UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

PROYECTO FIN DE CARRERA
Ingeniería de Telecomunicación

DESARROLLO DE HERRAMIENTAS DE PROCESADO Y VISUALIZACION PARA AUDIO 3D CON AURICULARES

Julio Magro Sastre

Junio de 2016

DESARROLLO DE HERRAMIENTAS DE PROCESADO Y VISUALIZACION PARA AUDIO 3D CON AURICULARES

AUTOR: Julio Magro Sastre
TUTOR: Dr. Ing. Joaquín González Rodríguez

Dpto. TEC
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio de 2016

DESARROLLO DE HERRAMIENTAS DE PROCESADO Y VISUALIZACIÓN PARA AUDIO 3D CON AURICULARES

Julio Magro Sastre

sujeto al programa de Licenciatura en Ingeniería de Telecomunicación,
Escuela Politécnica Superior
para optar al título de ingeniero licenciado (M. Ing.) de Telecomunicación
por la Universidad Autonóma de Madrid

Resumen

La *Auralización* o "realidad virtual acústica" es un término relativamente nuevo. Integra métodos de la física y la ingeniería acústica con la teoría de la Psicoacústica y de reproducción electroacústica [1].
El término Auralización es el análogo de la técnica de "visualización" en video 3D para el audio. En este Proyecto Fin de Carrera se describe el proceso de visualizar ciertas características, efectos o señales del sonido.
Los sistemas estéreo convencionales son capaces de posicionar la *imagen sonora o evento auditivo* solamente en el arco virtual que une los dos altavoces. Una extensión directa de estos sistemas fueron los sistemas de sonido envolvente o sonido Surround, en donde se usan más de dos altavoces para crear una imagen sonora que se puede mover por todo el círculo que contiene a los altavoces.

Por otro lado, los nuevos sistemas de audio 3D pueden posicionar la imagen sonora, usando solo altavoces (o unos auriculares), en cualquier punto de un espacio tridimensional alrededor del oyente.
La Auralización describe el proceso de generación, procesado y playback de audio Surround a la altura de los oídos del oyente. Aplicaciones comunes son la resolución de un problema de Acustica, la mejora de una sala, la simulación de la respuesta en frecuencia de los altavoces para escucha con auriculares, la construcción de un edificio, un coche u otros productos.
Ya que el fin último de los sistemas de audio 3D es convencer a los usuarios u oyentes de que el sonido es emitido desde una posición en la sala donde no existe físicamente una fuente o altavoz, no solo los parámetros físicos sino también los psicoacústicos juegan un papel fundamental en el diseño del sistema.

El concepto de conseguir sonido tridimensional fue investigado por primera vez en relación con la modelización de campos sonoros en salas en 1929. Spandöck procesó señales derivadas de medidas en un modelo a escala de la sala con el fin de poder escuchar la acústica de la sala en el mismo laboratorio. La idea fue bien recibida, pero en esa época no había medios para ponerla en práctica.
Veinte años después, en 1949, se inventaba la cinta magnética. Spandöck presenta finalmente su sistema basado en señales ultrasónicas, modelos de salas a escala y un grabador de cinta trabajando a diferentes velocidades. Los elementos básicos de la auralización se pusieron de manifiesto con este trabajo: modelado de campos sonoros, procesado y reproducción del sonido.

Con el tremendo desarrollo de los computadores, el concepto de simulación y auralización fue reinventado por Schroeder a principios de 1960. No es hasta después, en la década de 1990 para cuando la era del procesado digital (DSP), las velocidades de procesador y espacio de memoria se hacen suficientemente potentes como para procesar simulaciones en ordenadores personales, el momento donde se introduce oficialmente el término Auralización. Otros campos de la acústica también han incluido este término en su argot para referirse a fenómenos relacionados con la *espacialización* del audio, particularmente en los ámbitos de ingeniería de sonido y acústica arquitectónica. Desde entonces, el software y hardware se ha perfeccionado considerablemente y hoy en día el software comercial para la simulación de salas acústicas se considera incompleto sin una opción de auralización mediante la tarjeta de sonido del PC o una interfaz de audio DA/AD.

Buena parte del desarrollo de sistemas de audio 3D se ha basado en un único oyente posicionado en entornos *anecoicos*, lo que simplifica el análisis considerablemente. Sin embargo, esto acarrea normalmente que el sistema solo funcione debidamente en estos entornos aislados acusticamente. Para evitar este condicionamiento, se piensa en que los espacios de escucha sean salas reverberantes y por ello se caractericen con una *respuesta al impulso de la sala (RIR)* o su análogo en frecuenca la *respuesta en frecuencia de la sala (RTF)* de larga duración, debido a la reverberación. A una frecuencia de muestreo de 44.1 kHz (éstandar *de facto* y también usada a lo largo de todo este proyecto) se necesitan miles de coeficientes para los filtros FIR que modelen fehacientemente una RIR. Es por ello que los sistemas de audio 3D requieren de una gran capacidad de cómputo por parte del host. Se hace indispensable aplicar la teoría de Fourier, en concreto algoritmos FFT, para trasladar el problema al dominio frecuencial con el fin dereducir la complejidad computacional.
Aunque estas respuestas al impulso de larga duración puedan dificultar la implementación en tiempo real, permiten estudiar los efectos de un entorno/sala en el rendimiento del sistema.

Los sistemas de audio 3D filtran señales de audio monofónicas mediante una matriz de filtros digitales que depende de la posición de la fuente sonora relativa al oyente; esto es, dependiente de las coordenadas polares (θ, φ, r). En general, las soluciones de estos filtros se componen de dos partes.
La primera es la matriz de *respuestas en frecuencia relacionadas con la cabeza (HRTFs)*, que contiene la información direccional que el oyente debe percibir. Los coeficientes de esta matriz se obtienen normalmente de funciones de transferencia generalizadas y medidas previamente, p.ej. mediante una base de datos. La segunda es la red de *cancelación de Crosstalk (cancelación de XT)*, que invierte la matriz de funciones de transferencia acústicas (entre altavoces y oídos del oyente) de la manera más realista y eficiente posible.
Ya que las HRTFs varían considerablemente de un humano a otro debido a la compleja estructura de estas funciones, que dependen de la complexión física y psíquica así como de la estructura geométrica única de cada oído humano, calcular los filtros mediante HRTFs generalizadas degrada la imagen sonora percibida.

En este Proyecto Fin de Carrera se desea describir en profundidad el estado del arte de estos sistemas así como crear un sistema de audio 3D de estas características usando el software Matlab® R2014b. Para ello, se calculan RIRs mediante una función específica para ello y las HRIRs se obtienen de bases de datos; estas ultimas se implementaron de cuatro formas. La primera es mediante un sencillo modelo matemático que modele una HRTF. Las dos siguientes son dos bases de datos de HRTFs, una elaborada en el MIT Media Lab [1] en Estados Unidos de América y otra por la universidad de Peking PKU en China, la última con la ventaja que depende también de la distancia fuente-receptor y que incluyen HRTFs para cada oído izquierdo (L) y derecho (R). El número de muestras y la frecuencia de muestreo para cada HRTF son fijas y valen 512 muestras y 44.1 kHz, respectivamente. Cada una de estas funciones corresponde con una respuesta al impulso finita (filtro FIR) con 512 coeficientes o taps. La última de las cuatros formas en la que se implementaron HRTFs en este Proyecto Fin de Carrera fue interpolando en las tres coordenadas (θ, φ, r) las HRTFs de la base de datos de la PKU.

Si el sistema de auralización convoluciona un sonido con una BRIR que corresponda, por ejemplo, a un entorno reverberante con un tiempo de reverberación de aprox. 2 segundos, cada BRIR tendrá aproximadamente 23000 coeficientes a 44.1 kHz. Por tanto, se precisan métodos de convolución eficientes, procesadores potentes así como sistemas de interpolación y extracción de características binaurales para reducir el volumen de información en la medida de lo posible.
Un sistema de auralización en tiempo real de alta calidad se presenta como un verdadero reto para la tecnología actual disponible. La solución es encontrar nuevas teorías y aproximaciones de simulación acústica de entornos y auralización con un balance entre precisión y tiempo de computo requerido para obtener el efecto 3D deseado.

En este software de audio 3D desarrollado, la Auralización del audio original se consigue troceando por bloques la señal y dejando que el oyente defina una trayectoria en el espacio que la fuente trazará. Cada bloque de audio (que corresponde a un punto en la trayectoria) se convoluciona con una *respuesta el impulso binaural de la sala (BRIR),* obtenida de la convolución de la HRIR con la RIR.
Los bloques procesados se solapan y suman usando el *algoritmo de Solapamiento y Suma (Overlap and Add Algorithm OLA).* Así se consiguen dos señales, una para cada oído. Estas señales tendrán que ser reproducidas con auriculares para la mejor experiencia.

Summary

The Auralization of sound or Acoustic Virtual Reality or 3D Audio are new methods that use Physics and Sound Engineering together with the Psychoacoustic theory. Auralization of sound is the analogous of Visualization in the area of 3D Video. In this M Sc Thesis, the process of visualization of certain characteristics, effects and audio signals are investigated and developed.
Conventional stereophonic systems are capable of positioning the sound image (or auditory event) only between the arc spanned by the two loudspeakers. The Surround systems were an extension of the stereophonic systems, where two or more loudspeakers were used to create an auditory image that can move through the whole circle spanned by the various loudspeakers. However, the newer 3D audio systems are capable of positioning the sound image at any point in a three-dimensional space using only two loudspeakers (or headphones).

The process of auralization is, indeed, the generation, processing and playback pf surround sound at the listeners' ears. Common applications of auralization are the simulation of a loudspeakers frequency response over headphones, the acoustic treatment of a room or building and also the acoustic simulation in a car or other systems.

The goal of a 3D audio system is to trick the perception of the listener in order to make the sound emanate from a position in the room where a loudspeaker isn't really there. Therefore, not only the physical but also the psychoacoustic parameters play a role in the system design.

Three-dimensional sound was first investigated in 1929 related with the modeling of sound fields in rooms. Spandöck built small rooms such that the tests were carried out on a natural scale model. Doing so, the sound signals could be heard in the lab relatively easy. The idea was subtle, but in 1929 there wasn't really a technology to put this in practice.
Twenty years later the magnetic tape was invented. Spandöck finally brought forward his system based on ultrasonic signals and scaled room models. The fundamental elements of auralization were defined with his work: modeling sound fields, processing and reproduction of sound.

With the formidable development of computers, the concept of simulation and auralization was re-written by Schroeder in the 1960s. Nevertheless, it is not until the 1990s when the DSPs, computer run-times and memories were big enough to run simulations in personal computers. It is then when the term Auralization is officially introduced. Other fields in Acoustics like in Audio Engineering have also introduced the term auralization to refer to the *spatialization of sound.* Since the 1990s, software and hardware have been improved considerably and nowadays commercial software for the simulation of sound in rooms is considered incomplete without an option of auralization via a sound interface or an AD/DA card.

Much of the development of 3D audio systems has been based on a single user/microphone positioned in anechoic environments. This makes the analysis much easier but makes the system only usable in these acoustic isolated environments. To prevent this, one thinks in environment as a reverberant room modeled via a *room impulse response (RIR)* or the equivalent *room transfer function (RTF).*

The RIR has a considerable duration because of the reverberation. Establishing the sample rate to 44.1 kHz (standard de facto and also the standard for this Thesis) one requires thousands of taps for the FIR filters that correctly model the RIR. This is the reason that 3D audio systems need great amounts of computing capacity by the host. Because of it the Fourier theory is indispensable: FFT algorithms for looking at the problem in the frequency domain and so reduce the complexity.
Although these RIRs may cause difficulty in the implementation in real time, they enable to study the effects of a room in the global system.

3D audio systems filter audio signals using a matrix of filters that account for the position of the sound source relative to the receiver. That is, dependent on the polar coordinates (r, θ, φ). Generally speaking, the solutions to these filters are made up of two pieces.
The first one is the Head-Related Transfer Functions (HRTF) matrix, which holds the directional information for the receiver. The matrix coefficients are derived from transfer functions which were previously measured or from a data base.
The second one is the Crosstalk Cancelling Network. It reverts the acoustic transfer functions matrix (between loudspeakers and the ears of the listener) in the most realistic and efficient way.
Because HRTFs vary a lot between humans, using generalized HRTFs degrades the perceived sound stage.

The goal of this Thesis is to widely describe these 3D audio systems and also to develop a system using Matlab® R2014b. To this end, RIRs are computed using a function and HRIRs are extracted from data bases in four ways. The first way is to use a simple mathematical model. The second and third way are two HRTFs data bases, one developed at the MIT Media Lab in the USA [1] and the other at the Peking PKU in China. They include HRTFs for each ear left (L) and right (R). The third way has the advantage that it also depends on the source-receiver distance. The number of samples as well as the sample rate are fixed and of value 512 samples and 44.1 kHz, respectively. Each HRTF corresponds to a finite impulse response (FIR filter) with 512 samples or taps.
The fourth way that HRTFs were obtained was by interpolating the HRTFs of the PKU database in the polar coordinates (r, θ, φ).

Efficient convolution methods are required, powerful processors as well as interpolation systems to minimize the amount of data. The reason is that if an auralization system convolves an input sound with a BRIR that corresponds to a reverberation room with a reverberation time of let's say, 2 seconds, each BRIR will have approx. 23000 taps at 44.1 kHz.
An auralization system that operates in real time is a real challenge with the actual technology.

Keywords

Auralización, binaural, oído externo u oreja o pabellon auricular (pinna), oído (canal de entrada al sistema auditivo), objeto auditivo, localización, posicionamiento, fuente virtual, evento sonoro, Interaural Time Difference (ITD), Interaural Level Difference (ILD), Interaural Coherence / Interaural Cross Correlation (ICC),
Head-Related Impulse Response/Transfer Function (HRIR / HRTF),
Room Impulse Response/Transfer Function (RIR / RTF),
Binaural Room Impulse Response/Transfer Function (BRIR / BRTF),
Discrete Fourier Transform / Fast Fourier Transform (DFT / FFT), algoritmo Overlap and Add (OLA),
Digital Signal Processing (DSP), Graphics Processor Unit (GPU).

Agradecimientos

Gracias al Dr. Joaquín González Rodríguez, co-director del departamento TEC de la Escuela Politécnica Superior de la Universidad Autonoma de Madrid, España. Por su apoyo y sus clases magistrales sobre Ingeniería Acústica que me sirvieron para afianzar mi camino.

Dedicado a mi familia y a mi mujer.

ÍNDICE

Lista de símbolos de la memoria y del software en Matlab® adjunto

BRIR	*respuesta al impulso binaural de la sala*
BRTF	*respuesta en frecuencia binaural de la sala* $\mathrm{BRTF} = \mathrm{RTF} \cdot \mathrm{HRTF}$
c	*velocidad del sonido en el aire a temperatura Tc ($c \cong 343$ m/s)*
DA/AD	*conversor digital-analógico/analógico-digital*
eL(n)	*señal de entrada al oído izquierdo (componente izquierda L)*
eR(n)	*señal de entrada al oído derecho (componente derecha R)*
f	*frecuencia en Herz*
FFT	*transformada rápida de Fourier*
fs	*frecuencia de muestreo (igual a 44.1 kHz), modificable*
h(t)	*respuesta al impulso, en general*
hrir_length	*número de puntos de la HRIR (512 u otros)*
HRIR	*respuesta al impulso relacionada con la cabeza*
HRTF	*función de transferencia relacionada con la cabeza*
j	*número imaginario (igual a la raíz cuadrada de –1)*
λ	*longitud de onda en metros*
L	*longitud de bloque de entrada x(n)*
L	*referencia al canal izquierdo*
M	*número de puntos de la* BRIR, $\mathrm{M} = \mathrm{hrir\_length} + \mathrm{rir\_length} - 1$
n	*tiempo discreto del muestreo en tiempo*
N	*número de puntos de la* FFT
Nsig	*número de muestras total de la señal de audio s(n)*
nt	*nº total bloques igual al nº puntos del muestreo espacial de la trayectoria*
Ny	*número de puntos de la señal binaural,* $\mathrm{Ny} = \mathrm{Nsig} + \mathrm{M} - 1$
r	*radio o distancia transmisor-receptor en metros*
R	*referencia al canal derecho*
rir_length	*número de puntos de la RIR (mayor que 20000)*
RIR	*respuesta al impulso de la sala*
RTF	*respuesta en frecuencia de la sala*
RX	*receptor de sonido*
s(t)	*señal de audio analógica (continua) de entrada*
s(n)	*señal de audio digital de entrada, muestreada*
t	*tiempo continuo en segundos*
TX	*transmisor de sonido*
θ	*ángulo azimutal u horizontal en sentido antihorario, de 0º a 360º*
φ	*ángulo de elevación o vertical, de -90º a +90º*
x(n)	*bloque de la señal de entrada s(n) de longitud L*
y(n)	*bloque de la señal binaural de salida,* $\mathrm{Y(k)} = \mathrm{X(k)} \cdot \mathrm{BRTF}$
yL(n)	*señal binaural de salida (componente izquierda L) de longitud Ny*
yR(n)	*señal binaural de salida (componente derecha R) de longitud Ny*
yLp(n)	*señal binaural de salida (componente izquierda L) post-procesada*
yRp(n)	*señal binaural de salida (componente derecha R) post-procesada*

1 INTRODUCCIÓN A LA ACÚSTICA Y LA PSICOACÚSTICA

1.1. Introducción

Las aplicaciones típicas de sistemas de audio 3D complementan, modifican o sustituyen atributos del sonido para poder tener control sobre la percepción auditiva espacial del individuo. Este control no puede obtenerse modificando sólo los atributos físicos; hay que tener en cuenta consideraciones del ámbito de la Psicoacústica. Estos dos factores juegan un papel importante en el análisis, diseño y puesta en marcha de cualquier sistema de audio 3D. Para manipular la percepción espacial auditiva, el ingeniero debe haber profundizado en los fenómenos psicoacústicos que ocurren en la situación de escucha espacial del mundo real. Modificando los parámetros asociados a estos fenómenos acústicos, se podrá modificar la percepción espacial del oyente [1, 2]. Los seres humanos localizan el sonido debido a la percepción binaural. El sistema auditivo humano recibe como entrada ondas acústicas y vibraciones elásticas de los solidos y fluidos a su alrededor con los que esta en contacto mecánico permanente. El camino a los órganos receptivos, los "micrófonos" del sistema auditivo humano, es bien a través de los canales auditivos o bien mediante conducción por los huesos del cráneo, aunque está última no se suele tener en cuenta si el medio de transmisión es el aire. Se puede escuchar sólo con un oído, pero la escucha mediante los dos oídos, la *escucha binaural*, ofrece una serie de ventajas con respecto a la *escucha monoaural*. Esto se debe al hecho de que la escucha binaural provee información adicional al sistema auditivo, que está codificada en las diferencias entre las señales de entrada a cada oído. Además, el tener dos orejas posicionadas en sitios diferentes en el campo de onda acústico permite al cerebro concentrar la atención en procesar el sonido del oído con mejor relación señal a ruido.
El término "Psicoacústica" engloba todo lo relacionado con la descripción y modelado del sistema auditivo humano. Los modelos psicoacústicos intentan extraer datos característicos relacionados con las distintas dimensiones de escucha (tales como volumen, calidez, envolvimiento, brillo, …). Estas dimensiones son descriptores físicos directos del sistema auditivo y forman en conjunto el "carácter" del sonido.
Ademas, podemos juzgar la "calidad" del sonido. Para ello, nos basamos en procesos neuronales que usan nuestra memoria, lo que hemos aprendido y lo que esperamos escuchar. No obstante, para esto se tienen que tomar en cuenta el contexto del evento sonoro y una referencia para el análisis. Asumiendo que estas diferencias entre las señales izquierda y derecha se representen mediante un sistema lineal e invariante con el tiempo (sistema LTI) las diferencias interaurales más importantes son *las diferencias interaurales en el tiempo de llegada (interaural time difference ITD) y las diferencias interaurales de nivel/intensidad (interaural level/intensity difference ILD o IID)*. En general, ambas ITD e ILD dependen de la frecuencia de la señal. Hay que notar que la condición de sistema LTI no siempre se cumple; por ejemplo, cuando los objetos (fuentes sonoras, superficies reflectoras u oyentes) en el campo sonoro se mueven rápidamente.

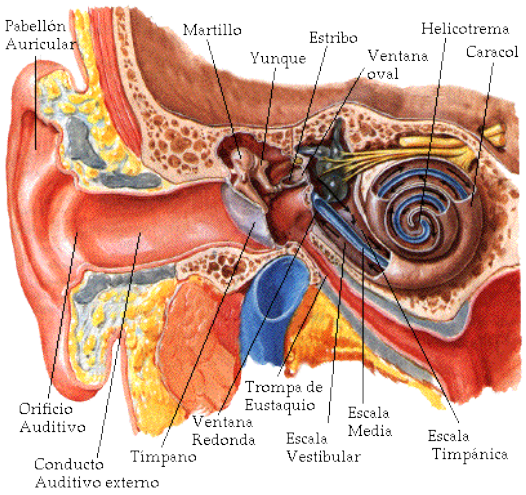
Este capítulo describe brevemente las partes fundamentales para la exposición de este trabajo sobre el sistema auditivo humano y la Psicoacústica. La información de esta sección se ha obtenido de diversas fuentes [1, 2, 3, 4, 5, 6, 7] y recopilado en este capítulo.

1.2. Fundamentos del sistema auditivo humano

El conocimiento del sistema auditivo humano y los correspondientes procesos relativos al tratamiento de las señales acústicas llevan directamente a modelos de percepción con una determinada resolución temporal y frecuencial, volumen sonoro, modulación y enmascaramiento temporal y simultáneo. Los modelos de los procesos troncales, en particular los que usan ambos canales de entrada (oído izquierdo y derecho) proveen las llamadas características "binaurales" e interaurales tales como la localización de la fuente, el efecto de precedencia y el envolvimiento (ver secciones siguientes).

La escucha binaural es una rama específica de la Psicoacústica. Para una auralización realista o al menos plausible, conocer los procesos de la escucha binaural es fundamental. Las consecuencias de la escucha binaural son la localización direccional y la percepción de distancia. Más aún, la escucha binaural activa un mecanismo de cancelación de ruido muy efectivo para entender voz hablada en entornos ruidosos, incluso en situaciones de reverberación en salas y en condiciones de señales de voz concurrentes (fénomeno conocido como *efecto cocktail party*).

El sistema auditivo se compone principalmente de tres partes: 1) oído externo, 2) oído medio y 3) oído interno. Las tres partes se explican en detalle a continuación.



1.2.1. Oído externo y oído medio

Los cambios en la presión sonora son recibidos por la oreja o *pabellón auricular (pinna)* que es la parte fundamental del oído externo junto con el canal auditivo que conduce a la membrana timpánica.
El oído externo consta del pabellón auricular (la oreja) y el canal auditivo. Este canal tiene forma tubular de aprox. 2.7 cm de longitud y 6-8 mm de diámetro. El canal auditivo se comporta como un resonador $\lambda/4$ [5] con una frecuencia de resonancia de aprox 3 kHz. La sensibilidad en la audición tendrá un máximo en ese rango frecuencial. La figura anterior es clara con respecto a la estructura general del aparato auditivo humano. Excepto para altas frecuencias, el pabellón auricular no tiene influencia ninguna sobre el sonido más que actuar como receptor.

Al final del canal auditivo se encuentra el tímpano, una piel de unos $0.8cm^2$ de área y 0.1mm de espesor. El oído medio empieza justo después del timpano. Engarzados con el tímpano se encuentran los huesos del oído medio: el más exterior se denomina martillo (maleus), luego se encuentra el yunque (incus) y finalmente el estribo (stapes). La parte inferior del estribo está conectada con la ventana oval que es parte del oído interno. La función principal del oído medio es asegurar la transmisión eficiente de ondas sonoras provenientes del aire a los fluidos de la cóclea. Actúa como un adaptador de impedancias que mejora la transmisión de sonido. El efecto principal del oído medio es actuar como un transductor

presión/fuerza. Es notable que la faringe conecte con el oído medio. Esta conexión permite igualar las presiones entre ambas caras del tímpano para que no se taponen los oídos en los cambios de presión atmosférica.

1.2.2. Oído interno

La ventana oval conecta el estribo como el oído interno. La cóclea es la parte fundamental del oído interno junto con los canales semicirculares que tienen la función de dotar al humano del sentido del equilibrio. La forma de ésta es similar a una espiral con 2.5 vueltas y una longitud de unos 3 cm. Está situada dentro del hueso petroso, extremadamente duro. La membrana basilar varía significativamente en anchura y rigidez. la rigidez en la parte mas cercana a la ventan oval es mayor que la del final de la membrana. La cóclea, en concreto la membrana basilar de su interior, realiza la transformación de ondas sonoras en impulsos nerviosos para ser procesados por el cerebro. Es un laberinto de múltiples huesos.

La cóclea se divide en tres tubos llenos de fluido. Estos tubos están separados por membranas: entre el primero y el segundo se encuentra la membrana de Reissner; entre el segundo y el tercero se encuentra la membrana basilar. El primer tubo se denomina escala vestibular, el segundo tubo escala media y el tercer tubo escala timpánica. En la figura se puede ver un esquema general del aparato auditivo humano con varios detalles del oído interno.

Los movimientos mecánicos de los osículos (martillo, yunque y estribo) debidos a la presión sonora mueven los fluidos cocleares, la ventana oval y la membrana basilar. La onda de sonido que se propaga dentro del fluido excita a la membrana basilar y crea ondas transversales. Cuanto mejor sea la adaptación de impedancias entre la onda viajera y la masa y rigidez local de la membrana basilar, mayor amplitud tendrán estas ondas transversales. En consecuencia, la amplitud de vibración en la membrana basilar varía con la frecuencia a lo largo de su longitud. Para un tono puro, la amplitud de la onda resultante en la membrana basilar será máxima en una determinada posición de la membrana, para colapsar seguidamente. La posición de este máximo depende de la frecuencia del tono; las altas frecuencias tienen el máximo cerca del estribo y las bajas cerca del ápice del canal. La cóclea es, en resumen, un analizador de espectro compuesto de un banco de filtros paso-banda. Este fenómeno se conoce como *transformación frecuencia-espacio (frequency-space transformation).*

Posterior al análisis, se tendrán que convertir las vibraciones mecánicas a impulsos nerviosos.
El llamado *órgano de Corti* gira en espiral por la cóclea y mide de 25 mm a 35 mm. Se encuentra entre la membrana basilar y la tectorial. Los desplazamientos locales que se producen en la membrana basilar se detectan mediante células sensoriales en el órgano de Corti. Se distinguen dos tipos de células (ciliadas) sensoriales dependiendo de su posición relativa al órgano de Corti: *las células interiores (inner hair cells IHC) y las células exteriores (outer hair cells OHC)*. Las IHCs forman una sola fila de receptores y están conectadas con el nervio auditivo (que lleva al cortex auditivo en el cerebro) mientras que las OHCs son mucho más numerosas y suelen estar organizadas en tres filas. Cada célula nerviosa responda a una cierta frecuencia dependiendo de la posición de su conexión con las IHCs.

1.2.3. Nervio auditivo y núcleo coclear

Las vibraciones en la membrana basilar producen desviaciones de las estereocilia sensoriales, que actúan como un amplificador de banda corta, situadas encima de las células ciliadas. Estas desviaciones abren y cierran canales de transducción mecánico-eléctricos, produciendo cambios en las corrientes de estos canales. El cambio de voltaje en las membranas de las neuronas sensoriales hace que se liberen partículas neuroquímicas y que se estimulen neuronas de las células ciliadas interiores. El voltaje presente en las OHCs controla un mecanismo de generación de fuerza conocido como “amplificador coclear”; estas fuerzas aplicadas a las moléculas incrementan la vibración de la membrana basilar.

Otra característica característica importante del órgano de Corti es que provee de realimentación. Las OHCs tienen de poca a nada de influencia en la transducción del sonido directo hacia el cortex. No obstante, las OHCs juegan un papel importante en aumentar la amplitud efectiva de vibración de las IHCs. Mediante este proceso es posible incrementar la sensibilidad en varias decenas de decibelios, así como la resolución frecuencial.

El nervio auditivo transmite los impulsos generados por las IHCs al núcleo de la cóclea. Si la estimulación es con tonos puros individuales con niveles por encima de un cierto límite, el ratio de respuesta medio del nervio auditivo aumenta casi linealmente con el logaritmo de la presión sonora (en un rango limitado de intensidad). En la pegada o parte inicial del tono, las fibras del nervio auditivo se descargan muy rápidamente. Luego, más lentamente y de manera constante. La diferencia entre el pico de descarga y el estado estacionario de descarga se conoce como "adaptación". Ésta ocurre en tres fases: la fase inicial es de pocos milisegundos, la segunda del orden de decenas de milisegundos y la tercera, la más larga, puede durar incluso decenas de segundos.

El nervio auditivo conecta con el núcleo coclear y se dividen dos ramas: una rama ascendente que engarza con el núcleo coclear anteroventral y una rama descendente que engarza con ambos núcleos posteroventrales y dorsales.

1.2.4. Tronco encefálico

Ya en este nivel y en los siguientes, la información presente en el nervio auditivo ha sufrido y va a sufrir transformaciones importantes. La información que se ha procesado en ambos oídos es integrada para extraer las importantes y relevantes *características binaurales (ITDs, ILDs, ICCs, …).* También se extraen ciertas características monoaurales para la estimación del ángulo de elevación.

El complejo olivar superior es una colección de núcleos ubicado en la región más baja de la protuberancia en el tallo encefálico. Participa en múltiples aspectos de la audición. Es el primer punto donde convergen los aferentes de los dos núcleos cocleares del oído izquierdo y derecho. En los humanos y en la mayoría de mamíferos podemos encontrar dos tipos de células binaurales ubicadas aquí: las olivas centrales superiores y las olivas laterales superiores reciben impulsos binaurales y contienen neuronas sensibles a las ITDs.

Las células en las olivas centrales superiores reciben excitación de ambos nucleos cocleares y se denominan por ello exitatorias-excitatorias (EE). También se llaman *células detectoras de coincidencia*. Conceptualmente, un conjunto de neuronas EE con diferentes retardos caracteristicos se puede comparar a una funcion de correlacion cruzada. La tasa de descarga de estas células en respuesta a estimulos binaurales depende de la ITD. Si cierta neurona es activada por diferentes frecuencias, las diferentes curvas de descarga periódicas muestran un máximo de amplitud para el mismo retardo interaural que el del estimulo. Este retardo se conoce como el retardo característico de la célula y provee de una estimación del tiempo que tarda el estímulo en viajar de cada oído al detector de coincidencia.

Recientemente se han modelado mediante software todos los componentes del sistema auditivo humano y se han recopilado en una herramienta llamada "Auditory Modelling Toolbox AMT" [8]. Este software para Matlab®/Octave® merece una mención especial por la extensión y calidad de los modelos acústicos, psicoacústicos y neuronales implementados.

1.2.5. Características psicoacústicas

Resolución frecuencial y nivel de volumen

Como se ha explicado, en el oído interno ocurre un análisis espectral de los sonidos entrantes. Esto permite al sistema auditivo percibir frecuencias en un rango de aprox. 50 a 16000 Hz. El rango especifico depende de la edad, las influencias de ruido ambiental, habilidades, discapacidades, etc.
La *ley de Weber y Fechner* es una aproximación de primer orden para muchas percepciones sensoriales humanas. Dicta que las percepciones son proporcionales al cambio relativo del estimulo físico medible. La impresión de volumen sonoro debería por ello ser representada en una escala logarítmica dependiente de la presión sonora (en la escala de dBs). No obstante, notar que la escala de dBs no es exactamente una escala lineal subjetiva pues el doblar la impresión de volumen sonoro subjetiva requiere incrementar 10 dB. Hay que resaltar que la sensibilidad no es la misma para cada frecuencia. Para una descrpicion de la sensibilidad, se dibuja el *nivel de presión sonora (sound pressure level SPL)* en dBs que se percibe a un cierto *volumen sonoro o nivel de volumen (loudness)*, en función de la frecuencia. El resultado es el llamado *campo auditivo.* El campo auditivo tiene su tope en el umbral de dolor, que es aprox. 120-130 phon.

El *nivel de volumen* L_N se define como el nivel de un tono de 1 kHz que es percibido como si tuviera el mismo volumen. El valor de referencia es la presión sonora mínima a la que se puede escuchar un tono de 1 kHz, denotada por $\mathbf{p_0 = 2 \cdot 10^{-5}\ Pa}$:

$$\mathbf{L_N = 20 \cdot \log\left(\frac{p_{1\ kHz}}{p_0}\right)\ [dB]}$$

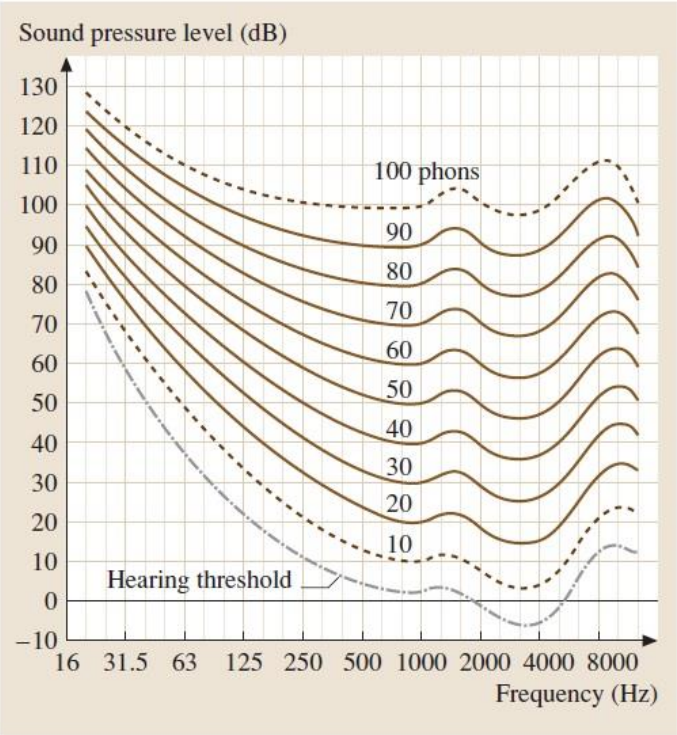
La figura de la página siguiente muestra un campo auditivo típico. La curva más inferior es el *umbral de audición (threshold of hearing)*. Es el SPL mínimo al que ese sonido es escuchado. Las curvas superiores son las *isófonas*, i.e. curvas con el mismo SPL o mismo nivel de volumen. Éstas se determinan comparando con un tono ajustable o predefinido de 1 kHz. Resalta que un tono con un SPL constante sea percibido como si tuviera un volumen diferente en cuanto se varia la frecuencia.

