J. Neurophysiology 17:97 1954.

E. Terhardt, *Oktavspreizung und Tonhöhen der Schiefbung bei Sinustonen*. In: Acustica 22, pp. 348-351. 1969.

E. Terhardt, *Pitch, consonance and harmony*. In: Journal of the Acoustical Society of America 55, pp. 1061-1069. 1974.

E. Terhardt, *Psychoacoustic evaluation of musical sounds.* In: Perception & Psychophysics 23, pp. 483-492. 1978.

J. R. Trotter, *The psychophysics of musical intervals: Definitions, techniques, theory and problems*. In: Australian Journal of Psychology 19, pp. 13-25. 1967.

T. Viitaniemi, A. Klapuri, A. Eronen, *A Probabilistic Model For The Transcription Of Singlevoice Melodies,* Institute of Signal Processing, Tampere University of Technology.

W. D. Ward, *The Subjective Octave and the Pitch of Pure Tones*. Tesis de Doctorado no publicada, Universidad de Harvard, Cambridge, Massachusetts. 1953.

W. D. Ward, *Subjective musical pitch*. In: Journal of the Acoustical Society of America 26, 369-380. 1954.

W. D. Ward, *Musical Perception*, In: J. Tobias (ed.) Foundations of modern auditory theory. Academic Press. 1970.

A. Wood, *The Physics of Music*. Dover

Yong Duk Cho & Hong Kook Kim , *Pitch Estimation using spectral covariance method For Low- Delay MBE Vocoder.*