El motivo para un nivel sonoro igual en rangos frecuenciales amplios se relaciona con la dependencia de estimulos físicos dependientes de la frecuencia para cada SPL. El rango audible se representa por un conjunto de curvas que tienen el mismo nivel de volumen para tonos puros. La unidad más antigua usada en este ámbito es el *phon*. Para un tono puro de 1 kHz, el valor en dB es idéntico al valor en phons.

Es preciso notar que el volumen subjetivo de un sonido no es directamente proporcional a su nivel de volumen medido en phons. Por ejemplo, un sonido con un nivel de 80 phons suena mas que el doble de alto que un sonido con un nivel de 40 phons. Se ha sugerido en base a experimentos que el volumen L es una funcion potencia de la intensidad de la onda sonora física I (*ecuación de la ley de potencia*) [3]:

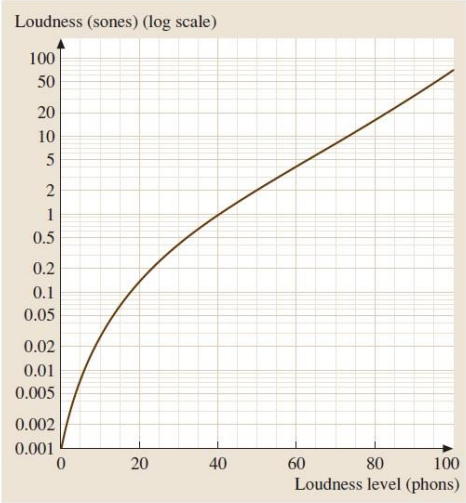
$$\mathbf{L = k \cdot I^{0.3}\ x}$$

donde k es una constante que depende del sujeto y unidades usadas. En otras palabras, el volumen de un sonido es proporcional a su intensidad elevado a 0.3. Por ello es que el volumen no se relaciona linealmente con la intensidad; más bien el volumen es una funcion compresiva de la intensidad.
Una aproximación a esta ecuación es que el volumen se multiplica por 2 cuando la intensidad se incrementa en un factor 10. Equivalentemente, cuando el nivel de volumen se incrementa 10 dB. En la práctica, esta relación solo es valida para niveles sonoros por encima de unos 40 dB SPL. Para niveles mas bajos que esos, el cambio de volumen con la intensidad es mucho mas pronunciado que el descrito con la ecuación de la ley de potencia y por tanto la ecuación no aplica a este caso.



Si uno examina el conjunto completo de curvas llega a la conclusión de que la sensibilidad máxima en la audición se produce a la frecuencia de 3-4 kHz. Resulta que este rango de frecuencias coincide con la banda en la que el oído externo (el pabellón auricular u oreja) es resonante. Teniendo en cuenta cuan pequeño es el timpano también se puede comprender por qué el humano tiene dificultades para responder a bajas frecuencias (por esto las curvas de igual volumen sonoro convergen a bajas frecuencias a altos niveles de presión sonora). La masa del timpano asi como otros elementos constituyentes del oído limitan la respuesta a las frecuencias mas altas (por esto las curvas tienden hacia arriba en las altas frecuencias) aunque con anomalías, quizás debidas a las limitaciones fisiológicas de lal oído interno asi como a resonancias particulares.El que cada curva del conjunto tengan curvaturas ligeramente diferentes pone de manifiesto la no linealidad de nuestra percepción auditiva; esto es, la sensibilidad varia con el nivel absoluto de volumen sonoro.

Una mejor estimación de la impresión subjetiva del nivel de volumen esta dada por una escala en unidades llamadas *sone*. Un sone se define arbitrariamente como el volumen de una sinusoide de frecuencia 1kHz a 40 dB SPL, que se transmite desde una posición fija enfrente del oyente en campo libre. La siguiente figura muestra la relación entre el volumen en sones y el nivel de volumen físico en phons de una sinusoide de frecuencia 1kHz. El nivel del tono de 1 kHz es igual a su nivel de volumen en phons.Ya que el volumen en sones se ha graficado con una escala logarítmica, y el dB es una unidad ya de por si logarítmica, la curva de la figura se aproxima a una línea recta para nivel de mas de 40 dB SPL.



Los ingenieros acústicos y psicoacústicos han desarrollado filtros en base a los contornos que se traducen de phons a sones, que aproximadamente tienen en cuenta la respuesta en frecuencia del oído. Una respuesta del filtro suave a niveles bajos se implemento en la llamada escala "A-weighed" y el nivel medido con un filtro A se expresa en dBA. Para expresar la sensibilidad frecuencial del oído a nivel mas altos, también se crearon filtros B, C y D. Debido al comportamiento no lineal del sistema auditivo (que es la razón por la cual se crearon diferentes curvas ponderadas para diferentes niveles) no se puede modelar la impresión subjetiva correctamente mediante un sistema lineal cerrado.

Enmascaramiento frecuencial

Otra consecuencia de los mecanismos de la coclea no lineales se da en el llamado efecto de *enmascaramiento (masking)* y que juega un papel fundamental en la producción de música pero también en los sistemas de detección de voz para entornos ruidosos y/o con múltiples hablantes [1, 3].
Cuando hay un sonido en una cierta banda y en las bandas adyancentes hay otros sonidos, uno o varios sonidos pueden enmascarar a los adyancentes. El cálculo de la presión sonora en dBA no es realmente suficiente para describir el volumen sonoro percibido. Se consiguen mejores resultados si se separa el espectro de la escena sonora completa en ciertas bandas críticas para tener en cuenta las máscaras entre bandas (estas máscaras se caracterizan por ser ruidos banda estrecha en las frecuencias 250 Hz, 1 kHz y 4 kHz). Es más normal que las bajas frecuencias oculten las altas frecuencias, aunque lo contrario también se da. Las máscaras representan el SPL mínimo requerido para percibir un tono que está superpuesto con otro tono o una máscara. Como aproximaciones a las curvas de enmascaramiento exactas del sistema auditivo humano, existen curvas estandarizadas en la norma ISO 532 B para evaluar el volumen sonoro total y también el local.

La banda de audio puede dividirse de manera eficiente en aproximadamente seis rangos distintos, cada banda con importancia capital en el sonido total. Estas seis bandas son:

1. Sub-graves. Las frecuencias muy bajas (entre 16 Hz y 60 Hz) son "sentidas" o "percibidas" más que oídas; tal como un rayo en la distancia. Estas frecuencias dan a la música una cierta potencia cuando están presentes, aunque sea por un breve periodo de tiempo. Demasiado énfasis en este rango de frecuencias hace que la música suene "turbia", como con ruido no audible.

2. Graves. Las frecuencias bajas (entre 60 Hz y 250 Hz) contienen las notas fundamentales de la parte rítmica. La ecualización de los graves permite modificar el balance musical; demasiada amplificación de este rango hace que la música suene como si retumbase.

3. Medios (bajos). Las frecuencias medias bajas (entre 250 Hz y 2 kHz) contienen los armónicos de menor orden de la mayoría de los instrumentos musicales; si se amplifican en demasia degradan la calidad hasta incluso al orden de la telefónica.

4. Medios (altos). Las frecuencias medias altas (entre 2 kHz y 4 kHz) pueden enmascarar las frecuencias importantes para el reconocimiento de sonidos y voz si están demasiado amplificadas. Además, aumenta la fatiga auditiva.

5. Presencia. Frecuencias mas altas que las medias (entre 4 kHz y 6 kHz) son las responsables de la claridad y definición en las voces e instrumentos. La amplificación en este rango hace que la música se "acerque" o "aleje" del oyente.

6. Brillo. El rango de frecuencias mas alto (entre 6 kHz y 16 kHz) controla la definición y claridad de los sonidos. Demasiado énfasis en este rango, no obstante, puede acarrear que las vocales se vuelvan sibilantes.

Enmascaramiento temporal

Si el audio tiene partes con transitorios muy rápidos en forma de por ejemplo una serie de impulsos, también se da el fénomeno de enmascaramiento. Si el oído humano recibe una onda sonora de una fuente que emita un impulso, hará falta un cierto tiempo en que se recargen las diferencias de carga de los potenciales electrostáticos de las células ciliadas IHCs y OHCs. Este efecto se conoce como *post-enmascaramiento*.
También existe el *pre-enmascaramiento* en el sentido contrario. No obstante la explicación de este efecto requiere de un conocimiento avanzado de los procesos neuronales más altos y sale del ámbito de este proyecto.

Volumen sonoro variable con el tiempo

Es más común que haya sonido fluctuante que sonido estacionario, aunque los estándares actuales estén orientados a sonidos estacionarios. El oído es capaz de discriminar niveles de volumen en intervalos de 10 ms. Para aplicaciones psicoacústicas este efecto será relevante en tanto se quieran tratar sonidos impulsivos o modulados, que pueden ser también de música o habla. Ver estándares más actuales en DIN 45631.

Nitidez

La percepción de un sonido nítido es una impresión subjetiva e independiente del carácter del sonido. La nitidez está determinada principalmente por el balance entre las componentes espectrales de baja y alta frecuencia e incremente cuando aumenta la densidad de componentes de alta frecuencia.

Se define 1 acum como la nitidez producida por un ruido paso-banda en la banda critica 1 kHz a 60 dB SPL. La nitidez de una señal arbitraria se obtiene calculando un centro de gravedad del nivel de volumen específico $\mathbf{N'}$ y se corresponde con la ecuación:

$$\mathbf{S = 0.11 \frac{\int_0^{24\ bark} N' g(z) dz}{\int_0^{24\ bark} N' dz}\ acum}$$

siendo S la nitidez, $\mathbf{N'}$ el nivel de volumen especifico en escala bark (z) y g(z) un factor de escala.

Intensidad fluctuante y aspereza

Si el sonido esta modulado en amplitud a baja frecuencia (hasta 20 Hz), los mecanismos de percepción humanos causan que percibamos fluctuaciones en el sonido.

La fuerza o intensidad de fluctuación se mide en *vacil* y se define la unidad como un tono puro a 1 kHz y 60 dB SPL con una modulación 100% en amplitud a 4 Hz. Se corresponde con la ecuación:

$$\mathbf{F = 0.08 \frac{\int_0^{24\ bark} \Delta L dz}{\frac{fm}{4} + \frac{4}{fm}}\ vacil}$$

siendo fm la frecuencia de modulación (se suponde conocida o medible filtrando L) y ΔL la profundidad de modulación del patrón de enmascaramiento frecuencial.

Las modulaciones en amplitud más rápidas (de 15 a 300 Hz) no se perciben como fluctuaciones en el nivel de volumen. En vez de eso, su efecto es una sensación de aspereza, como la voz de Louis Armstrong. La aspereza es un efecto claramente peculiar comparado con otros efectos psicoacústicos. Su unidad es el *asper* y la señal de referencia es un tono puro de 1 kHz y 60 dB SPL, modulado a fm = 70 Hz al 100%.
El cálculo de estas propiedades de fluctuación y aspereza es un proceso complicado debido a que, entre otros factores, la profundidad de enmascaramiento ΔL es un parámetro poco cuantificable.

1.3. Psicoacústica del audio 3D

1.3.1. Introducción al procesado de señales fisiológicas

El proceso de escucha humana se basa en el análisis de las características o pistas para la detección (ILD, ITD, ICC, filtrado direccional, etc…) de las señales de entrada a los dos oídos L y R.
Teorías que datan de fechas tan tempranas como 1882 [7] identifican dos mecanismos básicos como responsables de la localización de una fuente sonora: 1) las diferencias de tiempo interaurales (ITDs) y 2) las diferencias de nivel o intensidad interaurales (ILDs o IIDs).
Una teoría posterior desarrollada por Lord Rayleigh se baso en la combinación de estas dos características para operación a diferentes regímenes de longitud de onda λ [8]. Se demuestra que para λ cortas (f altas, 4 kHz - 20 kHz) la cabeza del oyente se comporta como una sombra o apantallamiento acústico para el oído mas lejano de la fuente sonora, dando lugar a una atenuación del sonido para ese oído (ILDs). Para λ largas (f bajas, 20 Hz - 1 kHz) las dimensiones de la cabeza son muy pequeñas comparadas con λ y la localización se basa en las difrenecias en el tiempo de llegada del sonido a los dos oídos (ITDs).
Estos dos mecanismos formaron la base para la teoría dúplex de localización del sonido. En el rango de f intermedias (1 kHz – 4 kHz) ambos mecanismos ILDs e ITDs están activos, lo que resulta en características que tienden a causar errores de localización. Esto hay que tenerlo en cuenta a la hora de seleccionar el tipo de fuentes sonoras que se van a usar.

Mientras que las diferencias ILD e ITD proveen información direccional en el plano azimutal (horizontal), en el plano mediano (vertical), las ITDs son constantes y la localización se basa en el *filtrado espectral.* La reflexión y difracción de las ondas sonoras debidas al cráneo, torso, espalda y pabellón auricular, combinadas con las resonancias del canal auditivo, forman la base física de las HRTF. El oído externo se puede modelar (en el caso estático) como un sistema lineal e invariante en el tiempo (LTI) que queda completamente caracterizado por esta función de transferencia HRTF. El rol del oído externo es superponer distorsiones lineales relacionadas con los ángulos y distancias al sonido incidente. La información espacial queda entonces codificada en las señales que se reciben en los tímpanos mediante una combinación de filtros complicados.
La magnitud y fase de las HRTFs varía significativamente entre direcciones de sonido pero mucho más importante es la variación entre persona y persona.

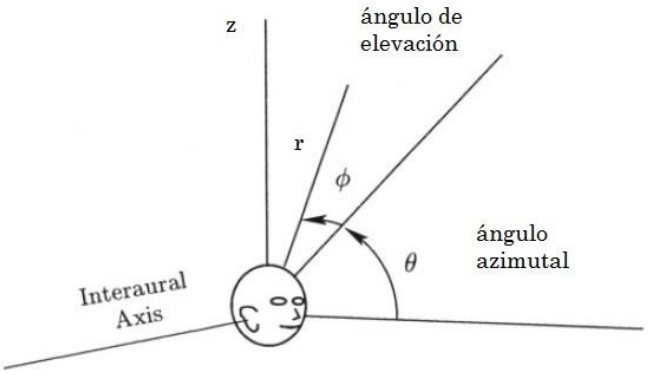
1.3.2. Localización. Características binaurales y HRTFs

La *localización* o el posicionamiento de una fuente se define como la relación entre la posición del *objeto auditivo* y uno o más atributos del *evento sonoro* (fuentes y sus señales, responsables de la percepción del objeto auditivo). El problema de la localización se puede describir como el problema de relacionar la dirección de incidencia del sonido de una fuente con la dirección del *objeto auditivo* correspondiente; es decir, la dirección percibida.

Cuando el ser humano localiza un sonido, básicamente hay tres mecanismos involucrados. Referente a los dos primeros mecanismos, el sonido proveniente de una fuente a la izquierda o derecha del plano mediano (i.e. con un ángulo azimutal $\theta \neq 0º$) llegará antes al oído más cercano a la fuente, creando una *diferencia interaural de tiempo* (*ITD*) entre ambas señales L y R. También tendrán un nivel de presión sonora diferente, mayor para la señal correspondiente al oído mas cercano a la fuente, resultando en una *diferencia interaural de nivel o intensidad* (*ILD o IID*).

Lo usual es usar un sistema de coordenadas esférico como el de la figura siguiente, donde θ denota el *ángulo azimutal u horizontal*, φ el *ángulo de elevación o vertical* y $\vec{\mathbf{r}}$ el vector distancia fuente receptor. El ángulo azimutal se mide desde el eje frontal en el sentido contrario de las agujas del reloj, de manera que la oreja izquierda estará situada en θ entre 80º-100º y la derecha en θ entre 260º-

280º. También se pueden usar ángulos medidos a favor de las agujas del relo, o tomando los ángulos del oído contrario como los mismos ángulos pero de signo opuesto.



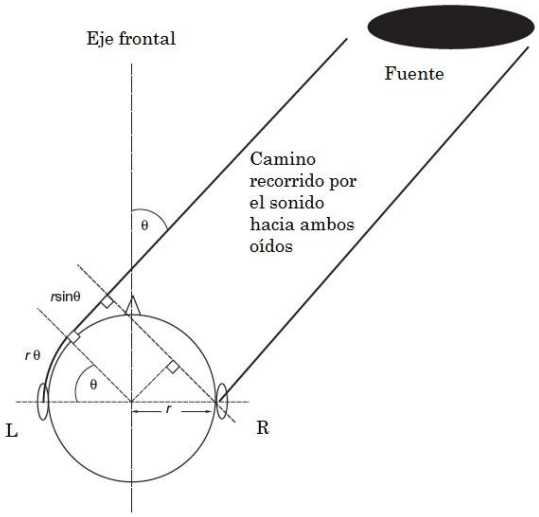
La sensibilidad a ITDs e ILDs en la escucha binaural se ha investigado extensamente y se describe por la lateralización (*lateralization blur*), que se define como el menor cambio en un parámetro que conlleva un desplazamiento perceptible de la fuente. En este caso la sensibilidad a la percepción de desplazamiento de una fuente. El mejor error en la localización se mide enfrente del oyente ($\theta = \varphi = 0º$) y diferentes publicaciones han concluido que la *menor diferencia perceptible jnd (just-noticable-difference)* para la localización es aprox. $\theta = \pm 0.75º$ en el plano horizontal y aprox. $\varphi = \pm 9º$ en el plano mediano.

Ya en el año 1920 se demostró que el error en la lateralización es de aprox. 2 a 10 μs para ITDs. Para las ILDs se estudió en el año 1942 que la jnd para esta característica es de unos 0.6 dB para tonos sinusoidales de frecuencia 2kHz. Con respecto a la percepción de *diferencias interaurales de fase (IPD),* se concluyó en 1958 que, la IPD más pequeña es de 2º a 4º para tonos sinusoidales de frecuencia 150 a 1000 Hz. Se demostró también que una fuente sonora difusa (o incluso dos fuentes diferentes) se puede percibir en 3D simplemente manipulando solo la ITD, o cuando solo la ILD es fuertemente dependiente de la frecuencia.

El tercer mecanismo se activa cuando no existen ITDs ni ILDs obvias. La coloración del espectro de la señal recibida, debida al torso, los hombros y a efectos como reflexiones, apantallamientos o difracciones, a la cabeza y en especial al oído externo (la oreja o pabellón auricular) proveen información útil para la escucha en 3D.

Como se discutirá más adelante, todos estas carácterísticas se pueden codificar en una *función de transferencia relacionada con la cabeza (HRTF)* o su variante temporal la *respuesta al impulso relacionada con la cabeza (HRIR).* Dependiendo de la posición de la fuente, ciertas frecuencias se amplificarán o atenuarán al convolucionar el sonido con la HRIR, consiguiendo la localización virtual en 3D.

La clave para sintetizar audio binaural es un correcto tratamiento para incluir en el sonido todos los efectos relevantes y características necesarias para el efecto 3D deseado. La manera más habitual es tomar las señales de entrada a los oídos como versiones filtradas de la señal original de la fuente. Los filtros que modelan el camino del sonido desde la fuente a los canales auditivos izquierdo y derecho y que dependen de los ángulos de incidencia del sonido proveniente de la fuente son precisamente estas funciones HRIRs/HRTFs.



Un punto de vista más intuitivo, aunque solo aproximadamente valido, de la relación entre el ángulo de incidencia de la fuente θ y las señales de entrada a los oídos es contemplar la diferencia entre los dos caminos a los oídos L y R. Como resultado de la diferencia de longitud de los caminos, hay una diferencia ITD en el tiempo de llegada entre los dos oídos.

$$\mathbf{ITD} = \boldsymbol{\kappa} \cdot \frac{\boldsymbol{\theta} + \mathbf{sin}(\boldsymbol{\theta})}{\mathbf{c}} \; [\mathbf{seg}]$$

es la fórmula más simple para describir esta característica de transmisión física. La variable κ es la distancia entre los dos micrófonos que modelarían las entradas a los oídos y, ya que se ignora el efecto de que el camino se curva debido al cráneo, será necesario para mejor aproximación que:

$$\boldsymbol{\kappa} > \mathbf{r_{head}}$$

El modelo de ITD más simple podría implementarse mediante un elemento de retardo simplemente como:

$$\mathbf{ITD} = \mathbf{ITD} \cdot \mathbf{fs} \; [\mathbf{muestras}]$$

La limitación importante en esta fórmula es que no se tiene en cuenta el efecto de apantallamiento debido al cráneo, i.e., el efecto de la cabeza en las intensidades de las señales de entrada a los oídos no es tenido en cuenta.

El rango fisiológicamente relevante y físicamente posible para cualquier ITD es de ± 1 ms debido a las *jnd* antes comentadas, además de que debido a la geometría de la cabeza el retardo temporal interaural máximo (llamado a veces retardo binaural) es de unos 0.65 ms [1]. Ya que la ITD es la diferencia en el tiempo de llegada del sonido en los dos oídos, otro modelo más complejo y uno de los métodos más extendidos para calcular las ITDs es obtener la *función de correlación cruzada normalizada*:

$$\Psi(\boldsymbol{\tau}) = \sum_{\mathbf{n}} \frac{(\mathbf{HRIR_L}(\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \mathbf{n}) - \boldsymbol{\mu_L}) \cdot (\mathbf{HRIR_R}(\mathbf{r}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \mathbf{n} + \boldsymbol{\tau}) - \boldsymbol{\mu_R})}{\boldsymbol{\sigma_L \sigma_R}}$$

donde $n = 1, 2, \ldots,$ hrir_length, la variable τ representa un intervalo del orden de milisegundos expresado en muestras y (μL, μR) y (σL, σR) representan las medias y desviaciones estándar de las HRIR_L y HRIR_R, respectivamente. De esta función de correlación cruzada Ψ se calcula su valor máximo en un

intervalo de aprox. 1 ms: $\tau = 1$ ms → fs · 1ms = 44.1 kHz · 1 ms ≅ 44 muestras. La ITD de este modelo más complejo queda así definida como:

$$\mathbf{ITD} = \mathbf{abs}(\boldsymbol{\tau}) = \mathbf{abs}(\mathbf{max}(\Psi(\boldsymbol{\tau})))$$

Aunque varios efectos de difracción, reflexión y resonancia debido a cabeza, torso y orejas resultan en ITDs y ILDs no solo dependientes del ángulo de incidencia sino también del tipo de fuente, normalmente se simplifica la situación a unas características interaurales sólo dependientes de la frecuencia, ángulos de incidencia y distancia a la fuente. Más aún, si sólo se consideran ITDs, existe un cono de probabilidades de posiciones de la fuente (*cono de confusión*) con ITDs "virtualmente" iguales para todas las posibles posiciones en ese cono. Es por ello que el sistema auditivo humano usa más características que la ITD para resolver la ambigüedad frontal/atrás. Ejemplos de estas características son movimientos de cabeza, referencias visuales y características espectrales.

Por supuesto, las diferencias temporales ITD pueden expresarse como diferencias de fase para el caso en que las señales provenientes de la fuente sonora sean tonos puros sinusoidales.

1.3.3. Localización del sonido en presencia de reflexiones. El efecto de precedencia y el efecto Haas

El sistema humano registra particularmente las características ITD al comienzo y final de los sonidos (*onsets* y *offsets*) y sobre todo para el contenido de bajas frecuencias. Las ITDs son útiles para monitorización de las diferencias en los onsets y offsets de la envolvente total de las señales para altas frecuencias. Es importante distiniguir entre el retardo binaural resultante de una sola fuente sonora y el retardo medido entre cada oído para dos o mas fuentes similares pero en posiciones diferentes. Es el ultimo caso el que forma el llamado efecto de precedencia.

Hasta ahora solo se han analizado los casos de escucha en condiciones de campo libre. Pero hay otras características relacionadas con "salas", las cuales por ejemplo añaden información de reverberación a la señal y que claramente también influyen en la escucha binaural. Estas características se pueden describir mediante los parámetros acústicos de la sala.

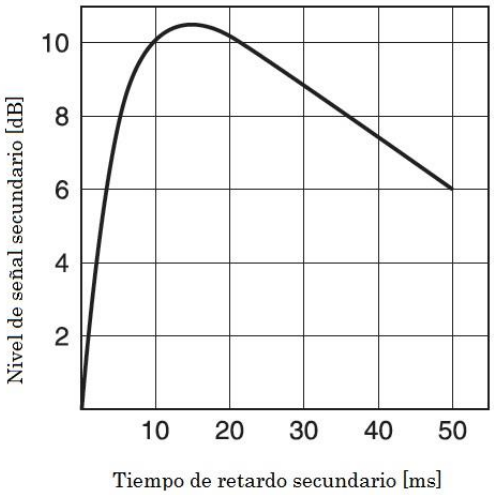
En casi todos los casos, el sonido directo de una fuente radiando en una sala alcanza los oídos del oyente antes que el sonido reflejado por paredes y mobiliario/objetos de la sala. Esto es debido a que el camino indirecto (asociado a una reflexión) es mas largo que el camino directo.
El *efecto de precedencia* o *ley del primer frente de onda* se refiere a un mecanismo de resolución de direcciones sonoras que favorece el primer frente de onda sobre los siguientes (que son debidos a reflexiones o a fuentes retardadas temporalmente). En otras palabras, la percepción direccional de reflexiones que llegan de microsegundos a milisegundos después del sonido directo es anulada y el sonido directo y estas reflexiones se funden en un único objeto auditivo con la dirección del sonido directo. El efecto de precedencia se compone de:

- · localización en suma
- · efecto de precedencia
- · localización del objeto auditivo primigenio y eco

El efecto de precedencia es absolutamente relevante en la reproducción con altavoces, donde se debe tomar más como una norma a seguir a la hora de diseñar la escena sonora. En el caso de altavoces se suele dar la escena de como minimo dos fuentes sonoras en sitios distintos pero emitiendo el mismo sonido en versiones diferentes (retardado, amplificado, filtrado, etc... para proveer información direccional). Ambos oídos escuchan todos los altavoces que haya involucrados y el cerebro tiende por naturaleza a localizar el sonido (la fuente) como proveniente de una direccion determinada por la señal que primero llego de entre todas las emitidas por los altavoces. Este efecto empieza a entrar en acción cuando los retardos entre las fuentes son algo mayores que el retardo binaural, del orden de milisegundos. Sonidos similar que llegan cun una diferencia no superior a 50 ms tienden a ser fusionados y ser percibidos como una sola fuente, de manero que una no se percibe como el eco de la otra. Este tiempo de fusión depende del tipo de fuente, con la característica de que los sonidos impulsivos tienden a separarse antes que sonidos mas

complejos como música o habla. No obstante, el timbre y cualidades espaciales de este sonido "fusionado" podrían ser afectadas positiva o negativamente.

Una forma del efecto de precedencia se conoce como *efecto Haas.* El efecto se identificó originalmente en experimentos realizados para determinar qué cambios se producen en la percepción del sonido cuando se añade un único eco al sonido original. Haas determinó que el eco retardado se podía conseguir hacer bastante mas alto que el sonido anterior antes de que fuera percibido como igualmente alto por el sistema auditivo, como muestra la aproximación de la siguiente figura.



En otras palabras, el efecto Haas predice el nivel de señal relativo requerido por una reflexión retardada (una fuente secundaria) para parecer de igual volumen sonoro al de una fuente primaria anterior.

El efecto Haas tiene implicaciones importantes en las técnicas microfónicas, donde se usan diferencias temporales y de intensidad entre los distintos canales (bien de forma separada o bien de forma conjunta) para crear características espaciales.

1.3.4. Características de distancia

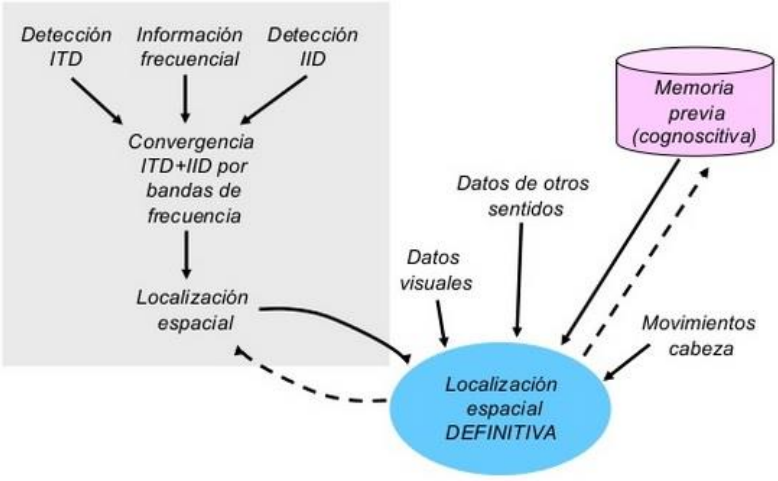
Se asocian varios fenómenos diferentes a la estimación de distancia de una fuente sonora por parte del sistema auditivo humano. El volumen o amplitud del sonido así como el cociente de energía directa a reverberante parecen ser los más efectivos en influenciar la percepción de distancia.
Las características de amplitud derivan del hecho de que el nivel de presión en el campo lejano decrece con el incremento en la distancia entre transmisor y receptor. Por ello, las fuentes cercanas son percibidas como si tuvieran un volumen mayor que fuentes lejanas que emiten la misma cantidad de energía acústica. El cociente entre las intensidades de dos fuentes a distancias r1 y r2 se conoce como *ley cuadrada inversa (inverse square law):*

$$\frac{\mathbf{I1}}{\mathbf{I2}} = \frac{\mathbf{r2^2}}{\mathbf{r1^2}}$$

Luego, el doblar la distancia hace que la intensidad a la altura de los oídos del oyente disminuya 6 dB. El sistema auditivo humano usa este hecho para estimar distancias. No obstante, la familiaridad con un sonido concreto y sus características influyen también en la estimación de distancia.
Considerando el mismo sonido cerca y lejos, uno se da cuenta de que el sonido mas lejano:

a) es mas bajo (por la distancia recorrida)
b) tiene un contenido frecuencial con menos altas frecuencias (debido a la absorción del aire)
c) es mas reverberante (si se considera la situación de un entorno reflectivo)

Más aún, la característica de nivel de volumen es solo válida en entornos anecoicos, puesto que en un entorno reverberante la distribución de sonido depende de las características de reverberación de la sala. Como ejemplo, en una sala reverberante el campo sonoro que vaya más allá de la distancia de reverberación, se percibirá como difuso y será teóricamente independiente de la distancia a la fuente. Esto explica la importancia del cociente de energía directa a reverberante como una característica de distancia.

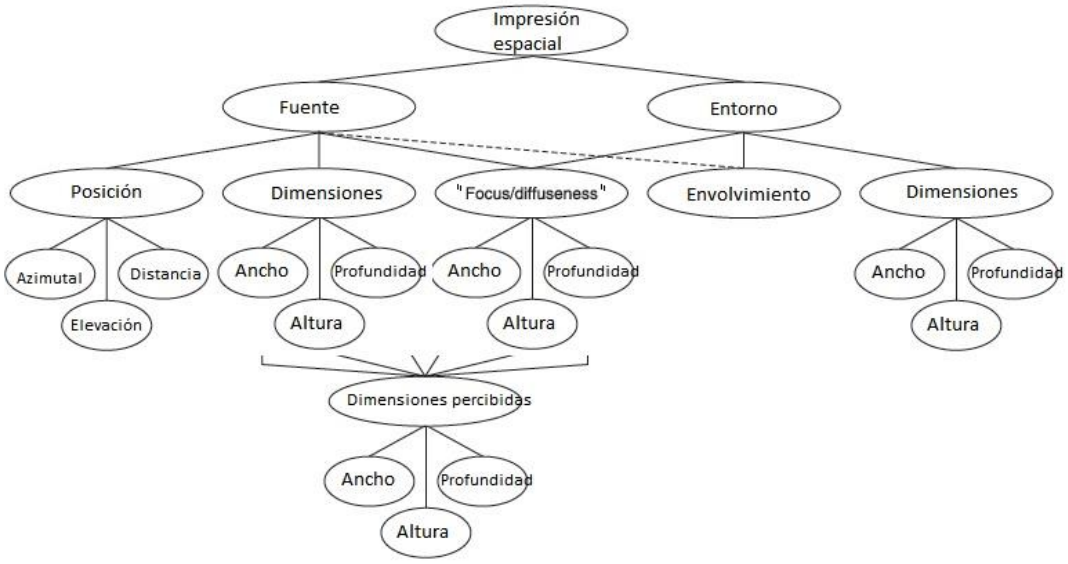


Percepción espacial

1.3.5. Características espectrales

Las características fundamentales usadas por el sistema auditivo para la estimación del Angulo de elevación se suelen clasificar como *monoaurales.* Esto contrasta con las características interaurales o binaurales usadas para la estimación del ángulo azimutal. Se forman características espectrales mediante las reflexiones de ondas sonoras de corta longitud de onda en la parte superior del oyente (torso y hombros) y orejas (pabellones auriculares).
El pabellón auricular, con su característica forma irregular y sus cavidades resonantes, actúa como antena receptora así como filtro acústico. Las ondas sonoras que se reflejan en el pabellón interfieren con el sonido directo que entra al canal auditivo. La interferencia puede ser constructiva a unas frecuencias y destructiva a otras. Esto resulta en picos (o máximos locales) en el espectro a las frecuencias donde ocurren interferencias constructivas, y hendiduras (o mínimos locales) a aquellas frecuencias donde se producen interferencias destructivas. Las frecuencias a las que estos fenómenos suceden dependen de la dirección del sonido. Las frecuencias donde se producen hendiduras en el espectro son de especial interés.
La primera hendidura espectral, conocida como *pinna notch,* parece ser la responsable más directa de la percepción de localización en elevación. La frecuencia a la que aparece esta hendidura varia de 6 kHz a 12 kHz conforme el ángulo de elevación cambia de -40º a +60º. Debido a que las características espectrales para elevación son debidas principalmente a reflexiones para las frecuencias altas, pequeños cambios en la forma de la oreja pueden originar cambios drásticos en la respuesta en frecuencia del oyente. Por ello, las características espectrales varían significativamente entre individuos debido a las diferencias geométricas y del tamaño de las orejas.

Como resumen, en la siguiente figura se muestra una jerarquía propuesta en 1999 para su uso en análisis y experimentos subjetivos relacionados con los atributos espaciales del sonido [7].



1.3.6. Escucha de material binaural con auriculares

La técnica mas sencilla para conseguir material binaural es grabar el sonido insertando micrófonos en miniatura en los canales auditivos de un oyente (o bien modificar unos auriculares para incluir micrófonos en miniatura), o bien usar un maniquin que incorpore micrófonos dentro de sus oídos. Es una técnica simple y que permite efectos muy realistas.

No obstante, hay ciertos problemas con estas técnicas. Primero, la calidad se degrada si la grabación no la hizo la misma que la va a escuchar (o si el maniquí usado no se corresponde con la anatomía del oyente). Segundo, el sonido puede parecer coloreado, las direcciones percibidas de las fuentes virtuales pueden invertirse con respecto al plano frontal y todo el sonido puede parecer como proveniente de dentro de la cabeza. Para solventar parcialmente estos problemas, la grabación y la reproducción deben ser ecualizadas cuidadosamente para obtener una respuesta en frecuencia plana del tímpano del objeto de grabación al tímpano del objeto de escucha u oyente. Tal ecualización requiere medidas muy cuidades y no se profundiza en más detalles en este proyecto.
No obstante, en este proyecto sí se han ecualizado los auriculares. Para ello, el autor contactó con la empresa Sonarworks® para hacer una ecualización acústica, Estos auriculares se usaron para la escucha del material binaural sintetizado durante el desarrollo del proyecto. La sección x.x.x. lo presenta de manera un poco más extensa.

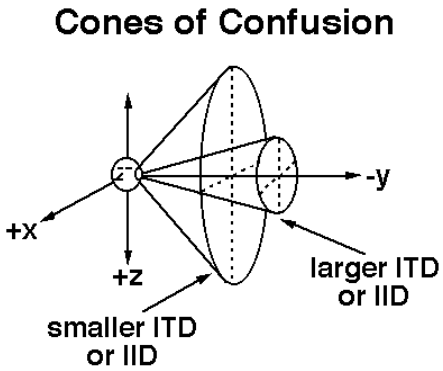
Otro problema adicional en la escucha de material binaural es el hecho de que los humanos usamos también características dinámicas para localizar sonidos. El uso de auriculares bloquea esta dinámica, ya que las características binaurales no cambian al mover el oyente su cabeza.

Otro problema es el hecho de que cada humano es único. Cada oyente tendrá orejas con tamaños y formas diferentes, diferente tamaño y forma de la cabeza. Cuando una grabación binaural realizada mediante un individuo es escuchada por otro individuo diferente, estos problemas aparecen naturalmente como en el material no ecualizado y son difíciles de mitigar.

1.4. Audio binaural

El audio binaural sea quizás la manera más eficiente de implementar audio 3D. Ya que percibimos el sonido 3D con nuestros dos oídos, toda la información relevante a la espacialidad del sonido esta contenido en las dos señales que llegan a nuestros oídos; de hecho, nuestra percepción es el resultado de interpretar la presión que recibimos en los dos tímpanos. Por ello, la grabación de estas señales con micrófonos en miniatura y la reproducción mediante auriculares (para no tener en cuenta funciones de transferencia acústicas) debería ser suficiente para recrear experiencias aurales del mundo real.

La percepción de direccionalidad de sonido en los seres humano se basa en las características relacionadas con las diferencias o similitudes entre las señales que llegan a cada oído, que nuestro cerebro decodifica e interpreta.
A finales del siglo XIX, Lord Rayleigh fue el primero que identificó dos mecanismos referentes a la localización del sonido: características temporales ITDs (que también son interpretadas como diferencias de fase), que se usan para determinar la dirección de llegada a frecuencias por debajo de los 700 Hz y características de intensidad ILDs (relacionadas con la energía de la señal) que dominan por encima de los 1.5 kHz [8].
En las frecuencias bajas del espectro audible, las longitudes de onda del sonido son grandes comparadas al tamaño de la cabeza; por ello, el sonido viaja casi sin ser afectado por el cráneo y llega a los dos oídos independientemente de la dirección de proveniencia del sonido. En estas frecuencias, a menos que una fuente este muy cerca de una oreja, la pequeña distancia entre las orejas no causa ninguna atenuación significativa de la presion sonora. A frecuencias bajas, la única diferencia entre las señales de cada oído es una diferencia de fase, relacionada con la diferencia en el tiempo de llegada del sonido. De acuerdo con [10] el problema de determinar qué oído tiene la fase dominante se puede determinar de manera exacta, pero para frecuencias por debajo de 700 Hz.
La diferencia interaural de tiempo ITD es usada por el sistema auditivo para detectar la proveniencia de un sonido con frecuencias inferiores a 1.5 kHz. La otra característica de localización importante, la diferencia interaural de intensidad o nivel ILD o IID lateraliza el sonido hacia el oído que reciba la señal de mayor intensidad. Esta ILD debería en principio funcionar para todo el rango de frecuencias; no obstante, el poco apantallamiento debido a la cabeza a bajas frecuencias no es suficiente para causar una diferencia de nivel significante para bajas frecuencias. Las ITDs y las ILDs son las características principales del sistema auditivo para el posicionamiento de sonidos en el espacio 3D.



En este punto hay que hacer una aclaración: para una dirección de proveniencia del sonido dada, en realidad existe todo un conjunto de direcciones que resultaría en las mismas ITDs e ILDs. Esto corresponde a un cono, obtenido de rotar la línea que conecta la fuente con la cabeza a lo largo del eje interaural (que une ambas orejas). Estos conos se llaman *conos de confusión.* En la figura siguiente se muestra un ejemplo. Para características de valor menor, el cono es más ancho y corto, mientras que para características de magnitud mayor el cono es más alargado y estrecho. Fuentes sonoras posicionadas en la superficie del mismo cono generaran las mismas ITDs e ILDs.

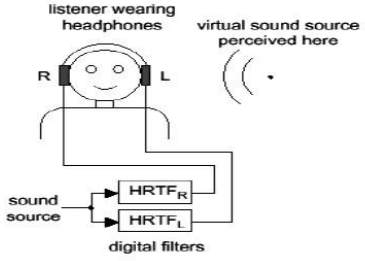
El ejemplo mas claro que refleja esta ambigüedad en las características interaurales es el de un sonido proveniente de frente comparado con otro proveniente de atrás.
Afortunadamente, nuestro sistema auditivo se basa en otros dos mecanismos adicionales para resolver esta ambigüedad y discriminar diferentes dirección para fuentes en el mismo cono de confusión: el primero es el rotar la cabeza (lo que causa una variación de la ITD e ILD que permite discriminar la dirección) y el segundo es el contenido espectral del sonido (que se modifica dependiendo de la dirección de incidencia, pues el oído externo o pabellón auricular o pinna introduce un filtrado dependiente de la dirección de llegada del sonido) [11].

El concepto básico detrás del audio binaural 3D es que si uno mide las presiones acústicas producidas por un campo sonoro en la posición de los oídos del oyente y posteriormente reproduce exactamente las mismas señales directamente en los oídos del oyente, la información espacial original será reconstruida por el cerebro. Las grabaciones binaurales se llevan a cabo con cabezas maniquí con orejas y canales auditivos [12]. Se insertan dos micrófonos que miden la presion sonora en cada oído del maniquí, con lo que se consiguen las señales que percibiría un humano.
La reproducción de audio binaural necesita usar auriculares para enviar a cada oído exactamente la señal grabada correspondiente. Esta técnica consigue efectos 3D realistas.

Vale la pena mencionar que mientras que la escucha de material convencional monofónico o estereofónico mediante auriculares recrea un escenario que se encuentra "dentro" de la cabeza del oyente, el uso de las técnicas binaurales reproduce el sonido, en mayor o menor medida, como proveniente de fuera de la cabeza. Este efecto se conoce como *externalización*.

En términos físicos, las señales que llegan a cada oído cuando una fuente emite un sonido desde cierta posición pueden expresarse como una **convolución** entre el sonido emitido y dos funciones de transferencia entre la posición del oyente HRTF y cada oreja (dejando de lado los efectos de la sala, que también influirá con su propia funcion de transferencia RTF).
Las funciones de transferencia relacionadas con la cabeza HRTFs (Head-Related Transfer Functions) dependen de la posición de la fuente y la forma particular del pabellón auricular usado durante la grabación de dichas funciones. La distancia no suele influir debido a que se graban en entornos anecoicos a una distancia de 1 metro. No obstante, recientemente se ha desarrollado otra base de datos que sí tiene en cuenta distancias de 0 a 2 metros [13]. Las bases de datos de HRTFs coleccionan las grabaciones de las respuestas al impulso, grabadas muestreando una esfera alrededor del oyente con la fuente moviéndose por esta esfera. Normalmente, se usan aproximaciones de campo lejano y la dependencia con la distancia se toma simplemente como 1/r a partir de cierta distancia.
Con las HRTFs se puede sintetizar material de audio binaural mediante convolución: una vez que se ha escogido la posición y tipo de fuente sonora, las señales binaurales izquierda y derecha se obtienen convolucionando la señal de la fuente con las HRIRs L y R que correspondan a la posición de la fuente. Denotando s(n) como la señal de audio a aurlizar:

$$\mathbf{y_L(n)} = \mathbf{s(n)} * \mathrm{HRIR_L}$$
$$\mathbf{y_R(n)} = \mathbf{s(n)} * \mathrm{HRIR_R}$$

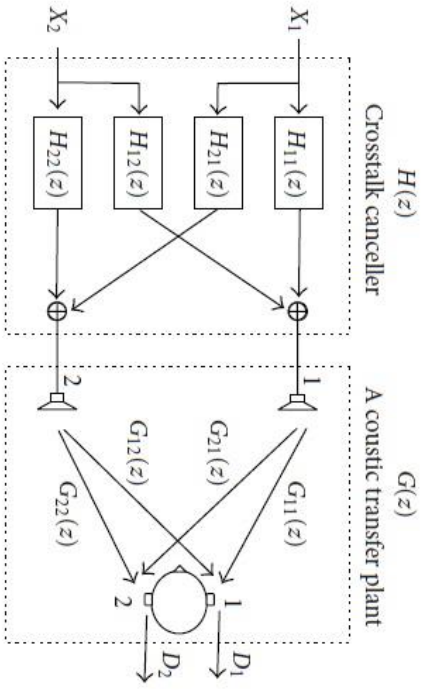


Una herramienta adicional del sistema de audio 3D es realizar seguimiento de movimientos o *head-tracking* , que resulta de un valor incalculable para detectar la orientación del oyente y adaptar la escena (los filtros HRTF) a los movimientos, solventando los problemas de incertidumbre relacionados con los conos de confusión, sobre todo de las ambigüedades frontal/atrás [10, 14]. La reproducción de material binaural también se puede hacer usando altavoces, pero en este caso las señales yL(n) e yR(n) no van a llegar intactas a los oídos del oyente. Las nuevas señales, eL(n) y eR(n), que llegan al oyente contienen diafonía del otro canal. Ya que para conseguir el mejor efecto binaural es necesario que:

$$\mathbf{e_L(n)} = \mathbf{y_L(n)}$$
$$\mathbf{e_R(n)} = \mathbf{y_R(n)}$$

se hace necesario implementar en el sistema de audio 3D un *filtro de cancelación de Crosstalk* para no colorear el espectro del audio sintetizado. Esta técnica, llamada *reproducción transaural o Ambiofónica,* se presenta extensamente en el capítulo 2 y se pueden encontrar los detalles de un sistema que implementa también head-tracking junto con cancelación de Crosstalk y que se usó como base para empezar el proyecto, en [9].

La reproducción binaural es ciertamente efectiva en aplicaciones para dispositivos móviles, donde el oyente casi siempre usa auriculares. El audio binaural o audio 3D, que consiste en dos canales como el estéreo convencional, puede ser distribuido como cualquier otro tipo de audio. Las desventajas más prominentes de la tecnología binaural son la necesidad de usar auriculares (los sistemas transaurales no están muy extendidos y demandan demasiado debido al pequeño punto de escucha optimo o *sweet spot*) además de que para una completa inmersión 3D se necesitan HRTFs individualizadas, algo casi imposible actualmente.



1.5. Escucha espacial (Spatial Hearing)

En situaciónes de escucha naturales, en el mundo real, el humano es capaz de estimar la posición de los objetos que emiten ondas acústicas. La precisión de esta estimación es, para fuentes situadas en el plano horizontal que contiene el eje interaural (los oídos del humano), bastante precisa como para haberse convertido en un hábito de la vida diaria.
Un ejemplo típico para explicar esta habilidad de localizar fuentes en el espacio 3D es fijarse cuando uno cruza una calle mirando hacia la derecha. Escuchando el sonido de un coche que viniese por la izquierda, sólo fiándose del sonido percibido, permite posicionar aproximadamente el coche, sin verlo con los ojos.

Mediante experimentos psicoacústicos, la ciencia ha relevado importantes características usadas por el sistema auditivo humano para localizar fuentes en el espacio. Es sabido que la habilidad de localización de fuentes en el plano horizontal que contiene los oídos del humano se debe a diferencias entre las ondas acústicas que llegan a cada oído. La localización o posicionamiento de fuentes que no están en el plano horizontal está fuertemente influenciada por el filtrado acústico que ocurre debido a la reflexión de ondas de corta longitud de onda en los pabellones auriculares (orejas), los hombros y el torso.

Estas y otras características de localización se detallan en el capítulo 2. El efecto de reverberación de una sala en la escucha espacial natural también es analizado en el capitulo 2.

1.6. Escucha espacial virtual

Todos los parámetros físicos asociados con las características necesarias para la localización que se encuentran en la escucha espacial natural son codificados en un par de funciones de transmisión acústica desde la fuente sonora a los dos canales auditivos (tímpanos) del oyente. Asuma el lector que estas dos funciones, usualmente llamadas un par de *funciones de transferencia relacionadas con la cabeza (Head-Related Transfer Functions HRTF)* puedan medirse y grabarse digitalmente en forma de filtros digitales. El filtrado de una señal monofónica que no contenga ningún tipo de información direccional, mediante este par de HRTFs, resulta en un par de señales eléctricas que contienen la información necesaria para la localización espacial. Al reproducir este par de señales eléctricas (p.ej. mediante auriculares) a la altura de los canales auditivos del oyente, se reconstruyen las características binaurales que permiten al oyente percibir el sonido como proveniente de la posición en la que se encontraba la fuente en el momento de la medición de ese par de HRTFs. Uno de los primeros artículos publicados relativos a la posición del sonido es [9]. Esto es en esencia el principio de los sistemas de audio 3D. En este contexto, la fuente sonora percibida se denomina *fuente virtual o fantasma.*

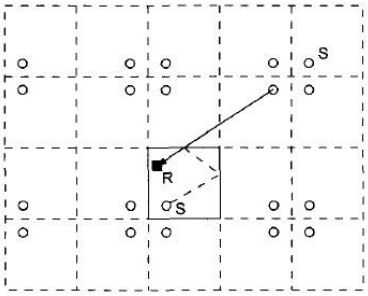
1.7. Reverberación y Acústica de salas. Impresión de espacialidad y parámetros acústicos de una sala.

1.7.1. Conceptos básicos de la Acústica de salas

Si en una sala cualquiera se dispara una pistola, apagándose repentinamente la fuente de sonido, el sonido del disparo se sigue escuchando un pequeño tiempo después. Esto es lo que se conoce como la reverberación del sonido en una sala. La duración de este eco (la reverberación) depende del tamaño de la sala y sus características.

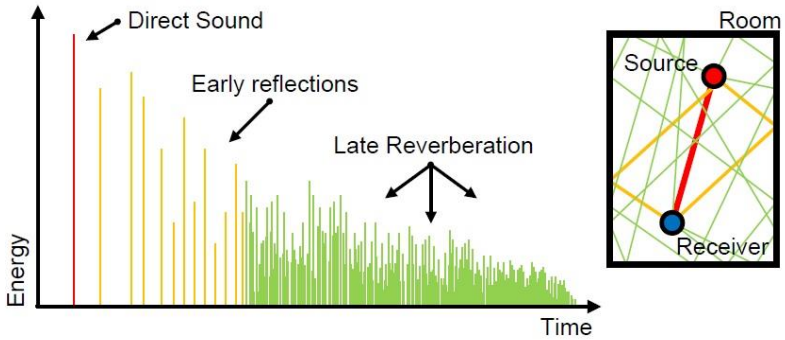
Cada reflexión del sonido en una pared plana completamente reflectante se puede ver como una fuente imagen. Para representar fuentes de orden superior, las fuentes imágenes deben reflejarse nuevamente en las paredes, produciéndose nuevas fuentes (imágenes de las imágenes)

Para una sala rectangular se obtiene una constelación como la ilustrada a continuación.



La extensión a tres dimensiones es directa. El campo sonoro de la sala se puede representar como la suma del sonido de la fuente mas el que proviene de todas las imágenes. La diferencia de tiempos entre las señales que van llegando al receptor está dada por su recorrido desde la correspondiente fuente al punto del receptor (oyente o micrófono).

Cuando la fuente genera un impulso, conforme se mide la amplitud con el paso del tiempo se obtiene un diagrama, conocido como *ecograma.*



La respuesta que caracteriza fundamentalmente cualquier acústica de una sala se denominda *respuesta al impulso de la sala (Room Impulse Response RIR)*. La RIR describe la acústica de la sala para una fuente sonora estática, fija, escuchada por un receptor también fijo. Los ingenieros dividen normalmente la RIR en tres partes: *el sonido directo, las reflexiones tempranas o primeras reflexiones y la reverberación tardía.* El ecograma de la figura muestra la subdivisión de una RIR en concordancia con la percepción y el sistema auditivo humanos. Mientras que el sonido directo y las primeras reflexiones contienen

información sobre la posición, nivel y ancho de la fuente sonora, la reverberación tardia se relaciona con la respuesta de la sala en cuestión.
La reverberación de una sala se compone de las reflexiones propias de un *entorno de no campo libre*, copias retardadas y atenuadas del sonido directo. El contenido frecuencial de cada reflexión es modificado debido a la directividad de la fuente y también debido al material de absorción de las superficies de la sala.

Sonido directo Es la parte del sonido que primero llega al receptor u oyente. El impulso que llega primero y con mas intensidad es evaluado por el sistema auditivo humano para localizar la fuente (también conocido como *efecto de precedencia* [11]). El sonido directo se retarda una distancia r entre la fuente sonora y el oyente y solo es atenuado por el gas del medio (el gas aire). Sólo la llegada de las primeras reflexiones dependen principalmente de la distancia fuente receptor y su ubicación en la sala.

Reflexiones tempranas o primeras reflexiones Las reflexiones de menor orden son añadidas por el sistema auditivo al sonido directo y no son percibidas por separado, incluso si su nivel fuera 10 dB mayor (efecto Haas [11]). Aunque las direcciones de cada reflexión individual no son percibidas debido al efecto de precedencia, sí contribuyen a la percepción del tamaño, ancho y volumen sonoro de la fuente, así como a la percepción del tamaño de la sala.

Reverberación tardía Se considera a partir de un cierto tiempo, que a veces se denomina *tiempo de mixing o mezcla,* cuando la densidad de las reflexiones es tan alta que las reflexiones individuales no se distinguirían en la RIR. Normalmente este tiempo esta entre los 50-80 ms.
Este *campo sonoro difuso* [17] forma una reverberación tardía que es aproximadamente independiente de la posición del oyente, ya que el sistema auditiva empieza a integrar la energía de la señal en un intervalo de tiempo determinado y en un intervalo de angulos. La reverberación es uno de los atributos más importantes y sobresalientes de una sala, pues ciertas características como el tamaño y volumen de la sala están en directa relación con estas reverberaciones tardías, dando a la sala su sonido característico individual.

Al emitir un impulso corto en una sala, el número de impulsos reflejados N que llegan en un intervalo temporal t ($t = 0$ corresponde al tiempo en que llega el sonido directo de la fuente) se puede estimar por medio del numero de fuentes ubicadas dentro de un circulo (2D) o esfera (3D) de radio $r = c \cdot t$, siendo c la velocidad del sonido. N es más o menos la razón entre el volumen de la esfera de radio $c \cdot t$ y el volumen de la sala original V:

$$\mathbf{N = \frac{V^{esfera}}{V} = \frac{4\pi}{3}\frac{(c \cdot t)^3}{V}}$$

Como ejemplo, en una sala con $V = 8\text{m} \times 6\text{m} \times 2.5\text{m} = 120\text{m}^3$, se tienen aprox. $N = 180000$ reflexiones en los primeros 0.5 segundos. Usando la ecuación anterior se puede calcular la densidad media de impulsos recibidos:

$$\mathbf{\frac{\Delta N}{\Delta t} \approx \frac{dN}{dt} = 4\pi c\frac{(c \cdot t)^2}{V}}$$

Esta densidad aumenta continuamente ya que la distancia a las fuentes imágenes se hace cada vez mas grande conforme avanza el tiempo. Por el contrario y como es lógico debido a los fénomenos de absorción del sonido, la amplitud de estos impulsos se va haciendo cada vez más pequeña, con una energía proporcional a $(c \cdot t)^2$:

$$\mathbf{E_{inc} \approx (c \cdot t)^2}$$

La media temporal de la energía E para un punto determinado de la sala, se obtiene como el producto del numero de impulsos por unidad de tiempo y su energía:

$$\mathbf{E = E_{inc} \cdot \frac{\Delta N}{\Delta t} = cte}.$$

La densidad de energía E alcanzara un valor constante transcurrido un breve periodo de tiempo. Y también ocurre esto para la distribución espacial de la energia. Se espera entonces una distribución de energía constante tanto en tiempo como en espacio. Debido a la absorción que experimentan las ondas por el amortiguamiento a lo largo del camino de propagación y por las paredes y elementos o superficies de la sala, en salas muy reverberantes ocurre que a distancias no muy cortas desde la fuente y también más largas, el nivel sonoro no cambia al cambiar la posición, sin embargo la reverberación siempre decae en el tiempo.

En la teoría de la Acústica de Salas es importante caracterizar las pérdidas por absorción. Con esto no se elimina la posibilidad de que hubiera un cambo sonoro difuso (con distribución espacial constante). Considerando el modelo de muchas fuentes imágenes, en el campo difuso las ondas inciden omnidireccionalmente (desde todos los ángulos) en cada punto del espacio y lo hacen con intensidad aproximadamente constante. Por tanto, el concepto de campo difuso se extiende a un campo sonoro donde tanto el nivel de presión sonora en los distintos puntos del espacio como las direcciones de incidencia tienen una distribución constante.

Un campo sonoro cuasi difuso ocurre en situaciones donde hay un amortiguamiento pequeño de las ondas. En el caso real de una distribución irregular (por ejemplo alta absorción en las paredes pero no en techo y suelo) no se cumplirá la suposición de incidencia en todas direcciones y podría producirse el denominado *eco fluctuante (flutter echo).* Al asumir una distribución estadísticas de absorción regular la ventaja que se obtiene es que se pueden estimar los campos de onda sonoros en las salas, lo cual se puede lograr de manera exacta solamente mediante procedimientos de calculo como los métodos de elementos finitos FEM o BEM, mucho mas engorrosos.

Cuando se consideran geometrías rectangulares sencillas (paredes paralelas) se pueden extraer algunas conclusiones usando la teoría de ondas, por lo menos en el caso de reflexión total en todas las paredes. Por el contrario, si se tienen muchas superficies absorbentes distribuidas espacialmente y elementos grandes que pueden introducir difracción y dispersión en las ondas sonoras, debido a la complejidad de la configuración no es posible realizar cálculos precisos usando la teoría de ondas. Es por esto que se suelen considerar para el análisis modelos de salas muy simplificados. Además las simplificaciones (por ejemplo asumir cierta distribución espacial) solo tienen sentido estadísticamente. Al hablar de un nivel de sonido constante en tiempo y espacio en un campo difuso, en realidad uno se refiere a la media espacial obtenida mediante varias mediciones en muchos puntos diferentes de la sala.

Antes de asumir cualquier distribución regular en una sala, es necesario considerar la teoría de ondas. Para una sala rectangular sin perdidas, de longitud, anchura y altura

$$\overrightarrow{\mathbf{rm}} = \left[\mathbf{l_x}\,,\mathbf{l_y}\,,\mathbf{l_z}\right],$$

con superficie total

$$\mathbf{S} = \left(\mathbf{S_{floor}} + \mathbf{S_{ceiling}}\right) + \mathbf{S_{walls}} = \left(\mathbf{2\,l_x l_y}\right) + \mathbf{2\,l_x l_z} + \mathbf{2\,l_y l_z}$$

y volumen

$$\mathbf{V} = \mathbf{l_x l_y l_z}\,,$$

al aplicar el método de separación de variables y el principio de superposición al problema de resolver la ecuación de onda en un recinto rectangular y aplicando las condiciones de contorno adecuadas se tiene una expresión para la presión sonora en forma de triple sumatoria [17]:

$$\mathbf{p(x\,,y\,,z)} = \sum_{\mathbf{n_x,n_y,n_z}} \mathbf{p_{n_x n_y n_z}} \cos\left(\frac{\mathbf{n_x \pi x}}{\mathbf{l_x}}\right) \cos\left(\frac{\mathbf{n_y \pi y}}{\mathbf{l_y}}\right) \cos\left(\frac{\mathbf{n_z \pi z}}{\mathbf{l_z}}\right)$$

ya que en todas las paredes deben existir máximos en la presión sonora. Existen los llamados modos espaciales, los cuales corresponden cada uno a una frecuencia de resonancia

$$\boldsymbol{\omega} = \mathbf{c}\cdot\mathbf{k} = \mathbf{c}\cdot\sqrt{\left(\frac{\mathbf{n_x \pi}}{\mathbf{l_x}}\right)^2 + \left(\frac{\mathbf{n_y \pi}}{\mathbf{l_y}}\right)^2 + \left(\frac{\mathbf{n_z \pi}}{\mathbf{l_z}}\right)^2}$$

Los modos normales de una sala pueden representarse mediande una malla tridimensional en las variables normalizadas $\left(\frac{c}{2l_x}, \frac{c}{2l_y}, \frac{c}{2l_z}\right)$. El número de frecuencias de resonancia existentes hasta una cierta frecuencia M se obtiene como el volumen de un octavo de esfera de radio la frecuencia máxima f divido por el volumen de un paralelepípedo:

$$\mathbf{M} = \frac{volumen\ de\ \frac{1}{8}\ esfera}{\frac{\mathbf{c^3}}{\mathbf{8\,V}}} = \frac{\frac{\boldsymbol{\pi}}{\mathbf{6}}\mathbf{f^3}}{\frac{\mathbf{c^3}}{\mathbf{8\cdot V}}} = \frac{\mathbf{4\pi}}{\mathbf{3}}\cdot\mathbf{V}\cdot\left(\frac{\mathbf{f}}{\mathbf{c}}\right)^{\mathbf{3}}$$

Como ejemplo, en el caso de un volumen $V = 120\ m^3$ hasta la frecuencia 554 Hz existen más de 2000 resonancias. La densidad de frecuencias de resonancia o propias se define

$$\frac{\Delta\mathbf{M}}{\Delta\mathbf{f}} \approx \frac{\mathbf{dM}}{\mathbf{df}} = \frac{\mathbf{4\pi}}{\mathbf{c}}\cdot\mathbf{V}\cdot\left(\frac{\mathbf{f}}{\mathbf{c}}\right)^{\mathbf{2}}$$

Para el ejemplo de $V = 120\ m^3$ se tendría para $f = 1000$ Hz una densidad $\Delta M/\Delta f = 38$. Esto es decir que en una banda de frecuencia de ancho 1 Hz en torno a $f = 1000$ Hz existen aproximadamente 38 resonancias.
Un campo difuso, que por definición implica un nivel de sonido independiente del espacio, solo se puede conseguir con señales de banda ancha. En una sala con poca absrocion un tono puro conduce a ondas estacionarias con marcados máximos y minimos. Solamente al excitar muchas de estas ondas estacionarias al mismo tiempo se puede conseguir un campo sonoro difuso con distribución independiente del espacio [17].

Normalmente, para mediciones de acústica de salas, se exige una relación entre ancho de banda y el volumen de la sala V tal que al excitar con ruido en bandas de octava o de tercios se tenga una densidad de aprox. 1 Hz:

$$\boldsymbol{\Delta M/\Delta f = 1/Hz\,.}$$

Usando la expresión

$$\frac{\Delta\mathbf{M}}{\Delta\mathbf{f}} \approx \frac{\mathbf{dM}}{\mathbf{df}} = \frac{\mathbf{4\pi\,V}}{\mathbf{c}}\cdot\left(\frac{\mathbf{f}}{\mathbf{c}}\right)^{\mathbf{2}}$$

con la ecuación anterior, se consigue obtener el rango de frecuencias permitido, para el cual

$$\mathbf{f} \geq \sqrt{\left(\frac{\mathbf{1}}{\mathbf{Hz}}\right)\frac{\mathbf{c^3}}{\mathbf{4\pi V}}} \approx \frac{\mathbf{1800\ Hz}}{\sqrt{\boldsymbol{V\,m^3}}}$$

Para el ejemplo de una sala rectangular con volumen $V = 120\ m^3$ se podrá empezar a medir correctamente solamente a partir de los 160 Hz.

Tiempo de reverberación de la sala (TR) Sin demostración se da aquí que, al considerar el estado estacionario para el cual la potencia radiada por la fuente es igual a la potencia absorbida por las superficies de la sala, el tiempo de reverberación TR de una sala se puede obtener usando la ***Fórmula de Sabine*** [19] desarrollada por W. C. Sabine a finales del siglo XIX:

$$\frac{\mathbf{13.8\cdot V}}{\mathbf{c\cdot TR}} = \frac{\mathbf{A}}{\mathbf{4}}$$

La cantidad $\mathbf{A}$ se define como el *área de absorción equivalente*, dada por la suma de todas las superficies con material absorbente $\mathbf{S_i}$ ponderadas por sus coeficientes de absorción correspondientes $\boldsymbol{\alpha_i}$:

$$\mathbf{A} = \mathbf{S}\,\bar{\boldsymbol{\alpha}} = \sum_{\mathbf{i}} \boldsymbol{\alpha_i}\cdot\mathbf{S_i}$$

Resulta útil deducir la expresión del coeficiente de absorción medio a partir de esta ecuación:

$$\bar{\boldsymbol{\alpha}} = \frac{\sum_{\mathbf{i}} \boldsymbol{\alpha_i}\cdot\mathbf{S_i}}{\mathbf{S}}$$

El tiempo de reverberación es, en suma, el tiempo que hay entre que se interrumpe la recepción de un sonido (normalmente impulsivo) y la recepción de las reflexiones de ese sonido.

Para incidencia normal, el coeficiente de absorción de un material poroso se puede obtener como [28]:

$$\boldsymbol{\alpha = \frac{4\,\sigma}{(1+\sigma)^2}}$$

donde σ representa el factor de porosidad del material en cuestión. La frecuencia a la que un trozo de material absorbente es mas efectivo se puede calcular aproximadamente como:

$$\boldsymbol{f_0 = \frac{c}{4d}}$$

siendo c la velocidad del sonido en m/s y d el espesor del material en metros. En [28] se puede encontrar información acerca de distintos materiales de absorción y difracción del sonido así como distintos tipos de resonadores para mejorar la escucha en salas o estudios de grabación pequeños.

En la práctica se considera que las reflexiones finalizan cuando el nivel sonoro que tienen es de una millonésima parte de su valor original, i.e. una atenuación de 60 dB. Por ello que a veces también se use la notación TR_{60} para llamar al tiempo de reverberación. En la mayoría de situaciónes se usa la fórmula adimensional válida para el aire:

$$\boxed{\boldsymbol{TR = 0.161\frac{V}{S\,\bar{\alpha}} = 0.161\frac{V\,[m^3]}{A\,[m^2]}\ [segundos]}}$$

Para recintos que tienen un fin musical es conveniente tener un TR de entre aprox. 0.5 a 2 segundos, mientras que para recintos con fines conferenciales se suele prefererir un TR menor, de entre aprox. 0.5 a 1 segundos. Los coeficientes de absorción se pueden medir u obtener de tablas estandarizadas para los materiales que se usen.

En la tabla de la página siguiente se listan las formulas preferidas para calcular el tiempo de reverberación aproximado de una sala para diferentes situaciones [18, 23]. La norma EN ISO 134 establece un procedimiento de medición para los coeficientes de absorción.

Todas las fórmulas suponen que el TR:

- · no varía dentro de la sala
- · no depende de la posición de la fuente de sonido
- · no depende de la forma ni la geometría de la sala (se supone una sala rectangular)
- · no depende de la posición de los diferentes materiales absorbentes en la sala

En muchas aplicaciones prácticas, asumir un campo difuso para poder aplicar la teoría de Sabine no es coherente con la distribución de absorción existente en la sala a analizar [23]. El campo sonoro será, en general, lo suficientemente difuso si no hay grandes diferencias en las dimensiones de la sala, las paredes no son paralelas, el material absorbente esta uniformemente distribuido. Como se puede comprobar, en la práctica casi ninguna de estas condiciones se cumple. No hay que olvidar que a medida que la abosrcion aumenta la condición de campo difuso va perdiendo significado, con lo que el tiempo de reverberación también. Es por esto que la formula del TR de Sabine o de Eyring no pueden ser aplicadas y esperar obtener resultados precisos. En 1959 D. Fitzroy publicó una investigación dedicada al problema de un calculo mas preciso del TR en una sala con absorción no uniformemente distribuida y que arrojaba resultadas mas cercanos a los TR medidos en salas reales [21]. Fitzroy postula que cuando uno considera las posibilidades que la Física Ácustica brinda, en comparación con el punto de vista geométrico, en general el campo sonoro tiende a establecerse en un patrón de oscilación simultanea a lo largo de los tres ejes principales de una sala rectangular, el vertical, el transversal y el longitudinal. El TR parece derivar o tener relación con los tres tiempos de decaimiento a lo largo de estos tres ejes principales.

Aunque los resultados presentador por Fitzroy fueron notables, se obvió esta formula por mas de 40 años. A principios del siglo XX en [23] se vuelve a analizar y a corregir esta fórmula, al igual que pasó con la fórmula de Eyring, descubierta en 1930 [20] que fue corregida mas de 40 años después en 1976 por Kuttruff [22] y al igual que la fórmula de Sabine que fue entendida por Eyring como una formula para salas “vivas” y que Eyring corrigió para salas “muertas” [22] también unos 40 años después del descubrimiento de Sabine [19].

La condición de "sala viva" (live room) se usa cuando una sala tiene absorción total en relación con la superficie total de la sala del orden del 2%. Por otro lado, *la condición de "sala muerta" (dead room)* aplica cuando la absorción total sea de aprox. 95%.

	Unidades MKS S = área total en m^2 V = volumen total en m^3	**Unidades inglesas** S = área total en ft^2 V = volumen total en ft^3
Sabine Da la mejor correspondencia con los coeficientes de absorción publicados con $\bar{\alpha} < 0.2$	$$\mathbf{TR = 0.16 \cdot \frac{V}{S \cdot \bar{\alpha}}}$$	$$\mathbf{TR = 0.16 \cdot \frac{V}{S \cdot \bar{\alpha}}}$$
Eyring Formula preferida para salas con "buen" comportamiento con $\bar{\alpha} > 0.2$	$$\mathbf{TR = 0.16 \cdot \frac{V}{-S \cdot \ln(1-\bar{\alpha})}}$$	$$\mathbf{TR = 0.49 \cdot \frac{V}{-S \cdot \ln(1-\bar{\alpha})}}$$
Millington Formula preferida para salas con todos los coef's de absorción menores que la unidad	$$\mathbf{TR = 0.16 \cdot \frac{V}{\sum_i S_i \cdot \ln\left(\frac{1}{1-\alpha_i}\right)}}$$	$$\mathbf{TR = 0.49 \cdot \frac{V}{\sum_i S_i \cdot \ln\left(\frac{1}{1-\alpha_i}\right)}}$$
Eyring-Kuttruff Formula preferida para salas en donde las i – 1 superficies tienen aprox. mismo coef. de reflexión $\rho = 1 - \alpha$ mientras que la superficie i-ésima tiene un ρ diferente (p.ej. el suelo)	$$\mathbf{TR = 0.16 \cdot \frac{V}{-S \cdot \ln(1-\bar{\alpha}) + \Delta + 4mV}}$$ $$\Delta = \frac{\sum_i \rho_i(\rho_i - \bar{\rho})S_i^2}{(\bar{\rho}\,S)^2 - \sum_i \rho_i^2\,S_i^2}$$	$$\mathbf{TR = 0.49 \cdot \frac{V}{-S \cdot \ln(1-\bar{\alpha}) + \Delta + 4mV}}$$ $$\Delta = \frac{\sum_i \rho_i(\rho_i - \bar{\rho})S_i^2}{(\bar{\rho}\,S)^2 - \sum_i \rho_i^2\,S_i^2}$$
Fitzroy Fórmula preferifa para salas rectangulares con absorción no uniformemente distribuida. las áreas para dos paredes paraleleas se denotan (Ax, Ay, Az) y los coef's de absorción medios respectivos $(\bar{\alpha}_x,\ \bar{\alpha}_y,\ \bar{\alpha}_z)$.	$$\mathbf{TR = 0.16 \cdot \frac{V}{S^2} \cdot \left(-\frac{A_x}{\ln(1-\bar{\alpha}_x)} - \frac{A_y}{\ln(1-\bar{\alpha}_y)} - \frac{A_z}{\ln(1-\bar{\alpha}_z)}\right)}$$	$$\mathbf{TR = 0.49 \cdot \frac{V}{S^2} \cdot \left(-\frac{A_x}{\ln(1-\bar{\alpha}_x)} - \frac{A_y}{\ln(1-\bar{\alpha}_y)} - \frac{A_z}{\ln(1-\bar{\alpha}_z)}\right)}$$
Fitzroy modificada Fórmula preferifa para salas rectangulares reales (se asume que la absorción mas grande esta siempre en el suelo o el techo), $\rho = 1 - \alpha$. $(\bar{\alpha}^*_{walls},\ \bar{\alpha}^*_{cf})$ denotan los exponentes de absorción efectivos de paredes y suelo+techo, respectivamente. $\bar{\alpha}$ es la media aritmética del coeficiente de absorción medio.	$$\mathbf{TR = 0.32 \cdot \frac{V}{S^2}\left(\frac{l_z(l_x + l_y)}{\bar{\alpha}^*_{walls}} + \frac{l_x l_y}{\bar{\alpha}^*_{cf}}\right)}$$ $$\bar{\alpha}^*_{walls} = -\ln(1-\bar{\alpha}) + \left(\frac{\rho_{walls}(\rho_{walls} - \bar{\rho})S^2_{walls}}{(\bar{\rho}\,S)^2}\right)$$ $$\bar{\alpha}^*_{cf} = -\ln(1-\bar{\alpha}) + \left(\frac{\rho_{cf}(\rho_{cf} - \bar{\rho})S^2_{cf}}{(\bar{\rho}\,S)^2}\right)$$	$$\mathbf{TR = 0.98 \cdot \frac{V}{S^2}\left(\frac{l_z(l_x + l_y)}{\bar{\alpha}^*_{walls}} + \frac{l_x l_y}{\bar{\alpha}^*_{cf}}\right)}$$ $$\bar{\alpha}^*_{walls} = -\ln(1-\bar{\alpha}) + \left(\frac{\rho_{walls}(\rho_{walls} - \bar{\rho})S^2_{walls}}{(\bar{\rho}\,S)^2}\right)$$ $$\bar{\alpha}^*_{cf} = -\ln(1-\bar{\alpha}) + \left(\frac{\rho_{cf}(\rho_{cf} - \bar{\rho})S^2_{cf}}{(\bar{\rho}\,S)^2}\right)$$

La RIR se puede medir con el software adecuado o sintetizar mediante modelos de acústica de salas. En este proyecto se ha modelado usando la función de Matlab "rir.m" de uso público, que usa el "método de imágenes" (ver sección 1.7.2. , "Dominio temporal: rayos acústicos"). Por completitud, la RIR también se midió para la sala donde se desarrolló el proyecto (sala poliédrica, de dimensiones 7m x 4m x 2.3m, dividida en dos por un arco saliente) mediante los softwares Sonarworks Reference 3 © y Room EQ Wizard © [29, 30].

Se ha preferido omitir todo lo referido a la medición de las RIRs a los resultados obtenidos, ya que excede la complejidad del proyecto. En su lugar, la función "rir.m" usada, que es bastante realista y permite un efecto de reverberación básico, permite definir RIRs para cada par fuente-oyente [26]. Como se explica en el capítulo 3, para cada trama de audio de entrada se calcula una nueva RIR en base a la posición actual de la fuente describiendo una trayectoria en 3D dentro de la sala, con el oyente siempre fijo. Esto, junto con las dos HRTFs L y R (también recalculadas para cada trama) permite obtener el efecto de auralización deseado.

En vez de usar RIRs y HRTFs, también se pueden usar las llamadas *respuestas al impulso binaurales (Binaural Room Impulse Responses BRIR)* si éstas estuvieran disponibles. Las BRIRs son una cuidada y complicada mezcla de las HRTFs (izquierda y derecha) y la RIR. Hay laboratorios que se dedican exclusivamente a medir BRIRs de todo tipo de auditorios, salas de conciertos y salas de grabación y post-producción alrededor del mundo, con el correspondiente desembolso económico por parte del usuario. Como ejemplo, Altiverb 7 es un plug-in para reverberación mediante convolución desarrollado por la empresa holandesa Audio Ease que contiene RIRs de algunos de los estudios, auditorios y las salas de conciertos más prestigiosos del mundo [25]. El software está pensado para productores musicales y de cine e ingenieros de post-producción, con posibilidad de mezclar también en 5.1 Surround. Hay también empresas nacionales dedicadas a este sector [24].

1.7.2. Consideraciones acústicas y perceptuales

Hasta ahora nos hemos centrado en las posiciones aparentes (virtuales) de las fuentes sonoras. En esta sección, centramos la atención en la reverberación del sonido como un efecto natural que ocurre cuando las ondas de sonido se propagan en un espacio cerrado. La reverberación contiene información sobre la naturaleza y texturas de los materiales de la sala, así como del tamaño y forma de esta y de los objetos contenidos en ella.
Con el fin de analizar los diferentes aspectos acústicos de la reverberación, en este proyecto se ha considerado una sala rectangular que contiene una fuente puntual omnidireccional. Es sabido desde la formulación del principio de dualidad en la era de la revolución de la Física Cuántica que cualquier onda (electromagnética, de presión, …) se puede considerar como onda y como partícula. En el ámbito de la Acústica, estas dos caracterizaciones permiten obtener diferentes resultados y ambos métodos de análisis van a resultar igual de útiles y en muchas ocasiones complementarios.

Fuente puntual en un espacio cerrado
Considerando la onda de presión sonora como una onda, sea una fuente esférica pulsante pequeña radiando a una frecuencia de ω rad/s. Considerando la onda de presión generada, el número de onda k correspondiente se obtiene de

$$k = \frac{\omega}{c}$$

Se puede demostrar que la velocidad de particula del aire depende del radio r (o distancia de la fuente al oyente), en fase con la presión sonora para distancias grandes ($r >> 1/k$, far-field) y en cuadratura de fase con la presión sonora para distancias pequeñas ($r < 1/k$, near-field). En el campo cercano, hay una componente de velocidad importante que, estando en cuadratura de fase, será responsable de algo de la energía reactiva que no es radiada. Una manera de análisis común es considerar que tan pronto las ondas han viajado un camino suficientemente largo desde la fuente (campo lejano), las ondas se pueden considerar planas. Así, se simplifica el problema de propagación esférica al de ondas planas. No obstante, en la proximidad de la fuente (campo cercano) va a ser necesaria una descripción basada en ondas esféricas.

El otro punto de vista es considerar el sonido como partícula. Esto es útil para los fenómenos que no son fácilmente identificables mediante los otros dos métodos de campo lejano y cercano. Por ejemplo, para la

difusión en paredes rugosas, un enfoque desde el punto de vista de la Óptica de Rayos resulta muy útil. También para la síntesis de respuestas al impulso de salas (RIRs), este punto de vista es mas adecuado.

Dominio frecuencial: modos normales

La acústica de un espacio cerrado puede describirse por la superposición de los modos normales. Los modos normales son ondas estacionarias que pueden darse en el gas (aire) que llena el espacio cerrado rectangular. Esta herramienta es útil para comprender el comportamiento frecuencial de la sala como si de un filtro común se tratara. Para una aproximación sencilla al calculo de modos normales, se simplifica la situación a ondas sonoras que se propagan en el campo lejano y después del transitorio inicial. Esta asunción permite considerar todos los frentes de onda como planos, de manera que se trabaje con ondas planas.

Como se ha visto, las frecuencias de los modos normales son:

$$\mathbf{f_{ijk}} = \frac{\mathbf{c}}{\mathbf{2}}\sqrt{\left(\frac{\mathbf{i}}{\mathbf{l_x}}\right)^2 + \left(\frac{\mathbf{j}}{\mathbf{l_y}}\right)^2 + \left(\frac{\mathbf{k}}{\mathbf{l_z}}\right)^2}$$

Cada modo normal se sustenta sobre una onda plana propagándose en la dirección

$$\overrightarrow{\mathbf{v_{ijk}}} = \left(\frac{\mathbf{i}}{\mathbf{l_x}}, \frac{\mathbf{j}}{\mathbf{l_y}}, \frac{\mathbf{k}}{\mathbf{l_z}}\right)$$

Se define el vector $\overrightarrow{\mathbf{n_{ijk}}} = (\mathbf{i}, \mathbf{j}, \mathbf{k})$ como un vector de enteros no negativos que caracteriza cada modo normal (i, j, k). De las ecuaciones anteriores se deduce automáticamente que con todos los modos con frecuencias múltiplo del mismo modo fundamental se asocia una misma dirección de propagación. En otras palabras, todas las 3-plas múltiplos de una 3-pla no reducible son asociadas con una serie armonica de frecuencias resonantes con misma dirección de propagación. Esta propiedad sufiere que dichas series se pueden reproducir usando un *filtro peine (comb filter),* i.e. usando una línea de retardo de realimentación con longitud

$$\mathbf{d} = \frac{\mathbf{1}}{\mathbf{fo}}$$

donde fo es la frecuencia fundamental de la serie armónica.
No todos los modos son excitados con la misma intensidad en cada posición (x, y, z) de la sala. La excitación más intensa de todas se da en las esquinas de la sala. Como ejemplo, en el punto central de una caja rectangular solo 1/8 de los modos son excitados (solo aquellos que tengan todas las componentes pares en la 3-pla $\vec{n}$, ya que todos los demás tienen un punto nodal en el centro).

Dominio temporal: rayos acústicos

Otra manera de describir la acústica de una sala es considerando la propagación de *rayos acústicos*: porciones de ondas esféricas con apertura pequeña. Esta es la base de la Acústica Geométrica. Es análoga a la Óptica Geométrica para ondas electromagnéticas y aplican todos los conceptos de la Óptica.

Para asegurar la validez de una descripción mediante rayos geométricos, hay que establecer primero una serie de bases. Lo primero que se asume es que las longitudes de onda que interesan para la propagación son mucho menores que la descripción geométrica mas precisa que se quiera hacer de la sala.

Lo segundo es confirmar que podemos ignorar todos los fenómenos de difracción e interferencia. Ya que trabajamos con un rango de frecuencias muy grande y la difracción se puede insignificante, sobre todo a altas frecuencias. Además, la ausencia de interferencias se puede verificar en la mayoría de los casos prácticos, donde los rayos suelen ser mutuamente incoherentes.

A lo largo del rayo acústico, la presión sonora decrece con 1/r ya que la energía en un frente esférico tiene que ser conservada durante la propagación y el área ocupada por el frente de onda aumenta con r^2.

Cuando un rayo choca con una superficie, se refleja proporcionalmente al material de la superficie. Es común asociar también cierto filtrado con el fenómeno de reflexión debido a las propiedades absorbentes

de los materiales. En general, hay una atenuación dependiente de la frecuencia y un retardo temporal variable. Tal filtrado se puede caracterizar mediante una función de reflexión R compleja

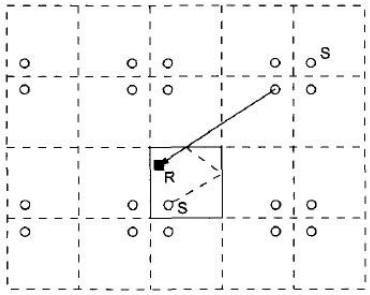
$$\mathbf{R(\omega, \theta)} = \frac{\mathbf{Z(\omega) \cdot \cos(\theta) - 1}}{\mathbf{Z(\omega) \cdot \cos(\theta) + 1}}$$

donde Z es la impedancia carteristica de la pared, que en general es compleja. Para una pared lisa y rigida, Z tiende a infinito (reflexión perfecta, $R = 1$ para toda ω y todo ángulo θ).

Hay dos métodos especialmente populares para modelar el campo de onda sonoro mediante rayos acústicos.

El método de imágenes o "Image Source Method" se ha explicado anteriormente. Por completitud, se repite el esquema de la situación de la sala, donde en la figura el punto S denota al transmisor (la fuente) y el punto R al receptor (el oyente o el micrófono). En el método de imágenes, la fuente S es reflejada simétricamente en los bordes y las imágenes de la fuente son reflejadas de nuevo y asi sucesivamente.

El mapa topográfico resultante es tal que cada rayo conectando una imagen de la fuente con el punto R corresponde a un camino acústico único que conecta S con R tras un cierto numero de reflexiones especulares. Este método es fácil de extrapolar a un espacio 3D. Para salas descritas por poliedros cualesquiera, con muchas complicaciones y complejidad, el método también es aplicable. El software



[CATT-Acoustics] es un ejemplo de uso de este método en un programa de simulación por ordenador.

El método de rayos o "Ray Tracing" ha sido ampliamente utilizado desde los años 1980 para computación gráfica y es una técnica que produce imágenes realistas. En la computación gráfica, una imagen se construye trazando rayos de luz desde el ojo (o plano focal) a la fuente de luz. En audio, una imagen acústica se construye trazando rayos acústicos desde la oreja (o micrófono) a la fuente de sonido. En ambos casos, a la escena la atraviesan los rayos y todas las reflexiones se pueden reproducir correctamente. Hay una diferencia importante para este método según se aplique a imágenes o a sonidos: mientras que en la computación gráfica la luz puede considerarse que cambia para todos los puntos de la escena por cada trama, en el audio la tasa de muestreo es mucho mayor y la velocidad de propagación mucho menor. Como consecuencia, cada rayo ha de asociarse a una línea de retardo, aumentando los requerimientos de memoria y complejidad a un método ya de por si pesado.

Ambos métodos son usados para temas de Acústica Arquitectónica por profesionales, implementados en programas de *"Computer Aided Design" (CAD)* que no precisan alto rendimiento en tiempo real. Por otro lado cabe destacar que, para este tipo de aplicaciones, la precisión de los resultados es muy importante para poder predecir la calidad acústica del auditorio o sala, antes de ser construida.

Un tercer método ha sido propuesto hace relativamente poco [27]. Este método, llamado de *transferencia de la radiancia acústica o "Acoustic Radiance Transfer"* se basa en la radiosidad progresiva aplicada al sonido y usa modelos de reflexión arbitrarios. La energía acústica se radia de la fuente a las paredes del modelo, que han sido dividas en un entramado de parches. La propagación de la energía de parche a parche es calculada y guardada. Finalmente, cuando se ha conseguido la precisión deseada, la energía se recolecta de los parches hacia el oyente [1].
Estos tres métodos tienen propiedades diferentes para la simulación de efectos de acústica de salas. Por supuesto, existen también métodos híbridos.
El método de "imágenes" solo puede modelar reflexiones especulares pero es muy eficiente encontrando

reflexiones tempranas perceptualmente importantes. Es aplicable para tiempo real pues las fuentes imagen (las reflexiones tempranas) pueden renderizarse espacialmente y luego añadir una reverberación tardía (con una red de retardo realimentada por ejemplo).
El método de "rayos" no es aplicable a software de tiempo real. No obstante, es preciso calculando la RIR, que puede ser luego convolucionada con un convolvotrón de tiempo real.
El método de "transferencia de radiancia acústica" es el más avanzado de los tres, en términos tecnológicos. Puede manejar reflexiones especulares y también reflexiones difusas. Aunque computacionalmente costoso, el uso de GPUs permite el renderizado de audio 3D en tiempo real, permitiendo así efectos de reverberación interactivos para entornos con características acústicas personalizables [1].

El sonido como particula: difusión

Hasta aquí se han planteado temas de acústica de salas desde el punto de vista macroscópico, contemplando rayos y modos. Si el fin de la aplicación a diseñar es conseguir el realismo (la complejidad), es necesario considerar también los fenómenos microscópicos que ocurren. El más importante de estos efectos es el de difusión de las ondas sonoras. La difusión se modela mejor usando una representación del sonido en forma de partículas; esto es aplicable debido a la dualidad onda-partícula. Comparando otra vez con los métodos de la computación gráfica se reconoce este mismo problema pero en el modelado de la difusión de la luz superficial y de las sombras. Para adaptar las herramientas usadas en computación grafica al ámbito de la acústica de salas, se necesita seguir considerando la asunción de señales de banda ancha. Y más aún, ya que estamos interesados en las interacciones locales con superficies irregulares (rugosas) comparables con las longitudes de onda que se manejan, es necesario asumir que las partículas acústicas son mutuamente incoherentes para prevenir cancelaciones entre particular. Una vez se han establecido estas hipótesis y se han asumido como validas, es lógico considerar una partícula sonora como una señal acústica de corta duración pero con gran ancho de banda. El modelo de difusión más usado en computación gráfica usa la función de distribución de reflexiones bidireccionales (BRDF) $f_r(\varphi_i, \varphi_0)$, que describe el cociente entre la *radiancia* reflejada a lo largo de la dirección φ_0 y la *irradiancia* incidente a lo largo de la dirección φ_i. La difusión se denomina *difusión lambertiana* cuando la BRDF es constante. En este caso las partículas sonoras se difunden en todas las direcciones con misma probabilidad, independientemente de la dirección. Y viceversa, si todas las partículas llegando de la dirección φ_i son redirigidas a la dirección especular $\varphi_0 = -\varphi_i$ se obtiene una *reflexión espejo* (no existe difusión ninguna).

1.8. Áreas de aplicación de los modelos binaurales y los sistemas de audio 3D

La asombrosa capacidad del sistema auditivo binaural ha atraído la atención de científicos e ingenieros por ya largo tiempo. Se han construido modelos de partes e incluso la totalidad del sistema auditivo humano. Aunque los modelos no son perfectos, permiten recrear ciertas funciones del sistema auditivo humano y ayudan a comprenderlo mejor e incluso impulsan nuevas investigaciones.

Aparte del ámbito de la investigación, estos modelos tienen el potencial de poder ser aplicados para el público ya que recrean funciones específicas y tecnológicamente interesantes del sistema auditivo.

El grupo de investigación AABBA [15] formado en 2009 por Jens Blauert debatió las siguientes aplicaciones potenciales para el audio binaural o audio 3D con el fin de definir los propósitos de cada aplicación. El grupo de investigación sugirió las siguientes categorías de aplicación [16]:

- **Tecnologías de audio**. Selectores de características binaurales, asesoramiento de la calidad de canales de audio, asesoramiento de la calidad de altavoces, automatización para evaluación de la calidad de transmisión, aprendizaje a distancia para temas relacionados con el audio 3D, edición profesional de audio para música, cine y TV.
- **Audiología.** Asesoramiento de desórdenes en la escucha binaural, asesoramiento para la dereverberacion binaural y decoloración binaural, displays acústicos para personas con problemas auditivos, asesoramiento para las capacidades de reconocimiento de voz en entornos acústicos adversos, medidor de volumen binaural.
- **Entornos virtuales para auralización**. Mapeo de escenas auditivas, identificación de fuentes sonoras virtuales, asesoramiento para el tamaño percibido de una sala, realidad virtual y realidad aumentada para sistemas comerciales de entretenimiento.
- Problemas de audición. Ayuda para problemas de audición, diagnósticos de disfunciones del sistema auditivo.
- **Control de Calidad de productos de sonido**. Asesoramiento para las propiedades espaciales de productos de sonido
- **Acústica de salas**. Detector de eco, medidor de espacialidad, detector de cambios en las imágenes sonoras, asesoramiento en la percepción de envolvimiento e inmersión, asesoramiento con el efecto de precedencia, asesoramiento con la "calidad global de la acústica" de una sala
- **Tecnologías de voz.** Posicionado del oyente por voz, inteligibilidad binaural del habla, asesoramiento con el reconocimiento de voz en entornos acústicos adversos, desarrollo sobre el efecto "cocktail-party", ·teleconferencia, tele presencia.
- **Modelos binaurales como herramienta de investigación.** Para ser empleados en la evaluación, asesoramiento y análisis de la escucha espacial en un mundo multimodal, p. ej. para oyentes moviéndose en el espacio y/o recibiendo características adicionales visuales y/o táctiles.

BIBLIOGRAFÍA

[1] "Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality", Vorländer, M. Springer, 2008.

[2] "Adaptive 3D Sound Systems", Garas.

[3] "Springer Handbook of Acoustics", cap's 12, 13 y 14, Rossing, M, Springer, 2000.

[4] "Spatial Audio Processing", Breebaart, J. and Faller, C. Wiley, 2008.

[5] "Speech and Audio Technology Script 2013/2014", cap. 5, Möller, S., Technical University of Berlin, 2013.

[6] "The Technology of Binaural Hearing", cap. 1, Blauert, J. 2013.

[7] "Spatial Audio", Rumsey, F. Focal Press, 2001.

[8] "The Auditory Modeling Toolbox," in "The Technology of Binaural Listening", Blauert, J. pp. 33-56. Søndergaard, P. and Majdak, P. Springer, 2013.
http://amtoolbox.sourceforge.net/

[9] "On sound localization." Wallach, H. AES, 1939.

[10] "3-D Audio using Loudspeakers", Gardner, W. Kluwer, 1998.

[11] "Spatial Hearing", Blauert, J. MIT Press, 1997.

[12] "HRTF Measurements of a KEMAR Dummy-Head Microphone", Gardner, W. et al. MIT Media Lab, 1994.
http://sound.media.mit.edu/resources/KEMAR.html.

[13] "Distance Dependent Head-related Transfer Function Database of KEMAR", Qu, T. et al. Peking University, 2008.

[14] "Computer Vision: Algorithms and Applications", Szeliski, R. Springer, 2010.

[15] "Aural assessment by means of binaural algorithms - The AABBA project". Blauert, J. Braasch, J. Bucholz, H.S. Colburn, U. Jekosch, A. Kohlrausch, J. Mourjopoulos, V. Pulkki, A. Raake.
Proceedings of ISSAR 2009: Binaural Processing and Spatial Hearing, 2nd International Symposium on Auditory and Audiological Research, 2009.

[16] "Binaural models and their technological application". Blauert, J. ICSV19, 2009.

[17] "Technische Akustik", 8. Auflage, Möser, M. Techinchal University of Berlin, Springer, 2009.

[18] "The Yamaha Sound Reinforcement Handbook", Davis, G. and Jones, R. Hal Leonard, 1987.

[19] "Collected Papers On Acoustics", Sabine, W. C. Cambridge: Harvard University Press, 1923.

[20] "Reverberation Time in "Dead" Rooms", Eyring, C. F.
The Journal of the Acoustical Society of America, Vol. 1, 1930.

[21] "Reverberation formulae which seems to be more accurate with non-uniform distribution of absorption", Fitzroy, D.
The Journal of the Acoustical Society of America, Vol. 31, 1959.

[22] "Nachhall und effective Absorption in Räumen mit diffuser Wandreflexion", Kuttruff, H. Acustica, Vol. 35, No. 3, 1976.

[23] "Estimation of Reverberation Time in rectangular rooms with non-uniformly distributed absorption using a modified Fitzroy Equation", Neubauer, R. O.
7th International Congress on Sound an Vibration, 2000.

[24] "IRIS. Medición de la respuesta al impulso en 3D", Alava Ingenieros.
http://www.alava-ing.es/ingenieros/productos/acustica-y-vibraciones/software-simulacion-acustica/software-de-medicion-de-respuesta-al-impulso-en-3d/, 2014.

[25] "Altiverb 7 © convolution reverberation plug-in", Audio Ease.
http://www.audioease.com/Pages/Altiverb/

[26] "A model for Room Acoustics", McGovern, S. 2003.
http://www.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator/content/rir.m

[27] "Frequency Domain Acoustic Radiance Transfer for Real-Time Auralization," Siltanen, S., Lokki, T. and Savioja, L. *Acta Acustica United with Acustica,* Vol. 95, No. 12, 2009.

[28] "Beyond Control. Acoustics of sound recording control rooms – past, present and future", van Munster, B. FAGO group, Eindhoven University of Technology, the Netherlands, 2003.

[29] "Room EQ Wizard REW ©", Mulcahy, J,
http://www.roomeqwizard.com/index.html#, 2004.

[30] "Sonarworks Reference 3 ©", Sonarworks ltd.
http://sonarworks.com/speakers/overview/, 2013.

2 ESTUDIO DEL ESTADO DEL ARTE

2.1. Introducción

En esta capitulo se presenta una visión general del estado del arte de los sistemas de audio 3D. Algunos temas se desarrollan más que otros por ser más interesantes desde el punto de vista del autor o porque el software desarrollado en este proyecto ha usado ese tema en concreto y ha sido necesario investigarlo en profundidad. Las tecnologías que se describen aquí son sobre todo técnicas de reproducción, aunque se menciona la grabación de señales para reproducción en algunos de estos sistemas. Paneo en amplitud, Ambisónica, HRTFs, audio 3D con altavoces o Síntesis de Campos de Onda (WFS) son algunos ejemplos de lo que se revisa en este capítulo. Se acompaña la investigación con ejemplos prácticos para Matlab®.

Este capitulo no pretende ser una guía de todas las tecnologías del audio 3D ni presentar novedades desarrolladas por el autor; todo este capítulo es una recopilación de información y no un desarrollo de nuevas tecnologías. Es un estudio en mayor o menor profundidad de las técnicas mas fundamentales y también de las mas relevantes en las investigaciones conducidas en el pasado y actualmente. La lectura de este capítulo, junto con el capitulo anterior, da al lector una visión general de la Acústica y la Psicoacústica, así como una referencia de apoyo para profundizar en los temas que se traten aquí.
Al final del capítulo se muestran las referencias o bibliografía que se mencionan durante la lectura.

2.2. Técnicas y efectos de reproducción mediante altavoces estéreo y auriculares

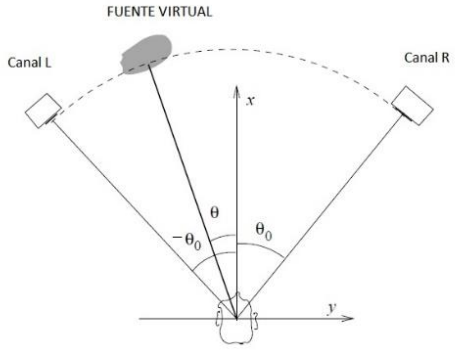
2.2.1. Audio estéreo

Estéreo de dos canales

Aunque muchos de los principios del sonido estéreo fueron desarrollados mediante investigaciones en los comienzos de 1930, sigue habiendo una confusión relacionada con el termino "estéreo". Generalmente asociado a la reproducción mediante dos altavoces, la palabra proviene del término griego *stereos,* que significa solido o tridimensional. La asociación del termino al uso común de dos canales de audio i zquierdo (L) y derecho (R) se genero en los años 1950 debido a las limitaciones tecnológicas impuestas por los discos de vinilo, que tenían solo dos survos para codificar la información. La configuración típica de un sistema de audio estéreo en el plano x-y (2D) se muestra a continuación. Las dos fuentes de sonido, *los altavoces* , radian sonido independientemente, denotado por "Canal L" y "canal R".

El oyente se suele situar en una posición simétrica respecto a los dos altavoces para percibir igual intensidad de sonido en ambos oídos. Los altavoces se posicionan en un arco llamado *arco activo* y dibujado con línea discontinua en la figura. También se pueden situar uno al lado del otro, con la línea recta que los une llamada *baseline.* Se elija la configuración que se elija, los altavoces estarán situados a un cierto ángulo del centro de la cabeza del oyente. Este ángulo se puede llamar *ángulo azimutal* aunque recibe otros nombres como *ángulo de apertura* o ángulo horizontal. Este ángulo queda especificado por la constante θ_0. Normalmente θ_0 se mide respecto a la orientación del oyente, en un ánglo comprendido entre los ejes x e y (por simplicidad, se suele posicionar el eje x coincidiendo con el eje interaural que pasa por los oídos del oyente). Comprende ángulos entre 10º y 60º, aunque debido a las numerosas pruebas que tanto oyentes como profesionales han llevado a cabo durante la historia de esta tecnología, se han probado infinitas combinaciones de ángulos y distancias para configuraciones de altavoces.

Cuando las fuentes (los altavoces L y R) radian las ondas acústicas que han sido decodificadas del material de audio, se crea un *campo de onda sonoro* alrededor de todo el oyente. Las soluciones del campo se pueden deducir de la ecuación de onda usando diferentes condiciones de contorno que constituirán, entre otros hallazgos más actuales, la *acústica de la sala*.



Las labores de investigación en el audio no se desarrollaron hasta 1930 en los laboratorios Bell Labs. No obstante en el siglo XIX sucedió la primera transmisión estéreo. Tuvo lugar en el año 1881 en una exposición en París. Clement Ader distribuyó una serie de pickups de teléfonos de la época (micrófonos) por el espacio del escenario de la Ópera de París y conectó éstos a receptores telefónicos situados en la

sala de la exhibición. Los visitantes, asombrados, escuchaban la ópera en directo y con cierto realismo espacial a varios metros de distancia de donde se tocaba la pieza musical.

La estereofonía tuvo un impulso inicial importante con el trabajo de Blumlein [1] en el Reino Unido. Blumlein reconoció acertadamente que era posible localizar un sonido en un rango de angulos azimutales usando una combinación apropiada de diferencias de nivel y retardo (las características binaurales ILD e ITD que se descubrieron posteriormente). Su trabajo se centró en la reproducción precisa del campo sonoro en cada oído y en el desarrollo de técnicas de grabación/microfónicas que permitieran grabar las diferencias de amplitud y fase necesarias para la reproducción estéreo.

En los laboratorios Bell en los Estados Unidos de América,los científicos Fletcher, Steinberg y Snow tomaron un camino diferente [2]. Consideraron una "pared de sonido" en la que un número infinito de micrófonos se usan para reproducir un campo sonoro con infinitos altavoces, similar al concepto del principio de Huygens [ver sección 2.8].
Aunque teoréticamente correcto, los laboratorios Bell se dieron cuenta que las implementaciones prácticas involucraban un número de canales mucho menor. Demostraron que un sistema de 3 canales (L, R y C) en el plano horizontal o plano azimutal podía representar la lateralización y profundidad del campo sonoro deseado con una precisión aceptable.
El primer sistema estereofónico (de tres canales) fue puesto en practica en 1934. La orquesta de Filadelfia tocó para una audiencia a distancia, situada en Washington, D.C., Estados Unidos, transmitiendo el sonido a través de líneas telefónicas de banda ancha.

Estéreo de cuatro canales (Quadraphonics)

Aunque los métodos estereofónicos pueden ser una herramienta útil para la representación de atributos espaciales de un campo sonoro, no son suficientes para reproducción 3D real. El sistema cuadrafónico intentó solventar, aunque sin éxito, estas limitaciones, capturando y transmitiendo la información del sonido directo y del sonido reverberante. Para transmitir los cuatro canales requeridos en las grabaciones cuadrofónicas sobre un medio de dos canales fue necesario desarrollar un *codec* apropiado. Se propusieron sistemas basados en los sistemas de codificación/decodificación matriciales 4:2:4 [3]. Los sistemas cuadrafónicos fueron capaces de reproducir imágenes sonoras bastante realistas en las sectores frontal/atrás del plano azimutal pero exhibieron limitaciones severas en el caso de imágenes sonoras situadas a los laterales del oyente. Además, la imagen frontal era bastante pobre, usualmente con un "agujero" en el medio.

Esta técnica no fue bien recibida por el publico en general, que parecían negarse a instalar los altavoces adicionales requerido. Además, parece que la codificación no se desarrolló de la manera correcta pues, aunque se suponía compatible con la reproducción mediante dos altavoces, se solian producir efectos laterales audibles.

Estereofonía

Desde el lanzamiento del audio estéreo, se han propuesto muchos procedimientos para intentar agrandar el campo sonoro. Campos sonoros en sólo el plano horizontal (llamados *pantofónicos*) se lograron usando varios altavoces y sistemas de codificación/decodificación matricial. En la mayoría de los sistemas desarrollados, los altavoces se situan en un plano bidimensional, horizontal normalmente.
Otros autores intentaron crear campos sonoros esféricos (llamados *perifónicos*) usando una configuración de altavoces en 3 dimensiones, tal como en la Holofonía, la Ambisónica y la WFS [ver secciones 2.4, 2.8].

En la mayoría de los sistemas, las posiciones de los altavoces son fijas. En la Ambisónica, el número de altavoces y su posicionamiento puede ser variable; no obstante, la mejor sensación de espacialización se consigue con un posicionamiento de altavoces ortogonal. Si se incrementa el numero de altavoces, la espacializacion no mejora apreciablemente.

Los campos sonoros pueden ser producidos por dos altavoces o con unos auriculares, filtrando el audio con modelos digitales de las funciones de transferencia relacionadas con la cabeza (HRTFs), que se explican detalladamente más adelante. La información espectral de la dirección de la fuente sonora es añadida al audio.

Una mejora posible y no tenida en cuenta en los años del desarrollo del VBAP hubiera sido un sistema de posicionamiento virtual de fuentes sonoras que fuera independiente de la configuración de posicionamiento de los altavoces y pudiera producir fuentes sonoras virtuales con muy alta precisión usando la configuración de altavoces disponible.

2.2.2. Surround matricial

Los sistemas matriciales han permitido que se extienda el uso del sonido envolvente o Surround. Uno de los motivos es la compatibilidad con los sistemas estéreo de dos canales. Los sistemas matriciales codifican múltiples canales de audio en dos canales. Dada una señal de audio codificada en forma matricial, un decodificador matricial es aplicado para aproximar la señal multicanal original.
Los primeros pasos dados en este tipo de sistemas fueron dirigidos a usarlos para codificar cuatro canales de audio en una señal estéreo, llamados sistemas 4-2-4 [ver subsección Quadraphonics]. Los sistemas más actuales, como Dolby® Prologic II™ son capaces de codificar señales de audio 5.1 (incluido el canal de LFE). Más recientemente se han desarrollado sistemas matriciales para señales 6.1 y 7.1 mediante el estándar Dolby® Pro Logic IIx™.
Lo que sigue es una breve descripción de cómo implementar un codec matricial 5-2-5 para codificar y decodificar los 6 canales de una señal de audio Surround 5.1.

Para la codificación, el esquema es simple; definiendo las señales codificadas estéro como l(n) y r(n), se demuestra que:

$$\begin{bmatrix}\mathbf{l(n)}\\ \mathbf{r(n)}\end{bmatrix} = \begin{bmatrix}\mathbf{1} & \mathbf{0} & \frac{\mathbf{1}}{\sqrt{\mathbf{2}}} & -\mathbf{j}\sqrt{\frac{\mathbf{2}}{\mathbf{3}}} & -\mathbf{j}\sqrt{\frac{\mathbf{1}}{\mathbf{3}}}\\ \mathbf{0} & \mathbf{1} & \frac{\mathbf{1}}{\sqrt{\mathbf{2}}} & \mathbf{j}\sqrt{\frac{\mathbf{1}}{\mathbf{3}}} & \mathbf{j}\sqrt{\frac{\mathbf{2}}{\mathbf{3}}}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{l_f(n)}\\ \mathbf{r_f(n)}\\ \mathbf{c(n)}\\ \mathbf{l_s(n)}\\ \mathbf{r_s(n)}\end{bmatrix}$$

donde los cinco canales son: $l_f(n)$ para el canal L frontal, $r_f(n)$ para el canal R frontal, $c(n)$ para el canal central, $l_s(n)$ para el canal Surround L trasero y $r_s(n)$ para el canal Surround R trasero.

En forma vectorial se puede escribir:

$$\begin{bmatrix}\mathbf{l(n)}\\ \mathbf{r(n)}\end{bmatrix} = \overrightarrow{\mathbf{v_{lf}}} \cdot \mathbf{l_f(n)} + \overrightarrow{\mathbf{v_{rf}}} \cdot \mathbf{r_f(n)} + \overrightarrow{\mathbf{v_c}} \cdot \mathbf{c(n)} + \overrightarrow{\mathbf{v_{ls}}} \cdot \mathbf{l_s(n)} + \overrightarrow{\mathbf{v_{rs}}} \cdot \mathbf{r_s(n)}$$

donde los cinco vectores $\overrightarrow{v_i}$ son las columnas de la matriz anterior. Estos vectores son unitarios ($||v|| = 1$) y mapean los cinco canales a las dos salidas l(n) y r(n).

Para la decodificación, la manera mas simple es aplicar una matriz de decodifiacion pasiva:

$$\widehat{\mathbf{l_f}}(\mathbf{n}) = [\mathbf{l(n)}\ \mathbf{r(n)}] \cdot \overrightarrow{\mathbf{v_{lf}}}$$

$$\widehat{\mathbf{r_f}}(\mathbf{n}) = [\mathbf{l(n)}\ \mathbf{r(n)}] \cdot \overrightarrow{\mathbf{v_{rf}}}$$

$$\hat{\mathbf{c}}(\mathbf{n}) = [\mathbf{l(n)}\ \mathbf{r(n)}] \cdot \overrightarrow{\mathbf{v_c}}$$

$$\widehat{\mathbf{l_s}}(\mathbf{n}) = [\mathbf{l(n)}\ \mathbf{r(n)}] \cdot \overrightarrow{\mathbf{v_{ls}}}$$

$$\widehat{\mathbf{r_s}}(\mathbf{n}) = [\mathbf{l(n)}\ \mathbf{r(n)}] \cdot \overrightarrow{\mathbf{v_{rs}}}$$

Cada canal de salida es la proyección de la señal codificada matricialmente en su correspondiente vector de codificación matricial.

Como se puede comprobar fácilmente, el sistema anterior reconstruye perfectamente un canal en el caso de que solo haya un canal activo. Los otros canales no serán cero en este caso; esto hace que este método no sea muy bueno porque provoca mucha diafónia o *Crosstalk*.

No obstante, la situación donde solo un canal está activo es fácilmente detectable y se puede implementar un algoritmo de detección para anular los demás canales en esta situación, eliminando buena parte el Crosstalk.

Habitualmente se filtran y retardan los dos canales de atrás. El filtrado se hace paso-bajo para simular la reverberación tardía, que tiene las frecuencias altas atenuadas debido a la mayor absorción del aire a frecuencias altas. El retardo previene que las componentes se muevan por los altavoces cuando el decodificador no pueda separar correctamente los canales. Debido al efecto de precedencia, las componentes de señales correladas frontales y de atrás siempre serán percibidas como viniendo de frente si se retardan los canales traseros. Para diálogos en el audio de una película esto es especialmente importante.

Aunque este esquema es muy práctico, la calidad del sonido envolvente siempre será muy limitada. Las limitaciones las imponen la dependencia de canales inherente asi como el alto Crosstalk entre canales.

Los esquemas de codificación espaciales consiguen una calidad de audio mucho mayor. Para ello, "guían" la mezcla hacia arriba dependiendo de las características binaurales que este sistema va extrayendo del audio.

2.2.3. Paneo en amplitud basado en vectores para altavoces (VBAP)

El *paneo en amplitud* y *Vector Base Amplitud Panning (VBAP)* son técnicas esterofónicas ambas desarrolladas por Ville Pulkki a finales de los años 1990 [4, 5, 6]. Permiten el uso de un número ilimitado de altavoces en una configuración 2D o 3D alrededor del oyente. Se requiere que los altavoces estén posicionados lo más equidistantes posibles del oyente y que la sala en la que se escuche el audio no sea muy reverberante (ver reverberación y Acústica de salas, sección 2.7).
El método más simple de paneo en amplitud es el paneo en amplitud en 2D, también conocido como *paneo en intensidad*. Es una aproximación a la localización de fuentes sonoras reales.
Como muestra la figura de la página siguiente, dos altavoces izquierdo (L) y derecho (R) radian señales de audio coherentes, con posiblemente diferente amplitud, hacia un oyente. El paneo consiste en dar más amplitud de señal (volumen) a un altavoz que a otro, para dar la sensación de que la fuente proviene de la dirección de ese altavoz con mas volumen. Para N altavoces, las señales de alimentación para cada altavoz se definen

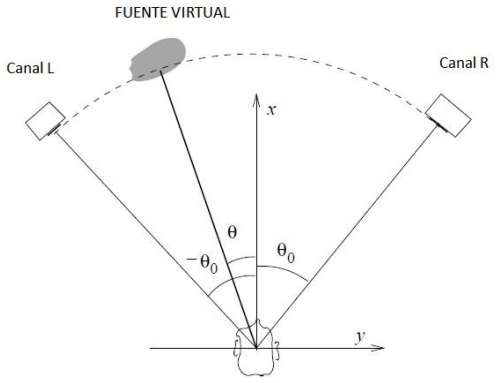
$$\mathbf{x_i(t) = g_i \cdot s(t)}, \qquad \mathbf{i = 1, 2, \dots, N}$$

El oyente percibe la ilusión de un solo *evento auditivo (fuente sonora virtual o fantasma)* que puede ser posicionada en un sector bidimensional definido por las posiciones de los altavoces y el oyente, controlando las amplitudes de cada señal. No obstante, esta técnica restringe el área de posicionamiento de las fuentes virtuales al área enfrente del oyente. Nunca se podrá conseguir sólo con un simple paneo de amplitud (sin filtrar la señal) que la sensación de posicionamiento este detrás del oyente sin tener que mover o añadir altavoces.
Como se muestra, se posicionan dos altavoces simétricamente con respecto al plano mediano. Las amplitudes se controlan con los factores de ganancia g1 y g2. Los altavoces se posicionan típicamente de forma simétrica, en un angulo $\theta o = 30º$ con respecto al oyente.

La dirección o posicionamiento de la fuente virtual depende de la relación entre amplitudes. Si la fuente se quisiera percibir como moviéndose pero con volumen constante, los factores de ganancia tendrían que normalizarse. Definiendo la potencia sonora mediante la constante C, la normalización requiere que siempre se cumpla:

$$\mathbf{g_1}^2 + \mathbf{g_2}^2 = \mathbf{C}, \qquad (\mathbf{g1}, \mathbf{g2}) \in [\mathbf{0}, \mathbf{1}]$$

El parámetro $C > 0$ puede considerarse como un control de volumen de la fuente virtual. La percepción de la distancia a la que se encuentra la fuente virtual depende, en mayor o menor grado, de C; cuanto más volumen tenga el sonido, más cerca será presentada la fuente virtual.



Para controlar la distancia de la fuente sonora virtual al oyente se deben tener en cuenta efectos psicoacústicos y se deben añadir otros elementos como reflexiones y reverberación. Esto se discute en las secciones posteriores. En el caso de que la distancia no se tenga en cuenta, la fuente se puede posicionar en un arco entre los altavoces; el radio queda definido por la distancia entre oyente y altavoces. Este arco se llama *arco activo* y en la figura está representado por un arco punteado entre ambos altavoces.
El cálculo de las ganancias {g1, g2} puede realizarse de varias formas, como se describe a continuación.

Formulación trigonométrica

La percepción de la dirección de una fuente sonora virtual producida por el método de paneo en amplitud sigue, aproximadamente, la *ley de senos estereofónica:*

$$\mathbf{sin}(\boldsymbol{\theta}) = \frac{\mathbf{g1} - \mathbf{g2}}{\mathbf{g1} + \mathbf{g2}} \cdot \mathbf{sin}(\boldsymbol{\theta o})$$

con $0º < \theta o < 90º$ y $-\theta o \leq \theta \leq \theta o$. Aquí, θ representa el ángulo de la dirección percibida de la fuente virtual y θo es el ángulo entre el eje central a los altavoces (se suponen simétricos).

Esta ecuación es válida pero solo si el oyente está mirando directamente al frente. Si el oyente gira su cabeza siguiendo a la fuente virtual, entonces es mas apropiada la *ley tangente*:

$$\frac{\mathbf{tg}(\boldsymbol{\theta})}{\mathbf{tg}(\boldsymbol{\theta o})} = \frac{\mathbf{g1} - \mathbf{g2}}{\mathbf{g1} + \mathbf{g2}}$$

con $0º < \theta o < 90º$ y $-\theta o \leq \theta \leq \theta o$.
Las dos leyes seno y tangente fueron desarrolladas restringiendo el problema a que el sonido solo difiriera en amplitud. Esto es valido para audio con frecuencias por debajo de los 600 Hz. Los factores de ganancia no se pueden resolver usando únicamente la ley tangente.
Para ser capaces de obtener los factores de ganancia, podemos formular una ecuación general para N altavoces que mantenga constante el volumen percibido de la fuente virtual:

$$\sqrt[\mathbf{k}]{\sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{g_n}^{\mathbf{k}}} = \mathbf{C}$$

Aquí, k se puede elegir según las características acústicas de la sala. En salas reales con algo de reverberación se escoge normalmente k = 2, para preservar la energía de la amplitud de las señales de la fuente virtual. Un valor k = 1 preservaría la amplitud y es más adecuado para situaciones de escucha en entornos anecoicos. Cuando se quiere mantener el nivel de potencia sonora constante, los factores de ganancia se resuelven usando una de las dos ecuaciones junto con la condición $\mathbf{g_1}^2 + \mathbf{g_2}^2 = \mathbf{C}$.

En principio, el método de paneo en amplitud crea un efecto de filtro peine (comb filter) en el espectro del audio ya que el mismo sonido llega de ambos altavoces a cada oído. No obstante, este efecto es relativamente suave y cuando se escucha en una sala normal, la reverberación de ésta suaviza lo suficiente la coloración del espectro debida al filtro peine. El que la coloración no sea muy fuerte y que el efecto direccional que provee el método de paneo en amplitud es bastante robusto, hacen que el método de paneo en amplitud se incluya en cualquier consola de mezclas habitual. Probablemente, sea la técnica mas usada para posicionar fuentes virutales.

En el siguiente ejemplo se muestra cómo hacer un paneo estéreo usando la ley tangente.

```
% stereopan.m
% Source: "DAFx", 2nd ed., cap. 5, Zolzer
% Paneo en amplitud estereo usando la ley tangente
function sig_altavoz = stereopan(signal, theta, theta_0, fs)

%fs = 44100;
%theta = -30;
%theta_0 = 30; % apertura de los altavoces

%Calculo en radianes
theta   = theta/180*pi;
theta_0 = theta_0/180*pi;

%Calculo de las ganancias usando la ley tangente
g(2) = 1; %el valor inicial ha de ser uno
g(1) =  -(tan(theta)-tan(theta_0))/(tan(theta)+tan(theta_0));

%Normalizacion de la suma de cuadrados
g       = g/sqrt(sum(g.*2));

%Señal a modificar
%signal = mod([1:20000]',200)/200;

%PANEO EN AMPLITUD
sig_altavoz = [signal*g(1) signal*g(2)];

%Reproducir audio
%soundsc(sig_altavoz, fs);
```

Formulación VBAP

El esquema del VBAP se aprecia en la siguiente figura y es igual que el de una configuración estéreo de dos canales. En el VBAP, el método de paneo de amplitud es reformulado usando bases vectoriales. Esta reformulación consigue ecuaciones sencillas para el paneo y el uso de vectores hace que estos métodos se puedan implementar con un costo computacional reducido. La base se define por los vectores de longitud unitaria

$$\overrightarrow{\mathbf{l_1}} = [\mathbf{l_{11}}\ \mathbf{l_{12}}]^{\mathbf{T}}, \qquad \overrightarrow{\mathbf{l_2}} = [\mathbf{l_{21}}\ \mathbf{l_{22}}]^{\mathbf{T}}$$

que apuntan hacia los altavoces desde el oyente. El vector que apunta en dirección a la fuente virtual $\vec{\mathbf{p}} = [\mathbf{p_1}\ \mathbf{p_2}]^{\mathbf{T}}$ se puede tratar como una combinación lineal de los otros vectores ((g1, g2) > 0):

$$\vec{\mathbf{p}} = \mathbf{g_1} \cdot \overrightarrow{\mathbf{l_1}} + \mathbf{g_2} \cdot \overrightarrow{\mathbf{l_2}}$$

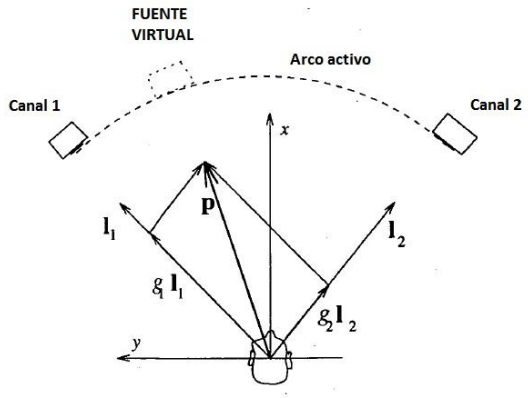
Esta ecuación se puede reescribir como:

$$\vec{\mathbf{p}}^{\mathbf{T}} = \vec{\mathbf{g}} \cdot \mathbf{L_{12}}$$

con $\mathbf{L_{12}} = [\overrightarrow{\mathbf{l_1}}, \overrightarrow{\mathbf{l_2}}]^{\mathbf{T}}$.

La ecuación anterior puede resolverse en el caso de que exista la inversa de **L**:

$$\vec{\mathbf{g}} = \vec{\mathbf{p}}^{\mathbf{T}} \cdot \mathbf{L_{12}}^{-1} = [\mathbf{p1}\ \mathbf{p2}] \begin{bmatrix} \mathbf{l_{11}} & \mathbf{l_{12}} \\ \mathbf{l_{21}} & \mathbf{l_{22}} \end{bmatrix}^{-1}$$



Además, la matriz inversa de **L** tiene que satisfacer $\mathbf{L_{12}} \cdot \mathbf{L_{12}}^{-1} = \mathbf{I}$ siendo $\mathbf{I}$ la matriz identidad. La inversa existe cuando siempre y cuando los altavoces estén posicionados con θo ≠ 0º y θo ≠ 90º (ambos además casos correspondientes a posiciones no muy relevantes). Para estos casos, VBAP 1-dimensional podría formularse (trivial). Los factores de ganancia, calculados mediante $\vec{\mathbf{g}} = \vec{\mathbf{p}}^{\mathbf{T}} \cdot \mathbf{L_{12}}^{-1}$, satisfacen la ley tangente. Cuando θo ≠ 45º, los factores de ganancia tienen que normalizarse usando:

$$\breve{\mathbf{g}} = \frac{\sqrt{\mathbf{C}} \cdot \vec{\mathbf{g}}}{\sqrt{\mathbf{g_1}^2 + \mathbf{g_2}^2}}$$

Ahora $\breve{\mathbf{g}}$ sí satisfará la condición

$$\breve{g_1}^2 + \breve{g_2}^2 = C$$

Un ejemplo de implementación en Matlab para VBAP en dos dimensiones se muestra a continuación.

```
% Calculo de ganancias para una configuracion VBAP en 2D
% Fuentes:
% DAFx: Digital Audio Effects, 2nd ed., Zoelzer, U.
%
% Computa las ganancias para VBAP-2D para una configuracion de altavoces
% horizontal. Las direcciones de los altavoces se dan en el sentido
% de las agujas del reloj.
function [gains] = VBAP2D(direction)


%Modifica estos valores para tu configuracion de altavoces
angles_loudsp = [30 -30 -90 -150 150 90];

N = numel(angles_loudsp);

angles_loudsp = [angles_loudsp angles_loudsp(1)]/180*pi; %angulos en radianes
% Se añade el angulo primero al final para el calculo de la matriz inversa.

%Direccion de paneo en coordenadas cartesianas (x, y)
vector_pan = [cos(direction/180*pi) sin(direction/180*pi)];

for i = 1:N
        % Calcular inversa de la matriz de altavoces
        Lmn_1 = inv([[cos(angles_loudsp(i)) sin(angles_loudsp(i))];...
        [cos(angles_loudsp(i+1)) sin(angles_loudsp(i+1))]]);

        % Hallar ganancias (no normalizadas)
        g_temp = vector_pan* Lmn_1;

        % Si las ganancias son no negativas, normalizarlas y terminar
        if(min(g_temp) > -0.001)
                g = zeros(1, N);
                g([i mod(i, N)+1]) = g_temp;
                gains = g/sqrt(sum(g.^2));
                return
        end
end
```

Formulación para configuraciones con más de 2 altavoces (VBAP 3D)

Una configuracion de altavoces en 3D es un sistema en el cual todos los altavoces no están en el mismo plano que el oyente. Esto suele implicar que haya algún altavoz más alto/bajo que la posición del oyente. Para estos sistemas, se puede usar un método de paneo por tripletas. Para cada instante, se aplica la señal de la fuente sonora a un máximo de tres altavoces que forman un triangulo si se mira desde la posición del oyente. Si la configuracion es de mas de tres altavoces, el sistema se divide en triángulos; cada uno de ellos será usado para el paneo de una fuente virtual en cada instante.

VBAP 3D es un método para formular la matemática que hay detrás de estas configuraciones de altavoces. Se formula equivalentemente al VBAP 2D discutido en las secciones anteriores. No obstante, el numero de factores de ganancia y altavoces es ahora tres. El angulo entre el plano mediano y la fuente virtual se suele estimar correctamente con VBAP para la mayoría de los casos. La percepción de elevación es, no obstante, individual para cada oyente.

Una implementación de VBAP 3D se da a continuación.

```
% Calculo de ganancias para una configuracion VBAP en 3D
% Fuentes:
% DAFx: Digital Audio Effects, 2nd ed., Zoelzer, U.
%
% Computa las ganancias para VBAP-3D para una configuracion de altavoces.
% Las direcciones de los altavoces se dan en el sentido de las agujas del reloj.
function [gains] = VBAP3(direction)


%Modifica estos valores (theta, phi) para tu configuracion de altavoces
angles_loudsp = [0 0; 50 0; 130 0; -130 0; -50 0; 40 45; 180 45; -40 45];

N = size(angles_loudsp,1);

%Define los triangulos para los altavoces en (x, y, z)
triangles=[1 2 6; 2 3 6; 3 4 7; 4 5 8; 5 1 8; 1 6 8; 3 6 7; 4 7 8; 6 7 8];

%Bucle por todos los triangulos
for T = 1:size(triangles, 1)
        actual = angles_loudsp(triangles(T,:),:);

        %Direccion de paneo en coordenadas cartesianas
        cos_E = cos(direction(2)/180*pi);
        vector_pan(1:2) = [cos(direction(1)/180*pi)*cos_E sin(direction(1)/180*pi)*cos_E];
        vector_pan(3) = sin(direction(2)/180*pi);

        %Calcular las tres componentes para la matriz de altavoces para el
        %triangulo actual
        for i = 1:3
                cos_E = cos(actual(i,2)/180*pi);
                Lmn(i, 1:2) = [cos(actual(i,1)/180*pi)*cos_E ...
                 sin(actual(i,1)/180*pi)*cos_E];
                Lmn(i, 3) = sin(actual(i,2)/180*pi);
        end

        % Factores de ganancia para el triangulo actual
        g_temp = vector_pan*inv(Lmn);

        % Si las ganancias son no negativas, normalizarlas y terminar
        if(min(g_temp) > -0.01)
                g_temp = g_temp/sqrt(sum(g_temp.^2));
                g = zeros(1, N);
                gains(1,triangles(T, :)) = g_temp;
                return
        end
end
```

2.2.4. VBAP no unitario (NVBAP)

Sin entrar en detalles, hay numerosos experimentos que delatan que el VBAP produce fuentes virtuales que están desplazadas respecto a la posición deseada original. Una manera de mejorar VBAP es implementar *VBAP no unitario (NVBAP).*

Para este método, se añaden ganancias adicionales una vez se han calculado las ganancias primeras según el procedimiento anterior. Las nuevas ganancias se pueden reformular:

$$g^*_1 = \frac{g_1}{\lambda_1}, \qquad g^*_2 = \frac{g_2}{\lambda_2}$$

donde λ es un numero real positivo. Las ganancias g* deben normalizarse usando la ecuación

$$\sqrt[k]{\sum_{n=1}^{N} g^*_n{}^k} = 1$$

para mantener el volumen constante. Usando las bases de vectores usadas anteriormente, las ganancias g* se pueden reformular como:

$$\left[\mathbf{g^*_1}\boldsymbol{\lambda_1} \quad \mathbf{g^*_2}\boldsymbol{\lambda_2}\right] = \vec{\mathbf{p}} \cdot \mathbf{L_{12}^{-1}}$$

que también puede expresarse como:

$$\vec{\mathbf{p}} = \mathbf{g^*_1} \cdot \overrightarrow{\mathbf{l_1}} \cdot \boldsymbol{\lambda_1} + \mathbf{g^*_2} \cdot \overrightarrow{\mathbf{l_2}} \cdot \boldsymbol{\lambda_2}$$

Para obtener las nuevas ganacias la ecuación es finalmente:

$$\overrightarrow{\mathbf{g^*}} = \vec{\mathbf{p}}^{\mathbf{T}} \cdot \boldsymbol{\Lambda}_{\mathbf{12}}^{-\mathbf{1}}$$

con $\overrightarrow{\mathbf{g^*}} = [\mathbf{g^*_1} \ \mathbf{g^*_2}]$ y $\boldsymbol{\Lambda} = [\overrightarrow{\mathbf{l_1}} \cdot \boldsymbol{\lambda_1} \ \overrightarrow{\mathbf{l_2}} \cdot \boldsymbol{\lambda_2}]^{\mathbf{T}}$. Esto puede verse como una modificación del VBAP original donde los vectores en la base vectorial ya no tienen longitud unitaria. La variable λ introduce una longitud a cada vector unitario de altavoz $\vec{\mathbf{l}}$. Esto se llama VBAP no unitario o *NVBAP*.

Para un par de altavoces, VBAP produce ganancias que cambian simétricamente con el punto medio entre altavoces. Las ganancias calculadas con NVBAP tienen un cociente distinto al de VBAP.

Para calcular los valores λ, se crea una función f dependiente de la dirección de los altavoces que produzca los λ que compensen el desplazamiento de la fuente virtual paneada en amplitud (el angulo es el del cono de confusión, ver sección 2.4.2.1):

$$\boldsymbol{\lambda} = \mathbf{f}(\boldsymbol{\theta}_{\mathbf{cc}})$$

Esto resulta en λ iguales en cada cono de confusión, lo que también implica simetría frontal/atrás. Además, en casi todas las configuracion se querrá también simetría izquierda/derecha, luego:

$$\mathbf{f}(\boldsymbol{\theta}_{\mathbf{cc}}) = \mathbf{f}(-\boldsymbol{\theta}_{\mathbf{cc}})$$

Y para ser consistentes con la ley tangente:

$$\mathbf{f}(\boldsymbol{\theta}_{\mathbf{cc}} = \pm \mathbf{30^{o}}) = \mathbf{1}$$

Esto, junto con otras consideraciones, deriva en un armónico esférico de primer orden:

$$\boldsymbol{\lambda} = \mathbf{f}(\boldsymbol{\theta}_{\mathbf{cc}}) = \mathbf{a} \cdot \mathbf{sin}(\boldsymbol{\theta}_{\mathbf{cc}}) + \mathbf{b} \cdot \mathbf{cos}(\boldsymbol{\theta}_{\mathbf{cc}}) \cong \mathbf{1.11} \cdot \mathbf{sin}(\boldsymbol{\theta}_{\mathbf{cc}}) + \mathbf{0.6} \cdot \mathbf{cos}(\boldsymbol{\theta}_{\mathbf{cc}})$$

2.2.5. Paneo en tiempo

Si se aplica un retardo constante a uno de los altavoces en la escucha estereofonica, la fuente virtual será percibida como proveniente del altavoz que radie la señal antes. Conforme el retardo se aproxime a 1 ms o más, el efecto tendera a su máximo (asintótico).
No obstante, diversos experimentos confirman que la dirección percibida para la fuente virtual depende de la frecuencia. Las características binaurales generadas varían con la frecuencia, y diferentes características sugieren diferentes direcciones de fuentes virtuales.

Por ello, este tipo de paneo puede generar una percepción "dispersa" de la dirección de sonido (que no obstante es interesante en ciertos casos).
El efecto también depende de la posición de escucha. Por ejemplo, si la señal de uno de los altavoces se retarda 1 ms, el oyente puede compensar este retardo moviéndose unos 30 cm hacia el altavoz con el retardo.

A continuación se muestra cómo cambiar la posición de una fuente sonora mediante la inclusión de un retardo en uno de los dos canales.

```
% delaypan.m
% Source: "DAFx", 2nd ed., cap. 5, Zolzer
% Crea una imagen virtual espaciada retardando uno de los canales
function sig_altavoz = delaypan(signal, tao, fs)

%fs = 44100;
%tao = 0.005; %retardo en segundos
tao_muestras = round(tao*fs);

%signal = mod([1:20000]',400)/400;
%Fade in
%signal(1:2000)=signal(1:2000).*[1:2000]'/2000;

%Retardar canal 1
sig_altavoz = [zeros(tao_muestras,1); signal(1:end-tao_muestras) signal];

%Reproducir
%audioplay(sig_altavoz, fs);
```

Un caso especial de paneo en tiempo en la reproducción estereofónica mediante altavoces es usar señales en contrafase. En esta técnica, la misma señal se aplica a ambos altavoces, pero la polaridad de una de las señales es invertida, creando una diferencia de fase entre las señales de 180º en todo el espectro frecuencial. Esto cambia la coloración del sonido percibida y dispersa las fuentes sonoras virtuales. La implementación de este método en Matlab® es muy básico y se omite la descripcion.

Otra pregunta a hacerse es cómo distribuir o dispersar la fuente virtual entre los altavoces disponibles. Para ello se modifica el espectro del sonido de manera distinta para diferentes frecuencias. Una forma es convolucionando la señal de audio para altavoces con dos ráfagas de ruido AWGN. Otra manera es aplicar retardos distintos para diferentes frecuencias. Esto esparce la fuente virtual entre los altavoces. El efecto es audible en un área de escucha amplia. El inconveniente es que este procesamiento cambia ligeramente la respuesta temporal, lo que puede ser percibido como emborronamientos o comportamiento caótico en los transitorios de la señal.
A continuación se muestra un ejemplo en Matlab® de cómo implementar una funcion para "esparcir" una fuente virtual entre N altavoces.

```
% Esparcir sonido por canales
% Fuentes:
% DAFx: Digital Audio Effects, 2nd ed., Zoelzer, U.
function loudsp_sig = spreadnoise(signal, Fs, Nchan)

% Generacion de rafagas de AWGN para todos los canales
ruido   = rand(round(0.05*Fs), Nchan) - 0.5;
% Convolucionar la señal con el ruido
loudsp_sig = conv(signal, ruido(:,1));
for i = 2:NChan
        loudsp_sig = [loudsp_sig conv(signal, ruido(:,1))];
end

loudsp_sig = loudsp_sig/max(max(loudsp_sig))*0.9;
audiowrite('spreaded.wav', Fs, [loudsp_sig]);
```

2.2.6. Auriculares para la reproducción de 5.0 Surround

Una aplicación interesante de las HRTFs es la escucha de material de audio multicanal con auriculares (sólo dos canales). En estos casos, cada altavoz de la configuracion multicanal se simula usando un par de HRTFs. Por ejemplo, una señal pensada para ser reproducida y percibida a 30º se convolucionara con el par de HRTFs medidas desde esa misma direccion. Las señales resultantes de la convolucion serán suministradas a los auriculares. El uso de HRTFs medidas en una cámara anecoica es suboptimo; el uso de BRIRs es mas beneficioso. Esto se puede hacer usando BRIRs medidas o simulando la reverberación mediante algún sistema.

Para el subwoofer (.1), bastaría con diseñar un filtro de audio para obtener las frecuencias más relevantes para el altavoz de graves elegido y alimentar a éste con el audio resultante.

Un ejemplo de escucha con altavoces 5.0 virtuales se da a continuación para Matlab®.

```
%% Reproduccion virtual de una señal 5.0 Surround con auriculares usando HRIRs
% Fuentes:
% DAFx: Digital Audio Effects, 2nd ed., U. Z?lzer
%
% sistema de referencia usado:
% el angulo azimutal de +0º a +180º en sentido de las agujas del reloj
% theta = 0 indica una fuente justo enfrente del oyente
% Usar el angulo -theta  para el oido derecho (o viceversa).
function out = virtual50(signal_50, angles_loudsp, Fs)

%Fs = 44100;

% Generar una señal 5.0 Surround como ejemplo
%cnt = [0:20000]';
%signal = [(mod(cnt,200)/200) (mod(cnt,150)/150) (mod(cnt,120)/120)...
% (mod(cnt,90)/90) (mod(cnt,77)/77)];

i = 1;

% Bucle por los canales de entrada
L = 0;
R = 0;
for angles_loudsp = [30 -30 -110 110 0]
        HRIR_L=simpleHRIR(angles_loudsp, Fs);
        HRIR_R=simpleHRIR(-angles_loudsp, Fs);
        L = L + fconv(HRIR_L, signal(:,i));
        R = R + fconv(HRIR_R, signal(:,i));
        i = i + 1;
end

out = [L R];

%Sound output to headphones
%soundsc(out, Fs)
```

2.2.7. Técnicas binaurales con altavoces mediante cancelación de Crosstalk

El material de audio binaural está pensado para ser reproducido de tal forma que el sonido que se origina para el oído izquierdo sea reproducido solo en el oído izquierdo y de manera análoga para el oído derecho. Si una grabación asi se reproduce con una configuracion de altavoces estéreo, el sonido del altavoz izquierdo también llegara al oído derecho (llamado *componente de Crosstalk*) y de manera análoga para el altavoz derecho, arruinando el efecto 3D conseguido mediante el procesamiento por tramas y convolucion con las HRTFs. Para proder reproducir audio binaural con dos altavoces, hay que usar métodos especiales, que se discuten detalladamente en las secciones 2.6 y 2.7. En ellos, se intenta alimentar los altavoces con señales tales que las componentes de Crosstalk se cancelen lo mas posible.

2.2.8. Escucha de material estereofónico de dos canales mediante altavoces y auriculares

La escucha mediante altavoces es diferente de la escucha mediante auriculares. En los auriculares, no existe el Crosstalk presente en la escucha mediante altavoces. Normalmente, los ingenieros crean el contenido de audio estéreofonico en estudios usando altavoces. Es importante preguntarse qué cambia en la percepción del contenido de audio si la escucha es con auriculares.

Con material que ha usado VBAP, la diferencia de nivel entre los canales de los auriculares se corresponde directamente con la ILD. La ITD permanece a cero. Esto es muy diferente de la escucha mediante altavoces, en donde las direcciones de las fuentes, paneadas en amplitud, dependen de las ITDs. En este otro caso, la ILD es cero para frecuencias bajas.
No obstante y curiosamente, la imagen espacial del audio es similar en los dos casos. No obstante, en el caso de auriculares existe el llamado efecto de "localización dentro de la cabeza" mientras que para altavoces las fuentes se perciben en un área comprendida entre los dos altavoces.

Este efecto de internalización se debe a dos factores: las características dinámicas hacen que las fuentes se escuchen como internas ya que las ITDs e ILDs no cambian cuando el oyente se mueve. El otro factor es que las características monoaurales tampoco permiten producir el efecto de fuentes exteriorizadas, ya que estas características monoaurales difieren mucho de las que se originan cuando se escucha con altavoces.

2.2.9. Sistemas de audio 3D actuales y formatos de reproducción

Aunque el estado del arte del audio 3D esté bastante avanzado, la expresión "sonido Surround" sique significando sonido horizontal para la mayoría de los consumidores y profesionales de la música y cine. Los formatos mas usados fuera del estéreo convencional son de hecho el 5.1 Surround y el 7.1 Surround.

El formato 5.1 usa un par de señales estero estándar izquierdo L y derecho R, una señal central C y dos canales traseros "Rear Left Surround" Ls y "Rear Right Surround" Rs, además de un canal para las bajas frecuencias "Low Frequency Effects" LFE mediante un subwoofer o altavoz de graves dedicado.

La configuración de la posición de los altavoces, especificada en el estándar ITU-R BS 775, establece posicionar los cinco canales a la misma distancia del oyente en circulo, en angulos azimutales $\pm 30º$ para los canales L y R, $0º$ para el canal C y $\pm 110º$ para los canales Ls y Rs. Esta es la configuración recomendada para la audición de música, mientras que para la de audio para cine los canales Ls y Rs se distribuyen por multiples altavoces uniformemente posicionados a lo largo y ancho de las paredes del teatro o cine.

La configuracion 7.1 para música añade dos canales Surround traseros cuyas posiciones se recomiendan entre $\pm 135º$ y $\pm 150º$, mientras que los canales Ls y Rs del 5.1 Surround se desplazan a posiciones entre $\pm 90º$ y $\pm 110º$. Para salas de cine y teatro, 7.1 Surround divide las líneas Ls y Rs existentes en dos, obteniendo cuatro canales para acomodar canales "Back Left Surround" y "Back Right Surround".

El formato 5.1 es parecido a una extensión del estéreo en el plano horizontal, pensado para contenido de audio de películas: la presencia del canal central C permite fijar los diálogos en la pantalla; el par (L, R) se usa para mover sonidos en la parte frontal, para contenido estéreo y para permitir la compatibilidad con estándares anteriores; el par trasero (Ls, Rs) se usa para proveer ambiente y sonido envolvente, con efectos de sonido ocasionales. La posición de los auriculares, en particular si se contempla todo el área mas cercana a la pantalla a expensas de una densidad menor en la parte trasera, produce una desadaptación en la precisión del posicionamiento del sonido para las posiciones traseras. Los angulos muy abiertos entre los altavoces traseros causan "huecos" percetibles en el campo sonoro cuando los sonidos estaban pensados para un área fuera del de la pantalla.
La posición de los altavoces traseros esta pensada para solventar deficiencias de nuestro sistema audiivo, que es peor posicionando sonidos provenientes de la parte trasera de la cabeza. Esto ayuda a la reproducción de campos sonoros difusos y efectos especiales, pero no ayuda a crear un escena sonora compleja y completa que envuelva completamente al oyente en 3D. Curiosamente, este limitación debido a las necesidades del audio para películas termino imponiendo limitaciones en el argot del mundo cinematográfico; ya que los sistemas no permiten un paneo preciso fuera del área de la pantalla, este concepto se prohibio explícitamente para los diseñadores e ingenieros de post-produccion. Los canales Surround han quedado relegados a recrear sonidos ambiente, reverberación y algún efecto especial.

Aunque no se haya extendido en el mercado, la tecnología de audio 3D se ha usado desde hace ya tiempo en la música electroacústica, donde el espacio se usa por los compositores como una herramienta artística y se le da misma importancia que al tono o al ritmo. Hay impresionantes instalaciones y trabajos de audio 3D de compositores como Egar Varèse, Karlheinz Stockhausen o Leo Kupper.

Recientemente, el potencial del sonido 3D se ha explotado para acompañar a la proyección semiesférica de exhibiciones como expos mundiales, museos y parques de atracciones. Hay muchas aplicaciones software y herramientas para la producción de audio 3D en el mercado, pero pocas han alcanzado la eficiencia necesaria para convertirse en un éxito. Esto se debe probablemente a la ausencia de un estándar de facto para la distribución del formato y para la configuracion de los sitemas de reproducción. Por ahora los medios han usado sistemas con posiciones de altavoces fijas y canales de audio discretos; esto no es compatible con el 3D Surround, básicamente porque seria demasiado optimista asumir que cada consumidor posicionara de la misma y difícil manera todos los altavoces.

Recientemente, algunos autores han tomado un camino diferente que parece haber solventado este problema. Presentan un cambio de paradigma importante, donde el ingeniero de sonido deja atrás el problema de pensar en términos de canales de audio de salida y solo considera cada fuente sonora en términos de su nivel de señal y posición en el espacio. Uno de los mejores y más versátiles softwares de audio 3D es *SoundScape Renderer (SSR)*, desarrollado por el QU Lab de la universidad TU Berlín [48]. SSR es una plataforma para renderizar una escena sonora basada en objetos a varios formatos de salida, incluyendo incluso WFS, Ambisónica y binaural. Además, SSR usa un servidor basado en sockets TCP/IP para poder monitorizar, mediante una app instalada en un teléfono móvil, la escena sonora que está siendo procesada y renderizada por la maquina host (p.ej. un ordenador portátil) y poder reproducirla también mediante el dispositivo móvil [48].

No obstante, en los últimos años se han producido importantes mejoras en el ámbito del audio 3D orientado a la industria cinematográfica. Actualmente, hay compañías que ofrecen soluciones que incluyen por ejemplo la configuracion de altavoces y equipamiento espacial para producir y reproducir las bandas sonoras. Dolby®, immsound, Auro3D o iosono son ejemplos de algunas de estas compañías.

En la siguiente página se muestra un diagrama en forma de árbol que resume todos los efectos de audio que se pueden conseguir mediante técnicas digitales. En rojo se han marcado los efectos que conciernen a este proyecto fin de carrera.


Audio Effects
Timbre
Loudness
amplification
Timbre compressor
Timbre expander
Timbre noise gate
Timbre limiter
Timbre contrast
nuance change
Rhythm tremolo
Timbre spectral tremolo
Space
Timbre granular delay
echo Room
reverberation
distance
height
Localization
panning
azimuth
Timbre spec. panning
Loudness
Pitch Doppler
3D binaural
3D transaural
directivity change Directivity
Directivity
Localization Leslie / Rotary
Pitch
Pitch
Formants no formant preservation pitch-shifting
spectral warping
detune Harmonicity
autotune
harmonizer
Time resampling
Time prosody change
Loudness intonation change
Time
Timbre
vibrato preservation
formants preservation
tremolo preservation
attack preservation
time-scaling
Duration
Pitch resampling
swing change Rhythm
Timbre time-shuffling
Timbre inversion
timbral metamorphosis
Time
Loudness
Pitch
Formants
timbre morphing
gender change Pitch
spectral envelope warping
vocoder effect
cross synthesis
Pitch
Loudness
hybridization
spectral interpolation
mutation
spec. env. modifications
scaling Voice quality
shifting
warping
Spectrum
harmonics generator Pitch
subharmonics generator Pitch
scaling Pitch
shifting
Pitch
Harmonicity
ring modulation
Formants
spectral ring modulation
SSB modulation Formants
warping
Harmonicity
inharmonizer Harmonicity
Quality
denoising
declicking
enhancer Brightness
distorsion Pitch
fuzz
vibrato Pitch
voice quality
robotization Pitch
whisperization
hoarseness
martianization Pitch
centroid change Brightness
Filter
comb filter Pitch
resonant filter Pitch
telephon effect
chorus, flanger, phase, wah-wah

2.3. Trayectorias sonoras

El primer intento en diseñar movimientos en el espacio como dimensión independiente de una composición sonora (musical) fue probablemente el de Stockhausen con su "Gesang der Jünglinge" [31]. En esta obra, el autor programo especialmente los movimientos y direcciones de cinco grupos de altavoces dispuestos en forma de circulo en torno a la audiencia. Esta obra es probablemente también el origen de la "utopía espacial" de buena parte de la música contemporánea ya que Stockhausen dejo abundantes explicaciones sobre el tema, describiendo velocidades y desplazamientos en el espacio así como duración de notas y pitch.

Desde 1970, muchos sistemas de computadores han sido construidos con el fin de ayudar al artista en la composición de su pieza musical usando efectos espaciales. La mayoría de estos sistemas están basados en una interfaz software que permite al músico mostrar y/o definir trayectorias sonoras de fuentes virtuales en el espacio de escucha. Aún más, la manipulación de las trayectorias sonoras pertenece a una capa de control construido justo encima del nivel de procesamiento de señal que implementa los efectos panorámicos o de distancia. Parámetros como la posición angular, la distancia a los altavoces, etc. son calculadas de la trayectoria a una tasa suficientemente alta como para permitir movimientos suaves y continuos y sin ruidos o distorsiones no naturales. Con respecto a este ultimo punto, si las trayectorias solo afectan a los efectos de distancia y de panorama, la tasa de control puede ser muy baja (unos 20 Hz), permitiendo la separación de la capa de control de la capa de procesamiento de señal mediante un canal de comunicación de bajo ancho de banda. Y viceversa, si el efecto Doppler se tiene en cuenta de alguna manera, la tasa de control tiene que ser mucho mayor que 20 Hz, porque el oído humano es sensible a variaciones minúsculas en el "pitch" del audio. Todavia se podrían transmitir las señales de control a una tasa baja si en el nivel de procesamiento de señal hubiera un mecanismo de interpolación que reconstruyera los valores intermedios de los parámetros de control.

Algunas veces, las trayectorias espaciales son una metáfora para la transformación guiada de material de audio. Por ejemplo, un autor usa epiciclos múltiples para controlar las transformaciones de sonidos continuos y proyectarlos en el espacio. La consistencia entre display espacial y transformación de sonido permite al oyente discriminar y seguir diferentes piezas musicales a la vez sin la barrera de patrones rítmicos o de pitch.

Los sistemas de audio software para la definición y proyección de trayectorias sonoras son bastante usados en la música electrónica en directo. Belladonna y Vidolin desarrollaron una interfaz para el control panorámico 2D para el lenguaje visual MAX [32]. Las trayectorias se pueden memorizar y usar y transformar en tiempo real a mensajes MIDI a enviar a un mixer de audio MIDI. También se pueden controlar y sincronizar dispositivos externos para reverb para recrear la impresión de profundidad y distancia. Este sistema ha sido usado en varias producciones musicales y en distintos espacios de escucha. En general, las trayectorias sonoras raramente son significantes per se, sino que pueden formar parte de un componente dramático crucial si sus variaciones son controladas de forma dinámica como función de otros parámetros musicales como la intensidad o el ritmo.

En Paris se desarrolló por parte del instituto IRCAM en los años 1990 un sofisticado sistema para procesamiento espacial de audio multicanal en tiempo real. En este sistema, se diseño un nivel intermedio entre el de procesamiento de señal y el de control. Este nivel intermedio contiene los parámetros físicos de las salas virtuales, que son simulados mediante un conjunto de atributos perceptuales como la calidez, el brillo, etc. de tal manera que el control sobre la composición espacial resulta más sencillo. Ya que la interfaz se desarrolló para el lenguaje visual MAX, es sencillo asociar conjuntos de parámetros a trayectorias en un espacio 2D.

2.4. Ambisónica

La Ambisónica se basa en una técnica microfónica especial; surgió en los años 1970 como una manera de codificar toda la información relevante a la grabación de propiedades espaciales del sonido en un único punto del espacio y la subsecuente decodificación mediante configuración de altavoces adecuadas. Después del audio binaural, fue la primera tecnología que tuvo en cuenta la componente vertical de los campos sonoros. No obstante también puede ser simulada para sintetizar audio espacial para grupos de altavoces en configuraciones 2D o 3D. En este caso se convierte en un método de paneo en amplitud. La señal de sonido sirve de alimentación para todos los altavoces (posicionados equiespaciados alrededor del oyente) pero con diferentes ganancias {gi} dadas por un método de paneo en amplitud:

$$\mathbf{g_i} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{m=1}}^{\mathbf{M}} (\mathbf{1} + \mathbf{2} \cdot \mathbf{p_m} \cdot \mathbf{Y_{mn}}^{\boldsymbol{\sigma}}(\boldsymbol{\theta}, \boldsymbol{\varphi}))$$

Donde gi es la ganancia del altavoz i-ésimo, N es el número total de altavoces, θ es el ángulo entre el altavoz y la dirección virtual o dirección de paneo, $\{\mathbf{Y_m}^{\boldsymbol{\sigma}}\}$ representan los armónicos esféricos de orden n, M es el orden de la Ambisónica y p_m son las ganancias para cada armónico esférico.

Este método difiere del método de paneo en amplitud en tanto en cuanto los factores (g1, g2) pueden tomar valores menores que cero. Cuando se codifica un campo decodificado, las señales en anti fase en un canal son aplicada a los altavoces en una dirección negativa respecto al eje. La fase de decodificación se hace con ecuaciones matriciales [33].

Cuando el orden M es bajo, la señal de sonido s(t) emana de todos los altavoces causando algunos efectos no deseados debidos al comportamiento no natural de las características binaurales.

Cuando aumenta el orden, el Crosstalk entre altavoz se puede minimizar optimizando las ganancias para los armónicos esféricos de cada altavoz de la configuración. El uso de armónicos de mayor orden aumenta tanto la calidad direccional como la timbral de la fuente virtual (ya que hay menos Crosstalk entre altavoces).

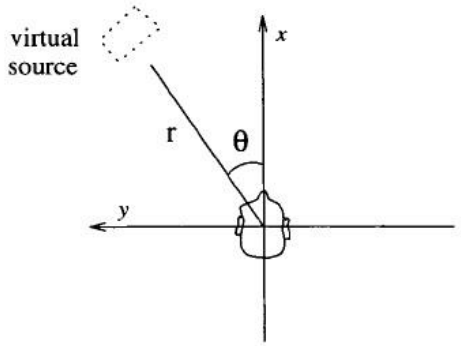
La Ambisónica se basa en la expansión del campo de onda en un punto en series de Fourier-Bessel; para un campo monocromático, después de manipular la ecuación de ondas en coordenadas esféricas, la presión se calcula como:

$$\mathbf{p}(\vec{\mathbf{r}}, \boldsymbol{\omega}) = \sum_{\mathbf{m=0}}^{\infty} \mathbf{i^m} \cdot \mathbf{j_m}(\mathbf{kr}) \sum_{\mathbf{0 \le n \le m}} \mathbf{A_{mn}}^{\boldsymbol{\sigma}}(\boldsymbol{\theta}, \boldsymbol{\varphi}) \cdot \mathbf{Y_{mn}}^{\boldsymbol{\sigma}}(\boldsymbol{\theta}, \boldsymbol{\varphi})$$

siendo $k = \omega / c$, $\{\mathbf{Y_{mn}}^{\boldsymbol{\sigma}}\}$ los armónicos esféricos y $\mathbf{j_m}(\mathbf{kr})$ las funciones de Bessel de primera especie. $\{\mathbf{A_{mn}}^{\boldsymbol{\sigma}}\}$ son los coeficientes de la expansión, que definen las propiedades 3D del campo de onda. En términos de magnitudes físicas, las componentes de orden 0 cooresponden a la presiones sonora, y las de orden 1 a las del gradiente de presión o componentes del vector velocidad acústico. Las componentes de mayor orden no tienen correspondencia física real; son combinaciones lineales de las derivadas del campo de presión sonora.

La primera formulación que se hizo de la Ambisónica se basó en las presiones acústicas escalares y tres componentes ortogonales del vector gradiente en un único punto del espacio, sin indicaciones que estas cantidades fueran usadas como términos de primer orden en la ecuación de expansión en series de Fourier-Bessel. El razonamiento para este método es que la percepción del sonido para un oyente depende de la presión y su gradiente (en particular, de su componente lateral para estéreo Surround u horizontal) a la altura de los oídos. En la parte de codificación (o grabación), la captura de estas componentes era el fin último de la Ambisónica. Se desarrollaron micrófonos para obtener cada componente por separado y corresponden a los transductores de presión omnidireccionales y los micrófonos de gradiente de presión bidireccionales. Ya que posicionar cuatro micrófonos en una

configuración coincidente, tres de ellos orientados en direcciones ortogonales, no era posible sin introducir interferencias acústicas graves entre los micrófonos, se pensó en una elegante solución que obtenía las señales indirectamente, de combinaciones de transductores de presión y gradiente de presión. El dispositivo resultante fue un micrófono tetraédrico, cuyos principios y características además de otras tecnologías de audio 3D se pueden leer con más detalle en [34].
Una vez se han conseguido obtener las componentes del campo de onda acústico, este será reproducido con altavoces para reproducir el campo sonoro original. Un decodificador de Ambisónica es un dispositivo que combina las señales acústicas ambisónicas disponibles de manera adecuada y a su salida ofrece las señales de alimentación de la configuracion de altavoces para reproducir correctamente el campo.



Hay dos maneras de decodificar Ambisónica, la manera "física" y la manera "psicoacústica" [34].

El fin de la decodificación según la manera "física" es, dada una cierta configuración de altavoces, combinar las señales ambisónicas para que las componentes reproducidas (los coeficientes de la expansión A) en el punto de escucha tengan la máxima similitud con las originales. En este caso ideal, uno podría grabar un campo sonoro, decodificarlo y reproducirlo de nuevo con altavoces y grabarlo otra vez con el mismo micrófono usado para la grabación original para comparar la versión obtenida con la original.

La manera "psicoacústica" de decodificación dicta construir el decodificador de tal manera que optimice la reconstrucción de ciertos parámetros relacionados con nuestra percepción. Para conseguir esto, se identifican dos criterios para una optima reproducción de material ambisonico: la reconstrucción precisa del vector velocidad a bajas frecuencias y del vector energía a altas frecuencias. Estas conclusiones se deducen directamente de las teorías de Lord Rayleigh discutidas, que decían que la localización del sonido a bajas frecuencias esta relacionado con la diferencia de fase entre las señales (vector velocidad) meintras que en altas frecuencias la característica principal es la direccion principal de llegada de la energía. Por ello, los decodificadores de Ambisonica suelen ser de doble banda con frecuencia de corte 700 Hz, aplicando diferentes coeficientes para cada banda.

Ambas maneras de decodificar señales ambisonicas resultan en decodificadores lineales, donde la señal de alimentación generada para cada altavoz es una combinación lineal de señales ambisonicas (aunque en el decodificador psicoacustico las componentes se dividan en dos bandas de frecuencias y una combinación lineal distinta se use para cada banda).
Considerando un sistema de N altavoces con posiciones $\hat{\mathbf{u}}_\mathbf{i}$ relativas a un sistema de coordenadas con origen en el punto de escucha del oyente, cada altavoz reproduce una determinada señale con ganancia $\mathbf{g}_\mathbf{i}$. El vector velocidad se define

$$\vec{V} = \frac{\sum_{i=1}^{N} g_i^2 \ \hat{u}_i}{\sum_{i=1}^{N} g_i} = \left|\vec{V}\right| \cdot \hat{u}_V$$

siendo $\hat{\mathbf{u}}_{\mathbf{V}}$ el vector unitario en direccion del vector velocidad. El vector energía se define de manera similar:

$$\vec{E} = \frac{\sum_{i=1}^{N} g_i^2 \ \hat{u}_i}{\sum_{i=1}^{N} g_i^2} = \left|\vec{E}\right| \cdot \hat{u}_E$$

Según el criterio psicoacustico comentado anteriormente, la direccion percibida de la fuente será, a altas frecuencias, $\hat{\mathbf{u}}_{\mathbf{E}}$ y a bajas frecuencias será $\hat{\mathbf{u}}_{\mathbf{V}}$. El objetivo de un decodificador de Ambisonica es pues, dada una fuente de sonido provienente de una determinada direccion, reproducir ambos vectores V y E de manera alineada con la direccion de la fuente y también maximizar los modulos |V| y |E|. Para configuracion de altavoces ordinarias, con igual distancia angular entre un altavoz y sus vecinos, los rquerimientos para decodificación optima se ha demostrado se pueden satisfacer y existen dos conjuntos de coeficientes que maximizan los módulos por separado y que en ese caso las direcciones de ambos vectores coinciden.

Aunque relacionadas, las partes de grabación y reproducción para Ambisonica son problemas separados que se suelen desarrollar con métodos diferentes. Como resumen, la grabación de señales ambisonicas intenta obtener las señales correspondientes a los armónicos esféricos de cierto orden mediante configuraciones adecuadas de micrófonos, mientras que la decodificación intenta reconstruir las cantidades grabadas ajustando las señales físicas o modificándolas para que cumplan los criterios psicoacústicos.

El punto fuerte de la Ambisonica es la completa codificación de campos 3D en un conjunto relativamente pequeño de señales. Para orden 1, las características espaciales del campo sonoro están contenidas en cuantro canales, mientras que para orden N el numero de canales es $(N+1)^2$ que es el numero de armónicos esféricos hasta orden N. Por otro lado, para ordenes bajos la precisión de localización obtenida es regular. Cuando se decodifican señales ambisonicas de orden 1 o 2 para una configuracion de altavoces dada, la mayor parte de los altavoces participaran en la reproducción del campo y reproducirán la señal con bastante intensidad. Mientras que esto renderiza una correcta reconstrucción en el punto de escucha (sweet spot), tan pronto el oyente se mueve fuera del punto de escucha el sonido tiende a ser percibido como proveniente del altavoz mas cercano; por ello conseguir la localización precisa en lugares fuera del punto optimo de escucha es un problema complejo.

2.5. Surround virtual usando auriculares. El estándar Dolby® Pro Logic™

2.5.1. Introducción a las tecnologías Dolby®

A comienzos de los años 1950 la 20th Century Fox desarrolla el primer formato para sonido multicanal. La combinación de formatos para pantallas gigantes como el CinemaScope 35mm y el Todd-AO 70mm con sonido multicanal fue la respuesta por parte de la industria cinematográfica a una creciente audiencia de TV. El sonido estereofónico para cine se solía reproducir con tres altavoces frontales, pero estos nuevos formatos incorporaron un canal monofónico adicional que era reproducido mediante 2 altavoces traseros (el canal de efectos). Este canal de efectos aumentaba la sensación de espacialidad pero tenia la desventaja de tener los problemas de localización en la cabeza y la ley del primer fuente de onda, destruyendo así la sensación de envolvimiento deseada. La solución a estos problemas fue introducir un segundo canal que era reproducido mediante un array de altavoces a los lados del teatro para recrear un campo difuso mas real.

A mediados de 1970 los laboratorios Dolby introducen la nueva tecnología de audio Dolby® Stereo™. Se basó en la tecnología óptica usada para el sonido para películas a lo largo de esos años y eliminó los problemas asociadas a la grabación multipista en cintas magnéticas. Dolby desarrollo un método para codificar cuatro canales (L, R, C y Mono Surround) en forma matricial en dos canales. Usaron una técnica derivada de los métodos matriciales anteriormente usados en los sistemas cuadrafónicos pero permitía la retro compatibilidad con sonido mono y estéreo.

En 1992 se dio un paso adelante y se lanzo el formato Dolby Stereo Digital™ (SR·D), que eliminaba la codificación/decodificación matricial y que usaba cinco canales (L, R, C, Ls, Rs) en una configuración conocida como Stereo Surround 5.1. Un sexto canal (.1) para efectos de baja frecuencia (LFE) prevenía que los altavoces se saturaran a frecuencias inferiores a 120 Hz. El canal LFE se limito de 0 a 120 Hz, pues es un ancho de banda que, para el ser humano, esta fuera del rango de localización en un entorno reverberante. Esto simplifico el posicionamiento del subwoofer para LFE, ya que ahora daba igual donde situarlo; la posición no afectaría a la calidad de escucha final (en proporciones apreciables).

Aun con todo, la mayoría de consumidores (particularmente aquellos que usan ordenadores) encuentran el uso de múltiples altavoces algo no muy práctico o inaccesible. Por ello es importante el desarrollo de sistemas de audio 3D para eliminar los requerimientos de los sistemas más complejos, permitiendo usar solo dos altavoces mediante técnicas de DSP para simular campos acústicos tridimensionales.

El *sonido Surround o sonido envolvente* es un método de reproducir audio con un conjunto de altavoces posicionados alrededor del oyente. Esto provee el efecto de 3D en un plano 2D, ya que la posición virtual de la fuente puede seleccionarse para que provenga de cualquier parte alrededor del oyente ajustando correctamente las señales de alimentación de los altavoces. Hay varias configuraciones de sonido Surround actualmente estandarizadas y en el mercado. Algunos ejemplos son 4.1, 5.1 y 7.1 y 7.2 [ITU-R BS 775-1, 1994]. El numero que precede al decimal indica el número de altavoces y el decimal cuántos subwoofers se usan en la configuración. El subwoofer, también conocido como *canal de efectos de baja frecuencia (Low Frequency Effects LFE)*, radia las frecuencias mas bajas a una potencia mayor.

Para este sistema, cada altavoz de la configuración se simula usando un par de HRTFs. Por ejemplo: una señal que se aplicaría a un altavoz posicionado en una dirección de 30º con respecto al oyente se convoluciona con un par de HRTFs para esa dirección y las señales resultantes son aplicadas a los auriculares. En caso de querer un entorno a un más realista, el uso de BRIRs es beneficioso. Mezclando audio estéreo a 5.1 Surround y emulando los efectos de las HRTFs para el posicionamiento de los altavoces Surround, se puede conseguir 5.1 Surround simulado en auriculares.

Lo más complejo de un sistema de estas características es desarrollar el mezclador y combinar los efectos con los efectos 3D conseguidos de cualquier otra forma. Esto es, una grabación de audio estéreo de dos canales se mezcla hacia arriba a seis canales Surround diferentes y cada uno de esos canales debe entonces someterse a un proceso de espacialización para posicionar los altavoces en el sitio apropiado. El resultado debe entonces mezclarse hacia abajo (mediante un downmixer) otra vez a estéreo de dos canales.
Este proceso debería, en teoría, dar al oyente la experience de escuchar audio en sonido envolvente usando solamente auriculares.

Una vez se tiene el audio, el modelo deseado para las HRTFs y el audio procesado para 3D (de la forma o formas que fuera, p.ej. en este proyecto de manera que describe una trayectoria 3D), se puede empezar a mezclar el audio a 5.1.

Con este sección se pretende, primeramente, presentar primero una breve historia de la trayectoria del estándar Dolby®. Por otro lado se decribe un upmixer (mezclador hacia arriba) 5.1. desarrollado por [9]. Se implementa un upmixer para sonido 5.1 Surround que sigue el estándar de Dolby® Pro Logic™,líder en la tecnología de reproducción de audio multicanal.
El mezclador usa ganancias estáticas y dinámicas, efectos de retardo, filtrado e isolación del subwoofer.

Dolby® Stereo™ y Dolby® Surround™

El líder de la industria del sonido envolvente son los laboratorios Dolby Laboratories Inc. Cuyas tecnologías son usadas de manera habitual en la industria cinematográfica. Dolby® Stereo™ fue originalmente desarrollado en 1976 para sistemas de sonido para cine (analógicos).
Dolby® Surround™ fue la primera versión de Dolby® Stereo™ para decodificar audio multicanal analógico para uso doméstico. Fue lanzado en 1982. El término se acuñó para no confundir “home stereo” de dos canales con “theater stereo” de cuatro canales (Dolby® SR™. Pro Logic™ es una tecologia de procesamiento de sonido envolvente diseñada para decodificar bandas sonoras codificadas mediante Dolby® Surround™.
Si se desea producir audio de una banda sonora que implemente Stereo™ o Surround™ (banda sonora porque la mayoría de aplicaciones se encuentran en el ámbito de la industria cinematográfica), se codifican cuatro canales de audio (L, R, C y Surround) de forma matricial a dos canales (L y R). La información envolvente esta contenida dentro de fuentes estéreo como TVs, DVDs o broadcasts desde donde puede decodificarse mediante un DSP para recrear el Stereo™ o Surround™ original. Sin el decodificador, el audio se reproducirá de manera estándar, en estéreo o mono.
En la tabla siguiente se muestran las ganancias que se aplicaban para este estándar.

Dolby® Surround Matrix	Left	Right	Center	Surround
Left Total (Lt)	1	0	$\frac{\sqrt{2}}{2}$	$j\frac{\sqrt{2}}{2}$
Right Total (Rt)	0	1	$\frac{\sqrt{2}}{2}$	$-j\frac{\sqrt{2}}{2}$

La tecnología de decodificación Dolby® Surround™ fue actualizada a mediados de 1980 y renombrada Dolby® Pro Logic™. Hay que notar que los términos Stereo™/Surround™ se siguen usando para denotar audio (bandas sonoras) que se ha codificado matricialmente usando esta técnica.

Dolby® Pro Logic™ y Pro Logic™ II

Pro Logic™ es la versión home entertainment de la versión para cines Dolby® Stereo™. También se basa en una tecnología de decodificación matricial. El estándar actual se llama Dolby® Digital EX™, que mejora y sustituye al antiguo sistema Pro Logic™ añadiendo una extensión del códec de Dolby® Digital en forma de canales traseros codificados matricialmente, pudiendo crear 6.1 y 7.1. Pro Logic™ usa la siguiente matriz para su códec. Un desarrollo extenso sobre esta tecnología se puede encontrar en [10].

Dolby® Pro Logic™	Left	Right	Center	Rear
Left Total (Lt)	1	0	$\frac{1}{\sqrt{2}}$	$-j\frac{1}{\sqrt{2}}$
Right Total (Rt)	0	1	$\frac{1}{\sqrt{2}}$	$j\frac{1}{\sqrt{2}}$

En el año 2000, Dolby ® Pro Logic™ II sale al mercado como una mejora sobre el estándar anterior. Este formato procesa cualquier fuente de audio estéreo de calidad a cinco canales independientes (L, R, C, Lr y Rr). Este método también decodifica cinco canales de señales estéreo codificadas en Dolby® Surround™. Pro Logic™ II implementa mejoras notables sobre Pro Logic™ consiguiendo un campo sonoro excepcionalmente estable que simula sonido envolvente de cinco canales.
Debido a la naturaleza limitada de Pro Logic™, muchas empresas dedicadas al entretenimiento introdujeron sus propios circuitos de procesamiento, tales como modos “Jazz”, “Hall” y “Stadium” que se pueden encontrar en los receptores de audio más comunes. Pro Logic™ II consigue evitar esto usando circuitos servo (de realimentación negativa) que son usados para obtener cinco canales. El contenido del canal extra se extrae usando la diferencia entre el contenido de audio espacial de los dos canales estéreo individuales o canales codificados Dolby® Digital™. Ademas de cinco canales de reproducción, Pro Logic™ II incluye el modo “Music” que incluye retardos para los canales optimizados.
Pro Logic™ II también implementó un modo diseñado específicamente para los videojuegos y fue usado de manera regular en títulos para las consolas Sony® PlayStation™ 2, Nintendo® Game Cube™ y Nintendo® Wii™.

Dolby™ Pro Logic II™	Left	Right	Center	Rear L	Rear R
Left Total (Lt)	1	0	$\frac{1}{\sqrt{2}}$	$-j\frac{\sqrt{19}}{\sqrt{25}}$	$-j\frac{\sqrt{6}}{\sqrt{25}}$
Right Total (Rt)	0	1	$\frac{1}{\sqrt{2}}$	$j\frac{\sqrt{6}}{\sqrt{25}}$	$j\frac{\sqrt{19}}{\sqrt{25}}$

Un desarrollo complete de esta tecnología se encuentra disponible en [11].

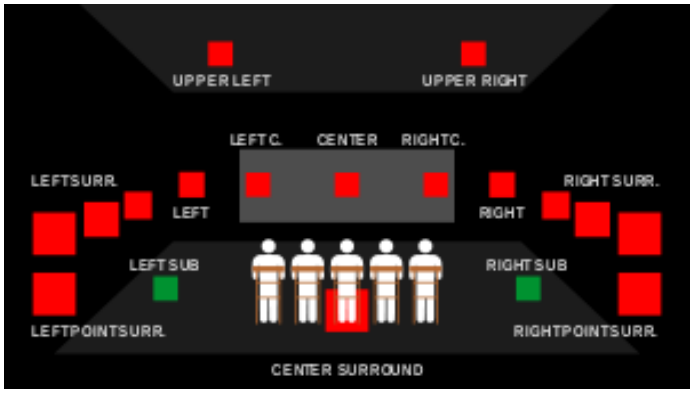
Se han implementado mejoras de este formato, mediante los estándares Pro Logic IIx™ y Pro Logic IIz™. No obstante, no se discuten aquí en mayor profundidad.

Dolby® Atmos™

En el año 2012 apareció un nuevo estándar de Dolby®. Es una nueva tecnología que permita hasta 128 canales de audio mas metadatos de paneo para distribuir los canales por la sala de cine de manera óptica, con renderizado dinamico para los altavoces. Dolby® Atmos™ permite al ingeniero de post-producción usar un plugin (disponible en la web de Dolby®) para operar con el estándar por excelencia en la industria audiovisual, Pro Tools © [49]. También puede usarse mediante una consola o mixer que incorpore Atmos™, para diseñar posiciones particulares en la sala de cine, mediante posicionado en 3D. Las fuentes que no se muevan dinámicamente, como los diálogos centrales o los sonidos ambiente en una película, se separan y pre-mezclan por separado en un formato multicanal tradicional.

Durante la reproducción del audio, cada sistema Dolby® Atmos™ de la sala de cine renderiza todos los sonidos dinámicos de los metadatos de paneo en tiempo real para que parezca que cada sonido viene de la posición elegida.

A finales de 2014, Dolby anuncio que Atmos™ pronto se integraría en los sistemas home cinema.
A fecha de febrero de 2015, sólo existían 2000 instalaciones con Dolby® Atmos™.

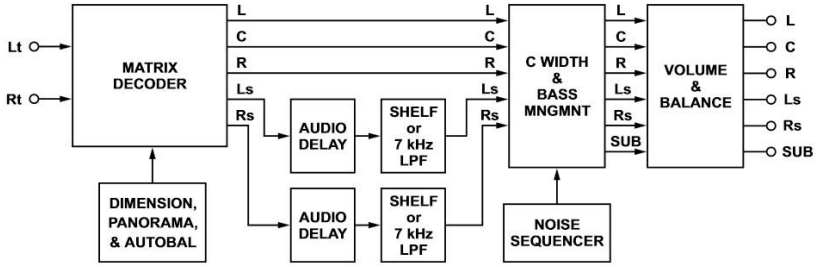


2.5.2. Una vista general sobre la historia de Dolby® Surround

Decodificador	Codificador	Año	Descripción (inglés)	Canales
	Dolby Stereo	1975	Cinema use with optical technology. Upmix stereo to Surround 4.0.	L R con C y MonoSurround codificados matricialmente
Dolby® Surround™	Dolby Surround	1982	First Home use. Analog. Upmix stereo to Surround 3.0	L R y MonoSurround codificado matricialmente
	Dolby Stereo SR	1986	Cinema use. Uses Dolby SR for noise reduction.	L R con C y MonoSurround codificados matricialmente
Dolby® Pro Logic™		1987	Improved Dolby Surround. Upmix Stereo to Surround 4.0.	L R con C y MonoSurround codificado matricialmente
Dolby® Digital™	AC3	1992 Film 1995 Laser Disc	Discrete channel encoder/decoder. Pro Logic Decoder can be used for downmixed stereo inputs.	L R C Lr Rr LFE
Dolby® Digital EX™/Dolby Digital Surround EX™		1999	non-discrete 6.1 or 7.1 (5.1 with Center Rear matrixed onto Lr & Rr)	L R C Lr Rr (con RearMono codificado matricialmente) LFE [7.1 no discreto: Left Back y Right Back]
Dolby® Pro Logic II™		2000	Improved Dolby Pro Logic. Upmix Stereo to Surround 5.1 in either Movie, Music, or Game mode.	L R C Lr Rr LFE
Dolby® Pro Logic IIx™		2002	Upmix Stereo or Surround 5.1 to 6.1 or 7.1 in either Movie, Music, or Game mode.	L R C Lr Rr LFE Left Back y Right Back
Dolby® Digital Plus™	Dolby Media Encoder	2005	Lossy compression codec; 48 kHz sampling frequency, 20-bit word length; supports data rates of 32 kbit/s – 6 Mbit/s, scalable, including 768 kbit/s – 1.5 Mbit/s on high-definition optical discs, typically, and 256 kbit/s for broadcast and online. 1.0- to 7.1-channel support for current media applications; extensible to 16 channels; discrete. Backward compatibile with Dolby Digital through S/PDIF connection up to 640 kbit/s. Supports Dolby Metadata.	L R C Lr Rr LFE Left Back and Right Back
Dolby® TrueHD™	Dolby Media Encoder	2005	Lossless compression codec; supports 44.1 kHz to 192 kHz sampling frequency up to 24-bit word length; supports variable data rate up to 18 Mbit/s; maximum channel support is 16 channels as presently deployed. Higher bitrate than Dolby Digital Plus. Blu-ray Disc channel support up to eight channels of 96 kHz/24-bit audio; six channels (5.1) up to 192 kHz/24-bit; and two- to six-channel support up to 192 kHz/24-bit maximum bit rate up to the maximum of 18 Mbit/s.	
Dolby® Pro Logic IIz™	Mohan	2009	Upmix Stereo or Surround 5.1/7.1 to 7.1 Height or 9.1 with the addition of front height channels. (Based on Dolby Pro Logic IIx.)	L, C, R, Ls, Rs, Lrs (Left Back), Rrs (Right Back), LFE, Lvh y Rvh

2.5.3. Mezclador estéreo a 5.1 con Dolby® Pro Logic II™

Un esquema del estándar Dolby® Pro Logic™ II se puede ver en la siguiente figura.



Hay que notar que los decodificadores Pro Logic™ encuentran su uso fundamental en la decodificación de audio que fue originalmente codificada en Pro Logic™ de dos canales (típicamente usado en CD, DVD, TV y otros) mientras que el propósito de este sistema mezclador es obtener audio Surround a partir de material de audio estéreo. Los decodificadores Pro Logic™ tienen también modos de decodificación para este propósito. Se distinguen dos modos: "Pro Logic" y "Music", paralelamente a las distinciones hechas por Dolby®.

Como se puede deducir del esquema, un sencillo *upmixer (*o *mezclador* hacia arriba*)* para codificar y decodificar sonido envolvente se puede implementar con los siguientes componentes:

· decodificador matricial de ganancias (puede incluir auto-panning)
· efectos de retardo
· efectos de filtrado
· isolacion del subwoofer (LFE)

El código Matlab para este upmixer ha sido desarrollado por [9] y se puede leer al final de esta sección. En lo que sigue, se describen los bloques fundamentales de un sistema de estas características.

Decodificador de ganancias matricial estático

El método más simple para decodificar los canales estéreo a cinco canales (y el punto de partida para métodos más complejos) es mezclar los canales L y R mediante una matriz de ganancias como la descrita en la subsección 2.5.2. Como se ha descrito, Pro Logic II™ tiene su propia matriz de decodificación estandarizada. La matriz que usa [9] difiere ligeramente de ésta última, pues se confirmó que un canal central con demasiado volumen hacía que todas las fuentes virtuales sonoras se posicionaran casi todas al centro, destruyendo la impresión de sonido envolvente. Por eso [9] elige un valor de 6 dB en vez de 3 dB

$$\mathbf{G_{c(n)}\,[dB] = 20log\left(\frac{1}{\sqrt{2}}\right) = -3dB}$$

para la atenuación del canal C, calculado como la simple suma del canal L y el R.

$$\mathbf{g_{c(n)} = 10^{-\frac{6\,dB}{20}} = 0.5012}$$

Notando que para generar los canales Surround Ls y Rs se necesita un cambio de fase de +90º para el canal L estéreo original y un cambio de fase de -90º para el canal R. Las ganancias se ponderan a la izquierda para el canal Ls y a la derecha para el canal Rs. La matriz resultante que se usa para el upmixer es:

$$\mathbf{[l_f(n)\ \ r_f(n)\ \ c(n)\ \ l_s(n)\ \ r_s(n)] = [l(n)\ \ r(n)]\begin{bmatrix} 1 & 0 & 0.5012 & j\,0.8165 & -j0.5774 \\ 0 & 1 & 0.5012 & j\,0.5774 & -j0.8165 \end{bmatrix}}$$

que difiere ligeramente como se aprecia de la de Pro Logic II™.

Auto-paneo

La técnica usada por Dolby ® Pro Logic II™ es el paneo automático de las fuentes según sus características (llamado auto-panning o auto-paneo). El auto-paneo se refiere a la detección del canal dominante en el momento del analisis (después del mezclado hacia arriba) y el consecuente paneo hacia ese canal. Como ejemplo, si el canal dominante es el R (que tiene que tener el mayor valor RMS) entonces el canal R se amplificará con una ganancia mientras que los otros cuatro canales se atenuarán. Este método de ajustar ganancias debe ser tal que la energía total RMS se conserve.

El auto-paneo solo es útil como una técnica dinámica; esto es, el canal que se amplifica en el momento del análisis debería ser realmente el canal dominante en ese instante. De acuerdo con [10,11], Pro Logic™ y Pro Logic II™ sólo aplican auto-paneo para el modo seleccionable "Pro Logic" pero no el modo "Music". Por ello los autores de [9] implementan un mezclador con esta misma distinción.
El modo "Music" no usa auto-paneo debido a que, en general, las voces de una canción se suelen posicionar en una posición virtual central dentro del audio estéreo; el paneo hacia otros lugares que no sean el centro daría como resultado efectos poco agradables para la escucha de archivos de música. Esto se puede confirmar pasando el upmixer por un archivo de música en modo "Pro Logic". Uno se da cuenta de que las voces se mueven constantemente de adelante hacia atrás, produciendo un efecto auditivo raro.

Una mejora posible no implementada en este mezclador 5.1 es el de hacer un auto-paneo por bandas de frecuencias, que es más útil y profesional que uno que solamente trabaje con el espectro entero.

Efectos de retardo

Un efecto importante que se usa en la reproducción multicanal es el de retardo temporal. Como se ha visto, si la misma señal llega a ambos oídos con una diferencia de tiempos suficiente, la ITD y el efecto de precedencia entran en acción como mecanismos de localización de sonidos. De hecho, algo común en las salas de cine y teatros es posicionar altavoces detrás de la audiencia con un retardo suficiente para no distorsionar la posición virtual de la fuente.
El mismo principio se aplica al sonido Surround. Los altavoces Surround tienen las señales de alimentación (Ls, Rs) retardadas para mantener la posición de la fuente frente al oyente (en la mayoría de los casos esta es la situación ideal). Dolby® recomiendo un retardo de 20 ms para el modo "Pro Logic". Para el modo "Music" no se retardan los canales Surround, pues el efecto que se quiere conseguir es el del sonido llegando a ambos oídos al mismo tiempo, dando las sensación de una llegada coincidente.

Además, los autores del mezclador [9] consideran bueno aplicar un retardo de 5 ms al canal central C para ambos modos Pro Logic" y "Music".

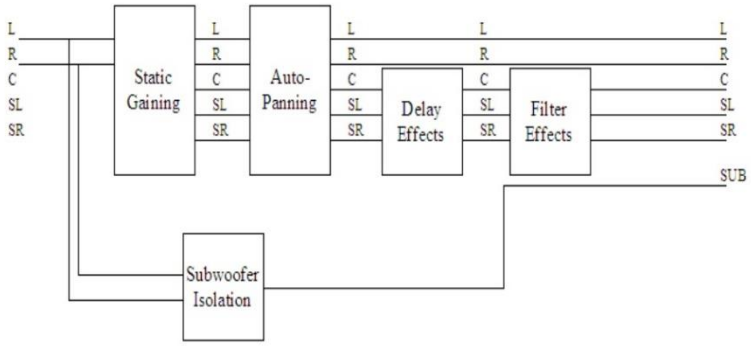
Efectos de filtrado

Además de los retardos, las señales de alimentación para los canales Surround también sobrellevan un filtrado particular. En el modo "Pro Logic" los canales Surround (Ls, Rs) se filtran paso-bajo hasta 7kHz. En el modo "Music", un *shelving filter* es aplicado en vez del paso-bajo para conseguir un sonido mas natural que emule la absorción para conseguir un sonido mas natural que emule la absorción y reflexión de altas frecuencias por las paredes de una sala.

Isolación del subwoofer (LFE)

Todas las consideraciones anteriores de cómo mezclar el audio estéreo a Surround no han tenido en cuenta el subwoofer o canal de graves (Low Frequency Effects LFE). De acuerdo con Dolby®, este canal tiene que aislarse de los demás canales de forma dinámica. Dolby® no proporciona más detalles al respecto y por ello esta implementación simplemente ha aislado el LFE del audio estéreo original directamente.

El canal LFE se obtiene filtrando ambos canales estéreo originales (L, R) con un filtro paso-bajo a 300 Hz y después mezclando los resultados con una ganancia de 3 dB para ecualizar el audio.



La siguiente figura muestra el esquema final del mezclador a 5.1 según la publicación descrita [9].

Codigo para Matlab® para implementar un upmixer 5.1

```
function [L, R, C, Lr, Rr, sub] = to_51Surround(x, Fs, mode)
% Mixing gains for xL and xR for the 5.1 channels
G = [1 0 10^(-10/20) 0.8165 0.5774 10^(3/20);...
0 1 10^(-10/20) 0.5774 0.8615 10^(3/20)];

% delays for the 5.1 channels
D = [0 0 0.005 0.02 0.02 0];

% Get xL and xR stereo channels
xL = x(:,1);
xR = x(:,2);

%% LP-Filtering simulating subwoofer
%[N_sub, Wn_sub] = ellipord(1000/(Fs/2), 1100/(Fs/2), 0.1, 60);
%[B_sub, A_sub] = ellip(N_sub, 0.1, 60, Wn_sub, 'low');
[B_sub, A_sub] = cheby1(5, 0.5, 300/22050, 'low');

subL = filter(B_sub, A_sub, xL);
subR = filter(B_sub, A_sub, xR);

%% STEREO-to-5.1 WITH THE DEFINED GAINS "G"
sub = G(1,6)*subL + G(2,6)*subR;
L = G(1,1)*xL + G(2,1)*xR;
R = G(1,2)*xL + G(2,2)*xR;
C = G(1,3)*xL + G(2,3)*xR;

% The Lr and Rr channels require phase shifts
shift_L = real((hilbert(xL) - xL)/i);
shift_R = real((hilbert(-1*xR) - (-1*xR))/i);
Lr_pre = G(1,4)*shift_L + G(2,4)*shift_R;
Rr_pre = G(1,5)*shift_L + G(2,5)*shift_R;

clear xL; clear xR; clear shift_L; clear shift_R; clear subL; clear subR;
```

```
%% PAN TO STRONGEST SIGNAL DINAMICALLY
% Iterate in blocks of length "L_secs" seconds
L_secs = 15e-3;

if(strcmp(mode,'prologic'))
        L = round(L_secs*Fs);
        nblocks = ceil(length(x)/L);

        for i = 1:nblocks
                start_IDX = (i-1)*L +1; % start sample
                end_IDX = start_IDX + min(L-1, length(x) - start_IDX);  % end sample

                % Find gains for the actual block
                [G_51upmix, max_IDX] = dominant_51Gains (L(start_IDX:end_IDX), R(start_IDX:end_IDX),
...
                C(start_IDX:end_IDX), Lr_pre(start_IDX:end_IDX), Rr_pre(start_IDX:end_IDX));

                focus(i, 1:2) = [max_IDX(1) length(max_IDX)];

                % Apply panning to this block
                L(start_IDX:end_IDX) = G_51upmix(1)*L(start_IDX:end_IDX);
                R(start_IDX:end_IDX) = G_51upmix(2)*R(start_IDX:end_IDX);
                C(start_IDX:end_IDX) = G_51upmix(3)*C(start_IDX:end_IDX);
                Lr_pre(start_IDX:end_IDX) = G_51upmix(4)*Lr_pre(start_IDX:end_IDX);
                Rr_pre(start_IDX:end_IDX) = G_51upmix(5)*Rr_pre(start_IDX:end_IDX);
        end
end

%% APPLY DELAYS
sub = delay(sub, Fs, D(6));
L = delay(L, Fs, D(1));
R = delay(R, Fs, D(2));
C = delay(C, Fs, D(3));
if(strcmp(mode, 'proLogic'))
        Lr_pre = delay(Lr_pre, Fs, D(4));
        Rr_pre = delay(Rr_pre, Fs, D(5));
end

if(strcmp(mode, 'proLogic'))
        %% LP-Filtering for ProLogic mode
        [N_LR, Wn_LR] = ellipord(7000/(Fs/2), 7500/(Fs/2), 0.1, 60);
        [B_LR, A_LR] = ellip(N_LR, 0.1, 60, Wn_LR, 'low');
        Lr = filter(B_LR, A_LR, Lr_pre);
        Rr = filter(B_LR, A_LR, Rr_pre);
elseif(strcmp(mode, 'music'))

        %% Shelving filter for Music mode
        [B_LR, A_LR] = shelving(Fs, 4000, -20, 'hc');
        Lr = filter(B_LR, A_LR, Lr_pre);
        Rr = filter(B_LR, A_LR, Rr_pre);
end
```

2.5.4. Reproducción de sonido 5.1 Surround con auriculares (Surround virtual)

Una vez se ha conseguido sonido 5.1 Surround, interesa combinar el audio binaural 3D sintetizado junto con el mezclador a 5.1 para crear Surround virtual en los auriculares.

Mezclando primero el audio stereo a 5.1 y mezclar seguidamente hacia abajo usando efectos de espacializacion mediante HRTFs para conseguir audio 3D para auriculares (audio binaural) es otra técnica que de provee muchas posibilidades al ingenierio de sonido.

Al usar la función quee se muestre "to_51Surround.m" sobre un archive de audio, se consiguen seis canales mono (para cada uno de los cinco altavoces y uno para el subwoofer LFE). Cada canal mono, excepto el LFE, se filtra con una HRTF para la posición del altavoz para simular su posición en el espacio virtual.

Ya que filtrar un canal mono con una HRTF (que es un par HRTF_L, HRTF_R) significa obtener dos canales, con este procedimiento se obtienen un total de 11 señales (10 mas la del subwoofer LFE que es mejor no filtrar). La señal binaural final para el oído izquierdo yL(n) se puede obtener simplemente mezcalando las cinco componentes L y de igual manera para la señal final al oído derecho yR(n). La mezcla no obstante es sin ganancias ya que la amplificación/atenuación la realiza el mezclador a 5.1.

```
function [y, L, R, C, Ls, Rs, LFE] = headphones51(x, fs, upMixMode)

% Define Constants
theta   = [30 330 0 110 250];
phi     = [0 0 0 0 0];
d       = 1.5;

%% 5.1 UPMIXING: Upmix stereo to 5.1 Surround
[up_L up_R up_C up_Ls up_Rs LFE] = to_51Surround(x, fs, upMixMode);
clear x;

%% SPATIALIZE EACH UPMIXED CHANNEL (except sub)

HRTF_up_L       = get_HRTF_d(phi(1), theta(1), d, 'L');
HRTF_up_R       = get_HRTF_d(phi(2), theta(2), d, 'R');
HRTF_up_C       = get_HRTF_d(phi(3), theta(3), d, 'R');
HRTF_up_Ls      = get_HRTF_d(phi(4), theta(4), d, 'L');
HRTF_up_Rs      = get_HRTF_d(phi(5), theta(5), d, 'R');

spatial_up_L = fconv(HRTF_up_L, up_L);
spatial_up_R = fconv(HRTF_up_R, up_R);
spatial_up_C = fconv(HRTF_up_C, up_C);
spatial_up_Ls = fconv(HRTF_up_Ls, up_Ls);
spatial_up_Rs = fconv(HRTF_up_Rs, up_Rs);

%% DOWNMIX TO STEREO

lengths = [length(spatial_up_L) length(spatial_up_R) length(spatial_up_C) length(spatial_up_Ls)
length(spatial_up_Rs) length(LFE)];
max_size = max(lengths);

y = zeros(max_size, 2);

% Mix zero-padded signals to outputs
y(1:lengths(1), :) = y(1:lengths(1), :) + spatial_up_L;
L = spatial_up_L;
y(1:lengths(2), :) = y(1:lengths(2), :) + spatial_up_R;
R = spatial_up_R;
y(1:lengths(3), :) = y(1:lengths(3), :) + spatial_up_C;
C = spatial_up_C;
y(1:lengths(4), :) = y(1:lengths(4), :) + spatial_up_Ls;
Ls = spatial_up_Ls;
y(1:lengths(5), :) = y(1:lengths(5), :) + spatial_up_Rs;
Rs = spatial_up_Rs;
y(1:lengths(6), :) = y(1:lengths(6), :) + [LFE LFE];
```

2.6. Audio 3D y funciones de transferencia relacionadas con la cabeza (HRTFs)

2.6.1. Introducción. Tipos de modelos para las HRTFs.

El motivo fundamental de usar HRTFs es el de presentar los efectos de elevación y giro de la fuente de una manera fácil y compacta. Siendo el modelado de HRTFs uno de los problemas actuales más importantes para la simulación de audio 3D eficiente y de calidad, la base de datos de respuestas al impulso tiene que ser muy precisa, requiriendo una frecuencia de muestreo para las medidas de 44.1 kHz como mínimo con longitudes de 512 taps normalmente. Esto impide el renderizado en tiempo real a menos que la complejidad de estas se reduzca inteligentemente. Y aún peor es la situación con las respuestas al impulso de sala (RIRs) que pueden tener longitudes de más de 20000 taps. Se distinguen claramente cuatro procedimientos diferentes hasta la fecha, que se describen por apartados a continuación. Estos son: 1) los modelos de HRTFs estándar, 2) los modelos de HRTFs obtenidos de una base de datos genérica, 3) los modelos de HRTFs individualizados y 4) los modelos paramétricos de HRTFs.

Modelos de HRTFs estándar
El uso de una HRTF estándar con las tecnologías actuales dará resultados de elevación pobres e incluso malos. Hasta la fecha, ni IEEE, ACM o AES han definido un estándar, pero parece que empresas multinacionales como Microsoft® o Intel® crearán un estándar de facto.

Modelos de HRTFs individualizados
Usar filtros individualizados produce los mejores resultados ya que se produce una adaptación a los filtros del propio sistema auditivo humano; sin embargo, requiere medir las HRTFs del oyente (suficientes para conseguir una continuidad en el audio que no sea audible o suficiente para poder interpolar y generar los filtros que falten; más adelante se discutirá por qué es necesaria esta continuidad [ver sección 3.7]), un inconveniente que consume mucho tiempo debido a las mediciones que se han de realizar.

Modelos de HRTFs de una base de datos genérica
Se elige el conjunto de HRTFs que mejor se adapte al oído del oyente de una base de datos con múltiples conjuntos obtenidos de las medidas realizadas a múltiples sujetos que representen las distintas complexiones auditivas de la población.
Para este Proyecto Fin de Carrera se usan dos de estas bases de datos, junto con un modelo paramétrico (ver siguiente apartado):

+ Base de datos de HRTFs del maniquí KEMAR, desarrollada por el MIT Media Lab, Massachusetts, Estados Unidos

La librería de HRIRs desarrollada por el MIT Media Lab se realizó en los años 1990 y está disponible para uso público en su página web [12]. Las respuestas al impulso (HRIRs) se midieron usando la cabeza artificial KEMAR Dummy Head.

+ Base de datos de HRTFs dependientes de la distancia, desarrollada por la universidad de Peking PKU, China

Más recientemente se ha desarrollado esta base de datos de filtros HRTFs para distintas distancias [13]. Para ello, se midieron las respuestas al impulso (HRIRs) posicionando las fuentes sonoras a diferentes distancias del maniquí. Esta base de datos en concreto diferencia filtros para distancias: 0,2, 0.3, 0.4, 0.5, 0.75, 1, 1.3 y 1.6 metros.

Modelos paramétricos de HRTFs

Los modelos paramétricos describen las funciones de transferencia mediante estudios y análisis de las características auditivas del ser humano. En este proyecto ha interesado utilizar un modelo paramétrico aproximado que se pueda adaptar para cada oyente, de la manera más general posible. En las publicaciones encontramos tres ramas de investigación para modelado de HRTFs:

+ Modelos estructurales de HRTFs

Tienen en cuenta las funciones de transferencia de los mecanismos físicos de transmisión del sonido, como el apantallamiento debido a la cabeza, reflexiones de hombros, torso, pinna, etc… El modelo estructural general de HRTF se puede estructurar en tres partes:

· Apantallamiento de la cabeza e ITD
· Eco de los hombros
· Reflexiones en el pabellón auricular (pinna)

Este ha sido el modelo paramétrico de HRTFs elegido para este Proyecto Fin de Carrera.

+ Modelos paramétricos de HRTFs con funciones racionales polos/ceros

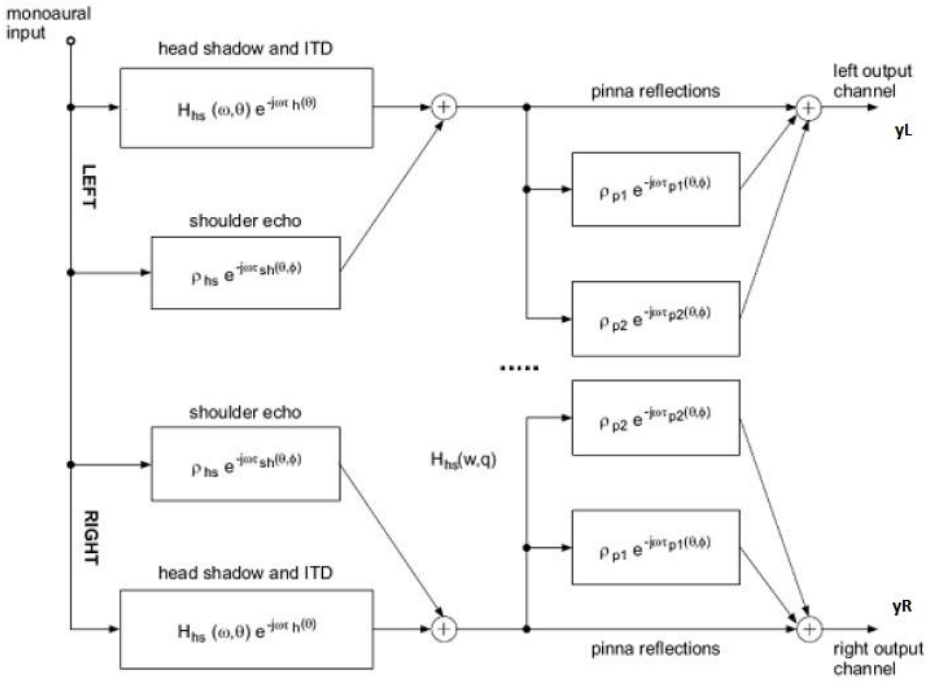
Modelan las funciones de transferencia mediante polinomios racionales y adaptando los coeficientes de dichos polinomios para crear los polos y ceros necesarios en el modelo de la HRTF. Están en estrecha relación con los modelos anteriores.

+ Modelos paramétricos de HRTFs usando expansiones en series

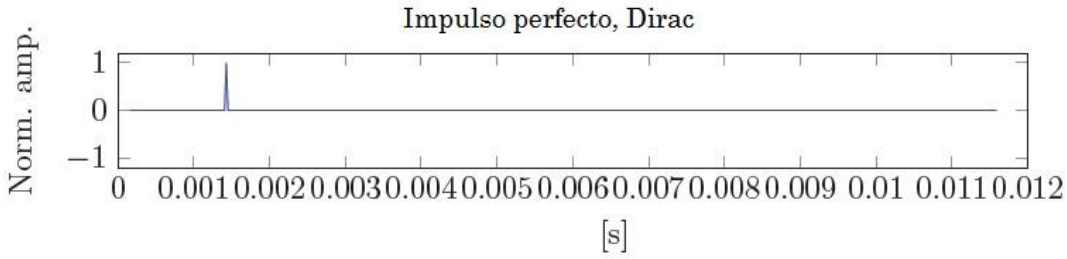
· Representación funcional de HRTFs mediante series de Fourier-Bessel para explotar la periodicidad de las HRTFs [15]
· Principal Component Analysis PCA (o expansión de Karhunen-Loéve) para reducción de complejidad y conseguir un conjunto independiente de la direccion de funciones base y un conjunto dependiente de la direccion de pesos para combinar las funciones base [16]
Estos dos métodos se discuten en este capítulo más adelante.

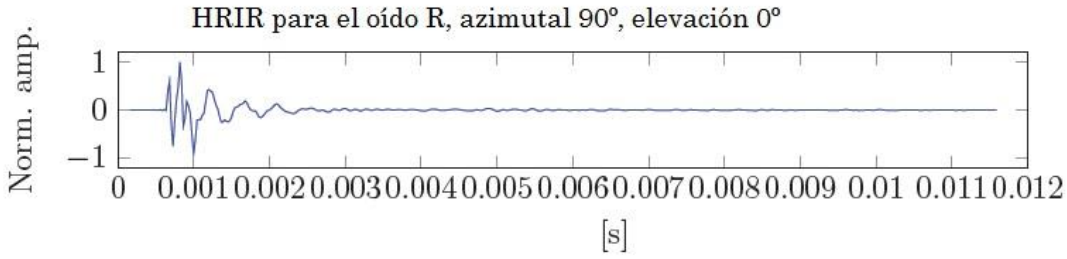
Merece mención especial un nuevo modelo de HRTFs desarrollado en los últimos años y al cual el lector puede referenciarse mediante la bibliografía: los modelos de HRTFs basados en la DWT y filtros sparse [17, 18, 19].

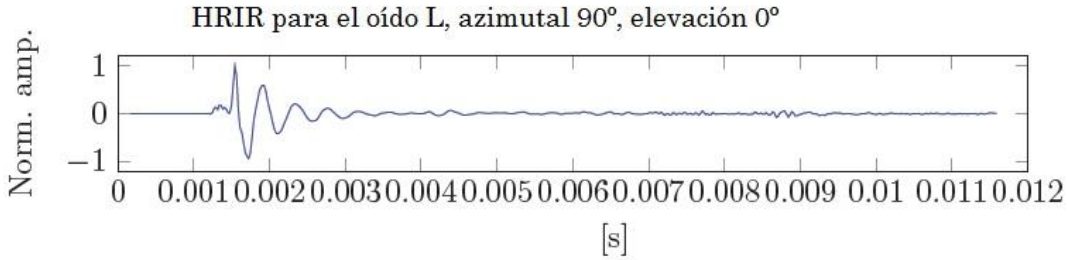
A continuación se muestra un modelo estructural tipo HRTF [9]. Los modulos separados tienen en cuenta la ITD, el apantallamiento debido a la cabeza, reflexiones en torso y hombros (shoulder echo) y efectos del pabellón auricular (pinna reflections).

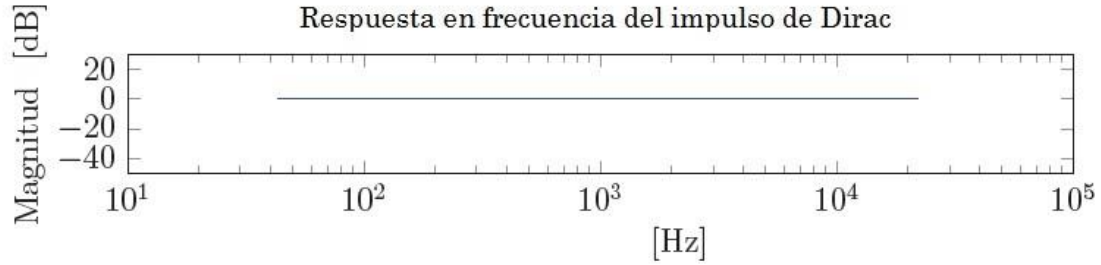


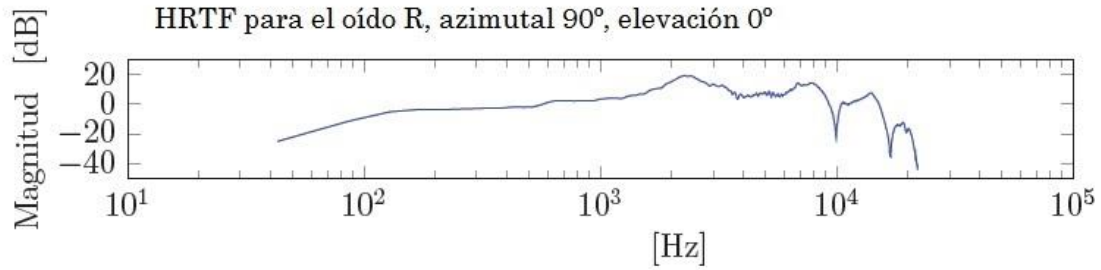
Para ilustrar cómo es un filtro relacionado con la cabeza (HRTF) o su correspondiente temporal (HRIR), se muestran a continuación primero HRIRs y luego HRTFs para el oído izquierdo (L) y derecho (R) para un ángulo horizontal o azimutal $\theta = 90º$ y un ángulo vertical o de elevación $\varphi = 0º$. Se aprecia un máximo a las frecuencias 2-3 kHz. Esto coincide con que la frecuencia de inteligibilidad del habla en los humanos está por entre ese ancho de banda. Las causas y efectos de los otros máximos/mínimos relativos o absolutos son más difíciles de identificar.

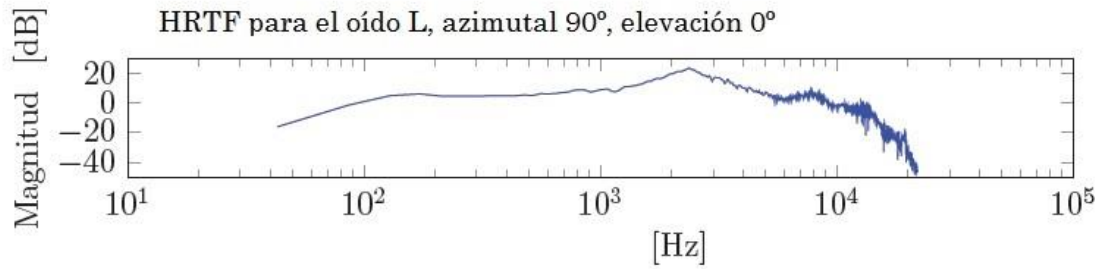

Impulso perfecto, Dirac
Norm. amp.
1
0
−1
0 0.001 0.002 0.003 0.004 0.005 0.006 0.007 0.008 0.009 0.01 0.011 0.012
[s]


HRIR para el oído R, azimutal 90°, elevación 0°
Norm. amp.
1
0
−1
0 0.001 0.002 0.003 0.004 0.005 0.006 0.007 0.008 0.009 0.01 0.011 0.012
[s]

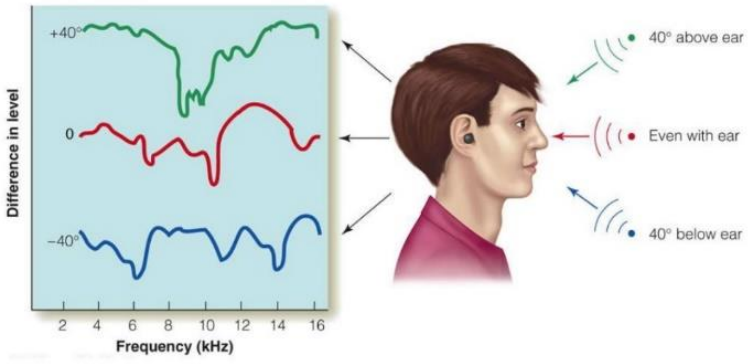

HRIR para el oído L, azimutal 90°, elevación 0°
Norm. amp.
1
0
−1
0 0.001 0.002 0.003 0.004 0.005 0.006 0.007 0.008 0.009 0.01 0.011 0.012
[s]


Respuesta en frecuencia del impulso de Dirac
Magnitud [dB]
20
0
−20
−40
10¹ 10² 10³ 10⁴ 10⁵
[Hz]


HRTF para el oído R, azimutal 90°, elevación 0°
Magnitud [dB]
20
0
−20
−40
10¹ 10² 10³ 10⁴ 10⁵
[Hz]


HRTF para el oído L, azimutal 90°, elevación 0°
Magnitud [dB]
20
0
−20
−40
10¹ 10² 10³ 10⁴ 10⁵
[Hz]

Los resultados se interpretan mucho más fácilmente usando la figura siguiente. Se dibujan tres HRTF (recordar que la transformada de Fourier de la HRIR es la HRTF) para tres ángulos de procedencia del sonido diferentes. El sonido proveniente de ángulos diferentes excitará el sistema auditivo humano, de distinta forma, creando grandes diferencias entre las funciones de transferencia.



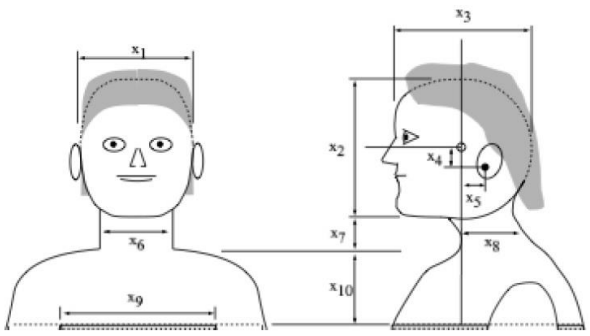
Los efectos de apantallamiento debido a la forma de la cabeza, junto a las resonancias en el canal auditivo y las reflexiones y difracciones en el torso y los hombros conformarán la forma específica de la HRTF.

2.6.2. Modelo para la característica ITD

Un modelo sencillo para la ITD se puede obtener mediante un simple retardo en segundos ya que es una funcion del angulo azimutal θ de la siguiente manera [20]. La formula para calcular el tamaño optimo que modele una esfera para la cabeza es:

$$\mathbf{r}_{\mathbf{head}}^{\mathbf{opt}} = \left(\mathbf{0.51} \cdot \frac{\mathbf{x_1}}{\mathbf{2}} + \mathbf{0.18} \cdot \frac{\mathbf{x_3}}{\mathbf{2}} + \mathbf{32}\right) \ [\mathbf{mm}]$$

con las medidas x1 y x3 la anchura y la profundidad de la cabeza del oyente, respectivamente, expresadas en milímetros, como muestra la siguiente imagen [43].



La ITD se genera debido a que la velocidad del sonido es de valor finito y relativamente bajo, además del hecho de que nuestros oídos se encuentran posicionados a una cierta distancia el uno del otro. La ITD se puede definir como la diferencia en distancia de una fuente sonora al oído mas cercano y al otro oído, dividido por la velocidad del sonido.
Esta ITD variará cuando cambien el angulo azimutal y/o el angulo de elevación. Debido al cono de confusión, no hay una correspondencia uno a uno entre un par (θ, φ) y una ITD. Multiples angulos diferentes pueden resultar en la misma ITD. La característica ITD es única para cada persona. Se relaciona con el tamaño y forma de la cabeza y la posición de las orejas. Es por esto que el uso de un modelo de HRTF fijo dará errores de localización para la mayoría de un conjunto de individuos.
El modelo de ITD más sencillo es para un modelo de cabeza esférico con las orejas en $\theta = \pm 90º$.

Asumiendo que la fuente esta a una distancia infinita, la ITD mas sencilla se puede aproximar como:

$$\mathbf{ITD} = \frac{\mathbf{r_{head}}(\boldsymbol{\theta} + \mathbf{sin}\boldsymbol{\theta})}{\mathbf{c}}$$

En esta fórmula no obstante el valor del radio de la cabeza irá expresado, como es lógico, en metros.
Otra formula para la ITD, más compleja, que depende de los dos ángulos azimutal y de elevación se da en [3D Audio using loudspeakers, Gardner] para un modelo de la cabeza esférico:

$$\mathbf{ITD} = \frac{\mathbf{r_{head}}}{\mathbf{c}}\left(\mathbf{arcsin}\left(\mathbf{cos}(\boldsymbol{\varphi})\,\mathbf{sin}(\boldsymbol{\theta})\right) + \mathbf{cos}(\boldsymbol{\varphi})\,\mathbf{sin}(\boldsymbol{\theta})\right)$$

El valor de la ITD normalmente varía entre 0-1 ms [ver sección 1.3]. La ITD es casi independiente de la frecuencia por debajo de 500 Hz y por encima de 3 kHz. Entre estas frecuencias, la ITD depende de la frecuencia y presenta mínimos entre 1.4 kHz y 1.6 kHz.
Para frecuencias normalizadas

$$\boldsymbol{\mu} = \frac{\boldsymbol{\omega} \cdot \mathbf{r_{head}}}{\mathbf{c}} > \mathbf{1}$$

con c la velocidad del sonido en m/s y $\mathbf{r_{head}}$ el radio efectivo de la cabeza en metros, la diferencia entre el tiempo en que la onda llega al punto de observación y el tiempo que tardaría en llegar al centro de una esfera de radio $\mathbf{r_{head}}$ en campo libre (ITD) se calcula, para un modelo de la cabeza esférico:

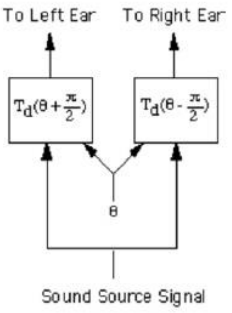
$$\mathbf{retardo\ de\ grupo} = \Delta\mathbf{T_d}(\boldsymbol{\theta}) = \begin{cases} -\frac{\mathbf{r_{head}}}{\mathbf{c}}\mathbf{cos}(\boldsymbol{\theta}), & \mathbf{si}\ \ \mathbf{0} \le |\boldsymbol{\theta}| \le \frac{\boldsymbol{\pi}}{\mathbf{2}} \\ \frac{\mathbf{r_{head}}}{\mathbf{c}}\left(|\boldsymbol{\theta}| - \frac{\boldsymbol{\pi}}{\mathbf{2}}\right), & \mathbf{si}\ \ \frac{\boldsymbol{\pi}}{\mathbf{2}} \le |\boldsymbol{\theta}| \le \boldsymbol{\pi} \end{cases}$$

Para un humano adulto general con un radio aproximado de 8.75 cm , $\mu = 1$ se corresponde con una frecuencia de aprox. 624 Hz.
No obstante, para $\mu < 1$ el retardo relativo aumenta mas alla de este valor, volviéndose aproximadamente un 50% mayor que el valor predecido por la ecuación para $\mu \longrightarrow 0$.

Para mantener este sistema como uno causal, se puede añadir un factor $\frac{\mathbf{r_{head}}}{\mathbf{c}}$ para conseguir un sistema causal y por tanto físicamente realizable, resultando en [43]:

$$\mathbf{T_d}(\boldsymbol{\theta}) = \begin{cases} \frac{\mathbf{r_{head}}}{\mathbf{c}} - \frac{\mathbf{r_{head}}}{\mathbf{c}}\mathbf{cos}(\boldsymbol{\theta}), & \mathbf{si}\ \ \mathbf{0} \le |\boldsymbol{\theta}| \le \frac{\boldsymbol{\pi}}{\mathbf{2}} \\ \frac{\mathbf{r_{head}}}{\mathbf{c}} + \frac{\mathbf{r_{head}}}{\mathbf{c}}\left(|\boldsymbol{\theta}| - \frac{\boldsymbol{\pi}}{\mathbf{2}}\right), & \mathbf{si}\ \ \frac{\boldsymbol{\pi}}{\mathbf{2}} \le |\boldsymbol{\theta}| \le \boldsymbol{\pi} \end{cases}$$

Como aproximación de primer orden a una HRTF, se puede modelar una funcion así mediante filtros lineales que den los retardos temporales relativos dados por esta ecuación. Esto generará buenas ITDs, pero no ILDs. Además, la ITD resultante será independiente de la frecuencia, que es contrario a las observaciones realizadas anteriormente. Un modelo simple como este obviamente no producirá ningún efecto de externalización ni posibilidad de discriminación frontal/atrás (anterior/posterior). Pero sí produce una imagen sonora que se mueve suavemente desde el oído izquierdo pasando "por dentro de la cabeza" hasta el oído derecho cuando el ángulo azimutal (θ) varía entre -90º y +90º. El esquema de implementación se muestra en la figura de la derecha.



2.6.3. Modelo para la característica ILD y modelo combinado

Ya a finales del siglo XIX, este mismo científico Lord Rayleigh había obtenido la solución en el dominio frecuencial para la difracción de una onda acústica debida a una esfera rigida. La solución analítica a la ecuación de onda para una onda plana incidente en una esfera rigida fue hallada por Rayleigh en 1904 y esta dada por [44]:

$$\mathbf{p_{esfera}} = \mathbf{p_0} \sum_{\mathbf{m=0}}^{\mathbf{N}} \mathbf{(2m+1)\, j^m\, J_m(kr)\, P_m(cos(\theta))} + \sum_{\mathbf{m=0}}^{\mathbf{N}} \mathbf{A_m\, h_m(kr)\, P_m(cos(\theta))}$$

con $J_m(\cdot)$ la funcion de Bessel esférica de orden m, $P_m(\cdot)$ la funcion polinomio de Legendre de orden m, $j = \sqrt{-1}$ es el numero imaginario, $h_m(\cdot)$ la funcion de Hankel esférica de primera especie de orden m y los coeficientes {Am} vienen dados por:

$$\mathbf{A_m} = -\mathbf{p_0(2m+1)\, j^m} \frac{\mathbf{m \cdot J_{m-1}(k \cdot r_{head}) - (m+1) \cdot J_{m+1}(k \cdot r_{head})}}{\mathbf{m \cdot h_{m-1}(k \cdot r_{head}) - (m+1) \cdot h_{m+1}(k \cdot r_{head})}}$$

Las soluciones de esta ecuación son las respuestas que se muestran para comparar la esfera rígida ideal con el modelo ILD presentado a continuación.

Para conseguir ITDs e ILDs realistas, también para bajas frecuencias y para incluir los efectos del apantallamiento debido a la cabeza, se puede realizar un filtro con respuesta en frecuencia aproximada con 1 polo y 1 cero [20, 45] que se ponga en cascada con el elemento de retardo para la ILD explicado anteriormente; dicho filtro de apantallamiento está caracterizado por:

$$\mathbf{s_{cero}} = \frac{-\mathbf{2\omega_0}}{\boldsymbol{\alpha(\theta)}}, \qquad \mathbf{s_{polo}} = -\mathbf{2 \cdot \omega_0}, \qquad \boldsymbol{\omega_0 = c / r_{head}}$$

$$\boldsymbol{\alpha(\theta)} = \left(\mathbf{1} + \frac{\boldsymbol{\alpha_{min}}}{\mathbf{2}}\right) + \left(\mathbf{1} - \frac{\boldsymbol{\alpha_{min}}}{\mathbf{2}}\right) \mathbf{cos}\left(\frac{\boldsymbol{\theta}}{\boldsymbol{\theta_{min}}} \mathbf{180^o}\right) \cong$$

$$\cong \mathbf{1.05 + 0.95 \cdot cos}\left(\frac{\boldsymbol{\theta}}{\mathbf{150^o}} \mathbf{180^o}\right), \qquad \mathbf{si\ \alpha_{min} = 0.1\ y\ \theta_{min} = 150^o}$$

$$\mathbf{0 \le \alpha(\theta) \le 2}$$

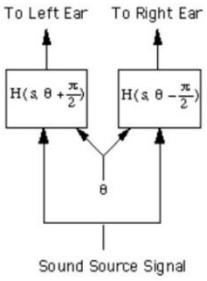
con la variable s la frecuencia de Laplace y ω0 = c / r_head. Como se puede apreciar, los polos se mantienen fijos y los ceros varían con la dirección de la fuente según la funcion (θ). Si (θ) = 2, hay una amplificación de 6 dB a altas frecuencias, mientras que si (θ) < 1 hay una atenuación. La elección de la función (θ) y de los valores

$$\boldsymbol{\alpha_{min}} = \mathbf{0.1}, \qquad \boldsymbol{\theta_{min}} = \mathbf{150^o}$$

que mejor se aproximan a la solución de la ecuación de onda para una superficie esferica de Rayleigh se calculan de forma experimental [45].
Los polos hacen la función de caracterizar las frecuencias de resonancia del oído. El par polo-cero se puede transladar directamente a un filtro digital IIR estable mediante transformación bilineal. El filtro resultante es:

$$\mathbf{H_{head\ shadow}(\theta, z^{-1})} = \frac{\mathbf{(\omega_0 + \alpha(\theta) \cdot fs) + (\omega_0 - \alpha(\theta) \cdot fs) \cdot z^{-1}}}{\mathbf{(\omega_0 + fs) + (\omega_0 - fs) \cdot z^{-1}}}$$

$$\mathbf{H_{head\ shadow}(\theta, \omega)} = \frac{\mathbf{1 + j \frac{\alpha(\theta)\, \omega}{2\omega_0}}}{\mathbf{1 + j \frac{\omega}{2\omega_0}}}$$



En la práctica, esta función realzará las altas frecuencias en un ángulo azimutal $\theta = 0º$ y atenuará las frecuencias para las que $\theta = 180º$, simulando así el efecto de apantallamiento. Hay que notar que a bajas frecuencias el filtro $\mathbf{H_{head\ shadow}(\theta, \omega)}$ también introduce un retardo de grupo:

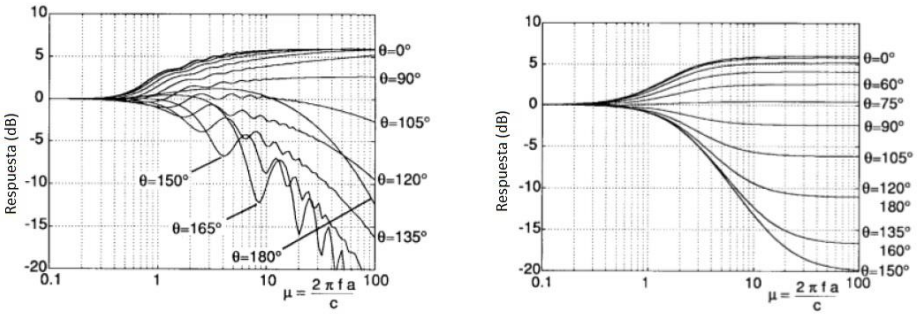
$$\mathbf{T_g = \frac{1 - \alpha(\theta)}{2\omega_0} = \frac{1}{2}\frac{r_{head}}{c}(1 - \alpha(\theta))}$$

que se añade al retardo de alta frecuencia $\Delta T_d(\theta)$.

A continuación, se muestra una función que usa lo anterior para generar una HRTF simple. Este modelo provee de una implementación aproximada pero simple en términos de tratamiento digital de señales para la solución de Rayleigh para una esfera.

```
%% Modelo simple de Funcion de Transferencia relacionada con la cabeza (HRTF)
% Fuentes:
% "DAFx: Digital Audio Effects", 2nd ed., Zoelzer, U.
function [HRIR] = simpleHRIR(azi, Fs)

theta           = theta + 90;
theta_0 = 150;
%alfa_min       = 0.05;
alfa_min        = 0.1;
c               = 343;   % velocidad del sonido
r_head  = 0.08;  % radio de la cabeza
w0              = c/r_head;
% ITD   es approx. ITD = 3 ms.
input           = zeros(round(0.003*Fs),1);
input(1)        = 1;

alfa    = 1 + alfa_min/2 + (1 - alfa_min/2) * cos(theta/theta_0*pi);
alfa2   = 1.05 + 0.95 * cosd(1.2*theta);

if any(abs(alfa - alfa2) > 0.00001)
    disp('Error in simpleHRIR.m. The filter is not well dimensioned.');
end

B       = [(alfa + w0/Fs)/(1 + w0/Fs), (-alfa + w0/Fs)/(1 + w0/Fs)];

A       = [1, -(1 - w0/Fs)/(1 + w0/Fs)];

if (abs(theta) < 90)
    gdelay = round(-Fs/w0 * (cos(theta*pi/180) - 1));
else
    gdelay = round(Fs/w0 * ((abs(theta) - 90) * pi/180 + 1) );
end

out_magn = filter(B, A, input);

HRIR    = [zeros(gdelay,1); out_magn(1:end-gdelay)];
end

%Dibujar la HRIR. Si se quiere la HRTF hay que transformarla al dominio de la frecuencia
%plot(HRIR)
```

En la figura siguiente se muestra la respuesta en frecuencia de una esfera rígida ideal y de la obtenida mediante el procedimiento anterior, concluyendo que los resultados son razonablemente buenos.



a) Esfera rígida ideal

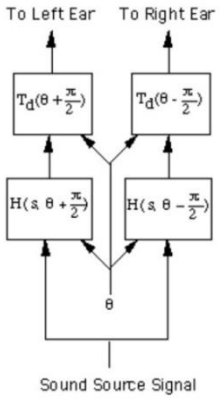
b) Respuesta dada por el modelo ILD

El filtro anterior es un ejemplo particularmente simple de una aproximación como funcion racional a una HRTF. Usado por si mismo, puede producir efectos bastante convincentes en el plano azimutal aunque sea un modelo que solo se adapte a las magnitudes de las características mas prominentes del espectro de la HRTF. Su efectividad se puede mejorar notablemente añadiendo una sección paso-alto que contemple el retardo de propagación y por tanto la ITD. Esto resulta en un modelo caracterizado por [45]:

$$\mathbf{H_{HRTF}(\omega, \theta)} = \frac{\mathbf{1 + j\frac{\alpha(\theta - \theta_{ear})\,\omega}{2\omega_0}}}{\mathbf{1 + j\frac{\omega}{2\omega_0}}}\mathbf{e^{-j\omega T_d(\theta - \theta_{ear})}}$$

con θ_{ear} el ángulo donde esta posicionada la oreja respecto al origen de coordenadas (normalmente entre 80º-100º).

Poniendo en cascada los dos filtros anteriores obtenemos un aproximado pero muy útil modelo de una cabeza esférica. Aunque sigue sin haber ningún tipo de externalización o elevación, este modelo elimina la imagen doble del modelo ITD y produce una imagen virtual bastante compacta.
Una desventaja es que este filtro solo produce efectos en el plano azimutal. Para conseguir también las características del plano de elevación, hay que usar una función de tres variables $\mathbf{H_{HRTF}(\omega, \theta, \varphi)}$. Se ha demostrado que cambiando solo el retardo de un modelo monoaural tan sencillo como



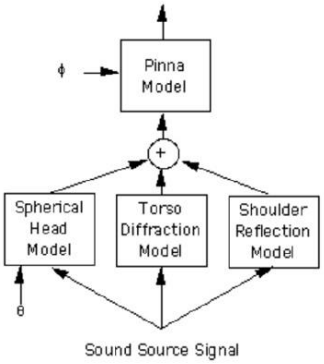
$$\mathbf{H_{HRTF}(\omega, \varphi) = 1 + \alpha_A \cdot e^{-j\omega T_A(\varphi)} + \alpha_V \cdot e^{j\omega T_V(\varphi)}}$$

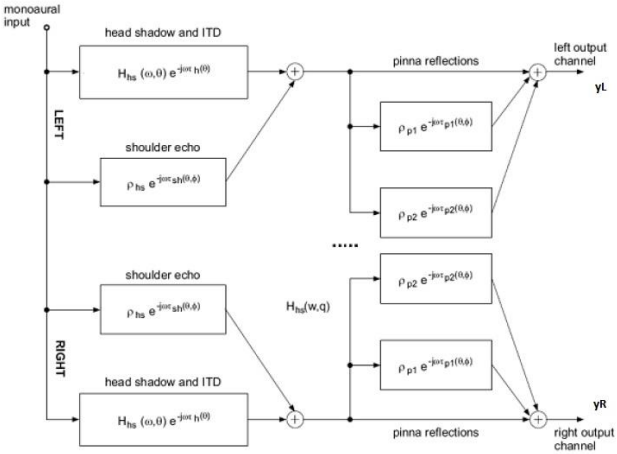
se puede producir un espectro con una característica de elevación. Los tests psicoacústicos llevados a cabo mostraron que esto producia un efecto de movimiento vertical en el plano frontal para angulos de elevación φ de por lo menos entre -45º y +45º [46]. Esto, junto con otros estudios realizados para la localización en el plano de elevación o mediano, muestran que es posible que las características azimutales sean binaurales mientras que las características de elevación sean monoaurales [47].

Debido a que es inviable factorizar la HRTF mediante un método de separación de variables, los científicos han aplicado varias técnicas de diseño de filtros, identificación de sistemas y técnicas de redes neuronales para intentar adaptar modelos de parámetros multiples a datos medidos. No obstante no se consigue un balance definitivo entre coste computacional y precisión para una implementación en tiempo real.

Combinando los modelos del pabellón del oído y la cabeza, así como añadiendo modelos de difracción del torso y modelos de reflexión de los hombros, resonancias del canal auditivo, modelos acústicos de la sala… podemos obtener aproximaciones sucesivamente mejores a la HRTF real, aunque más complejas. Este modelo está físicamente bien formado, es computacionalmente eficiente y provee de una adaptación flexible a distintos entornos para generar audio 3D.

La siguiente figura muestra, a la derecha, otra vez el modelo combinado más detallado de una HRTF a partir de tres componentes: un modelo de apantallamiento debido a la cabeza e ITD, reflexiones de torso y hombros y reflexiones del pabellón auricular (pinna). Estas se ven esquematizadas en la parte izquierda.





Un examen de las HRIRs muestra que la mayor actividad del pabellón auricular ocurre en los primeros 0.7 ms, que corresponden a 32 muestras a una frecuencia de muestreo de 44.1kHz. Por tanto, un filtro FIR de 32 taps puede elegirse para simular el modelo para las características de elevación. Como se puede apreciar, hay dos cantidades asociadas con cada evento, un coeficiente de reflexión ρ_{pn} y un retardo temporal τ_{pn}. Las investigaciones de [46] muestran que los valores de los coeficientes de reflexión ρ_{pn} no son un parámetro critico y se pueden asignar a valores constantes, independientemente del angulo azimutal, del angulo de elevación y del oyente. Los retardos temporales no obstante sí que varían dependiendo de estas ultimas tres variables. Ya que las funciones del azimutal y de elevación son siempre periódicas, es lógico aproximar los retardos temporales usando sinusoides. En estas mismas investigaciones se encontró empíricamente que la siguiente fórmula es una aproximación razonable para el retardo temporal del evento en el pabellón n-ésimo:

$$\boldsymbol{\tau_{pn} = A_n \cdot \cos\left(\frac{\theta}{2}\right) \cdot \sin(D_n \cdot (90º - \varphi)) + B_n}\,, \qquad \boldsymbol{-90º \le (\theta\,, \varphi) \le 90º}$$

siendo {An} coeficientes de amplitud, {Bn} coeficientes de offset y {Dn} factores de escala. La siguiente tabla muestra los coeficientes a usar en función del evento n-ésimo para el modelo de pabellón auricular (pinna) dado por la ecuación anterior [46].

n	ρ_{pn}	An	Bn	Dn (para los sujetos 1 "PB" y 3 "NH")	Dn (para el sujeto 2 "RD")
2	0.5	1	2	1	0.85
3	-1	5	4	0.5	0.35
4	0.5	5	7	0.5	0.35
5	-0.25	5	11	0.5	0.35
6	0.25	5	13	0.5	0.35

2.6.4. Apantallamiento acústico debido a una esfera

En la actualidad, hay otros modelos de HRTFs de esfera rígida disponibles más realistas.
La solución teórica al problema de Rayleigh reformulada a finales de los años 1990 por [14] para el modelo de HRTF basado en una esfera rígida que aproxime la cabeza humana queda parametrizado mediante las ecuaciones:

$$\mathbf{HRTF}(\boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\theta}) = -\left(\frac{\boldsymbol{\rho}}{\boldsymbol{\mu}}\right) \cdot \mathbf{e}^{-\mathbf{j}\cdot\boldsymbol{\rho}\cdot\boldsymbol{\mu}} \cdot \boldsymbol{\Psi}(\boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\theta})$$

$$\boldsymbol{\Psi}(\boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\theta}) = \sum_{\mathbf{m}=\mathbf{0}}^{\infty} (\mathbf{2m}+\mathbf{1}) \cdot \mathbf{P_m}(\mathbf{cos}(\boldsymbol{\theta})) \cdot \frac{\mathbf{h_m}(\boldsymbol{\rho}\cdot\boldsymbol{\mu})}{\mathbf{h'_m}(\boldsymbol{\mu})}, \qquad \boldsymbol{\rho} = \frac{\mathbf{r}}{\boldsymbol{r_{head}}} > \mathbf{1}$$

En esta representación, la variable r es la distancia desde el centro de la esfera a la fuente sonora, θ es el ángulo azimutal u horizontal, ρ = r/r_head es la distancia normalizada con respecto al radio de la cabeza, k = ω/c es el número de onda, μ = k· r_head y hm(x) es la función de Hankel esférica m-ésima:

$$\mathbf{h_m}(\mathbf{x}) = \mathbf{J_m}(\mathbf{x}) + \mathbf{j} \cdot \mathbf{Y_m}(\mathbf{x})$$

Jm(x) e Ym(x) son las funciones de Bessel de primera y segunda especie ordinarias. h'm(x) es la derivada de la función con respecto de su argumento (x).
La función **Pm(x)** es el polinomio de Legendre m-ésimo:

$$\mathbf{P_m}(\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{2^m m!}} \frac{\mathbf{d^m}}{\mathbf{dx^m}}((\mathbf{x^2}-\mathbf{1})^{\mathbf{m}}) = \frac{\mathbf{1}}{\mathbf{2^m}} \sum_{\mathbf{g}=\mathbf{0}}^{\mathbf{m}} \binom{\boldsymbol{m}}{\boldsymbol{g}}^{\mathbf{2}} (\mathbf{x}+\mathbf{1})^{\mathbf{m-g}}(\mathbf{x}-\mathbf{1})^{\mathbf{g}}$$

Este modelo, un poco aparatoso matemáticamente, permite modelar una cabeza esférica rígida con los pertinentes efectos de difracción del sonido incidente [14]. La función resultante, la HRTF, es una función de transferencia y relaciona la presión sonora que habría en el centro de la esfera en el espacio libre con la presión que se genera realmente en la superficie de la esfera. La transformada inversa de Fourier de la HRTF es la respuesta al impulso relacionada con la cabeza (HRIR).
En el caso de distancias muy grandes (r→∞), la HRTF queda como la solución de Rayleigh para una fuente infinitamente distante:

$$\mathbf{HRTF}(\infty, \boldsymbol{\mu}, \boldsymbol{\theta}) = \frac{\mathbf{1}}{\boldsymbol{\mu}^{\mathbf{2}}} \sum_{\mathbf{m}=\mathbf{0}}^{\infty} \frac{(-\mathbf{j})^{\mathbf{m-1}} \cdot (\mathbf{2m}+\mathbf{1}) \cdot \mathbf{P_m}(\mathbf{cos}(\boldsymbol{\theta}))}{\mathbf{h'_m}(\boldsymbol{\mu})}$$

Para el caso de las bajas frecuencias (para μ pequeño), este resultado se suele aproximar como:

$$\mathbf{HRTF}(\infty, \boldsymbol{\mu}, \boldsymbol{\theta}) \cong \mathbf{1} - \mathbf{j} \cdot \frac{\mathbf{3}}{\mathbf{2}} \cdot \boldsymbol{\mu} \cdot \mathbf{cos}(\boldsymbol{\theta})$$

Ya que la ecuación de la HRTF converge más lentamente conforme μ aumenta, el comportamiento para altas frecuencias es más complicado.
Para el caso especial de incidencia normal (θ = 0º) se puede argumentar que, físicamente, cuando la longitud de onda es pequeña comparada con el radio de la esfera r_head la solución tiene que reducirse a una onda plana sonora que incide normalmente sobre una superficie plana rígida, donde la presión en la superficie será dos veces la presión en el espacio libre. Por ello,

$$|\mathbf{HRTF}(\infty, \infty, \mathbf{0})| = \mathbf{2}$$

Estos resultados para casos particulares especiales sirven para definir límites fundamentales en la solución general de la HRTF. En general, se necesitan métodos numéricos complicados para evaluar la HRTF. Un algoritmo útil para implementar el modelo esférico de HRTFs en Matlab® se da a continuación. Más información sobre la función y cómo usarla en [14].

```
%Funcion para similar una HRTF en un modelo esférico%

function HRTF = esfera(r_head, r, theta, f, c, lim)

if(lim > 1e-4)
    lim = 1e-4;
end

x = cos(theta);
mu = (2*pi*f*r_head) / c;
rho = r / r_head;
zr = 1 / (j*mu*rho);
za = 1 / (j*mu);
Qr2 = zr; Qr1 = zr*(1 - zr);
Qa2 = za; Qa1 = za*(1 - za);
P2 = 1;
P1 = x;
sum = 0;
term = zr / (za*(za - 1));
sum = sum + term;
term = (3*x*zr*(zr - 1)) / (za*(2*za^2 - 2*za + 1));
sum = sum + term;
oldratio = 1;
newratio = abs(term) / abs(sum);
m = 2;

while((oldratio > lim) || (newratio > lim))
        Qr = -(2*m - 1)*zr*Qr1 + Qr2;
        Qa = -(2*m - 1)*za*Qa1 + Qa2;
        P = ((2*m - 1)*x*P1 - (m - 1)*P2) / m;
        term = ((2*m + 1)*P*Qr)/((m + 1)*za*Qa - Qa1);
        sum = sum + term;
        m =  m + 1;
        Qr2 = Qr1; Qr1 = Qr;
        Qa2 = Qa1; Qa1 = Qa;
        P2 = P1; P1 = P;
        oldratio = newratio;
        newratio = abs(term)/abs(sum);
end

HRTF = (rho*exp(-j*mu)*sum) / (j*mu);
```

2.6.5. Modelos paramétricos de HRTFs usando una representación funcional mediante series de Fourier-Bessel

Para sintetizar una escena auditiva filtramos la señal monoaural con una base de datos de HRTFs. No obstante, para representar la escena completa una manera común de abordar el procesamiento es tener una representación funcional de las HRTFs, tal como modelos de bancos de filtros o descomposición mediante PCA o armónicos esféricos. Ya que los pesos de estos modelos solamente están disponibles bien para las direcciones medidas o bien para los puntos en frecuencia muestreados, las HRTFs se tienen que interpolar entre dos posiciones de medida o dos frecuencias, ambas discretas. Durante el desarrollo de la tecnología muchos autores han investigado cómo interpolar HRTFs, encontrando métodos como la interpolación bilineal, modelos de aproximación polos/ceros y métodos basados en "splines" esféricos. Otra manera de hacer esto es si tenemos una representación de las HRTFs mediante un funcional continuo, lo que evita tener que interpolar.

Una representación en forma de funcional continuo es un modelo matemático o una ecuación que representaría la HRTF como una función de variables continuas (posiciones de la fuente y puntos en frecuencia). El plano azimutal u horizontal se representa como HRTF(f, θ).

El modelo que aquí se describe es el modelo funcional basado en series Fourier Bessel para separar la dependencia de la parte espacial de la de la parte espectral [15].
Debido a la fuerte correlación entre los puntos medidos de las HRTFs y la familia de funciones de Bessel de orden 1, se usa una serie de este Fourier-Bessel para representar los pesos de los coeficientes de la serie de Fourier (las componentes espectrales de las HRTFs). Aplicando estas dos series, cada HRTF se transforma en una matriz de coeficientes que se puede usar muy eficientemente para ahorrar espacio y para el procesado para predecir cualquier HRTF en cualquier punto arbitrario en el plano $(r, \theta, \varphi = 0)$.

Una de las características notables de la HRTF es que es periódica con periodo 2π en la variable θ. Y una función periódica se expande naturalmente usando una serie de Fourier, usando ortogonalidad para separar una función periódica arbitraria en una serie convergente a la media. Cuando la serie se trunca, el resultado provee de una muy buena aproximación a la función original (siempre que la HRTF tengo una característica paso bajo en la variable θ, lo que suele ser común).

Así, representamos una HRTF en función de la frecuencia y el ángulo azimutal como [15]:

$$\mathbf{HRTF}(\mathbf{f},\boldsymbol{\theta}) = \sum_{\mathbf{m}=-\infty}^{\mathbf{M}=+\infty} \mathbf{A_m}(\mathbf{f}) \cdot \mathbf{e}^{\mathbf{jm\theta}}$$

$$\mathbf{A_m}(\mathbf{f}) = \frac{\mathbf{1}}{\mathbf{2\pi}} \int_{\mathbf{0}}^{\mathbf{2\pi}} \mathbf{HRTF}(\mathbf{f},\boldsymbol{\theta}) \cdot \mathbf{e}^{-\mathbf{jm\theta}} \mathbf{d\theta}$$

No hay una elección obvia para la representación de las componentes espectrales. Una estrategia practica efectiva es representar usando un conjunto completo de funciones ortogonales que puedan servir como funciones base hasta cualquier orden K para los coeficientes **{Am(f)}**:

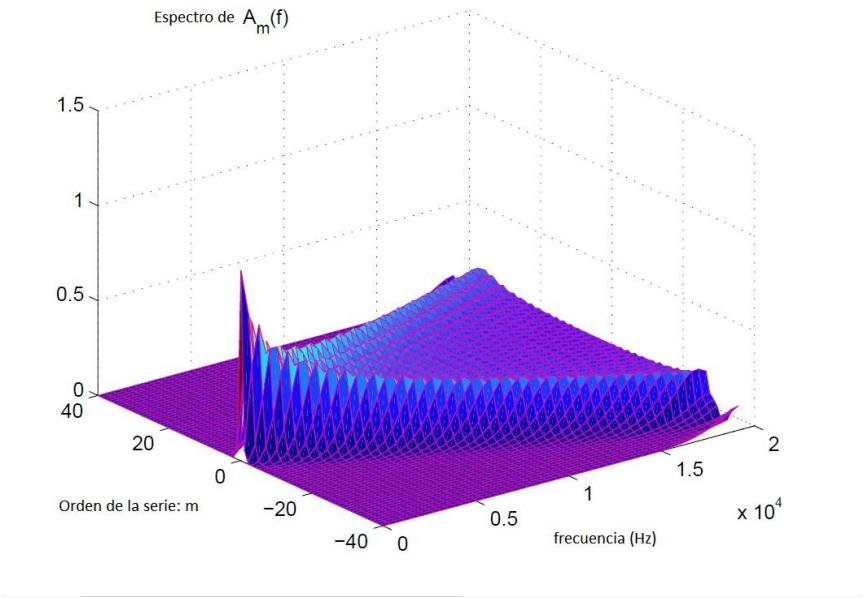
$$\mathbf{A_m}(\mathbf{f}) = \sum_{\mathbf{k}=\mathbf{1}}^{\mathbf{K}=+\infty} \mathbf{C_{mk}} \cdot \mathbf{B_k}(\mathbf{f})$$

donde **{Bk(f)}** es un conjunto de funciones ortogonales adecuado para la expansión, definido en el intervalo

$$\mathbf{f} \in [\,\mathbf{0},\ \mathbf{fmax}\,]$$

donde fmax es la frecuencia de medición máxima, usualmentenno más de 20kHz.
En principio cualquier conjunto ortogonal **{Bk(f)}** es válido; no obstante, debido a la truncatura impuesta, algunos conjuntos serán más apropiados que otros. Esto es: bajo truncatura, diferentes conjuntos ortogonales no serán equivalentes y habrá conjuntos preferidos para la representación.

En las siguiente figura se ha dibujado el espectro de Am(f) (calculado de un modelo simple analítico) para recalcar que *la energía de las componentes frecuenciales de las HRTFs no esta uniformemente distribuida sobre la expansión en serie de Fourier de orden m.* En vez de eso, el conjunto {Am(f)} es muy similar a una familia de funciones de Bessel de primera especie. Esto revela la fuerte correlación entre las componentes espectrales de las HRTFs y la familia de funciones de Bessel de primera especie.



Las series Fourier-Bessel usan la ortogonalidad inherente en las funciones de Bessel de primera especie, para un orden específico l en el intervalo (0, 1) , para expandir cualquier función. Esta expansión se escribe como:

$$\mathbf{A_m(f)} = \sum_{\mathbf{k=1}}^{\mathbf{K=+\infty}} \mathbf{C_{mk}} \cdot \mathbf{J_l}\left(\boldsymbol{\beta}_{\mathbf{k}}^{(\mathbf{l})} \frac{\mathbf{f}}{\mathbf{f_{max}}}\right)$$

Donde $\{\boldsymbol{\beta}_{\boldsymbol{k}}^{(\boldsymbol{l})}\}$ son las raíces positivas de $\boldsymbol{J_l(x) = 0}$ y (l) es el orden de la función de Bessel de primera especie. Los coeficientes de la expansión en serie de Fourier-Bessel se resuelven mediante:

$$\mathbf{C_{mk}} = \frac{\mathbf{2}}{\left(\mathbf{J_{l+1}}\left(\boldsymbol{\beta}_{\mathbf{k}}^{(\mathbf{l})}\right)\right)^{\mathbf{2}}} \int_{\mathbf{0}}^{\mathbf{f_{max}}} \mathbf{f} \cdot \mathbf{A_m(f)} \cdot \mathbf{J_l}\left(\boldsymbol{\beta}_{\mathbf{k}}^{(\mathbf{l})} \frac{\mathbf{f}}{\mathbf{f_{max}}}\right) \cdot \mathbf{df}$$

Todavía queda definir el orden l a usar. Ya que A0(f) es dominante, debemos incluir la funcion de Bessel J0(x). En el límite cuando f→0, solo $J0(0) = 1$ permite una aproximación eficiente.
Pero hay que incluir más términos para **{Am(f)}** con $m \neq 0$ para que la aproximación tenga sentido alguno. Notando que las series Fourier-Bessel de orden alto pueden corresponder a componentes de alta frecuencia, una representación l con dependencia $l = F(m)$ es más efectiva; una manera simple es

$$\mathbf{l = l(m) = |m|}$$

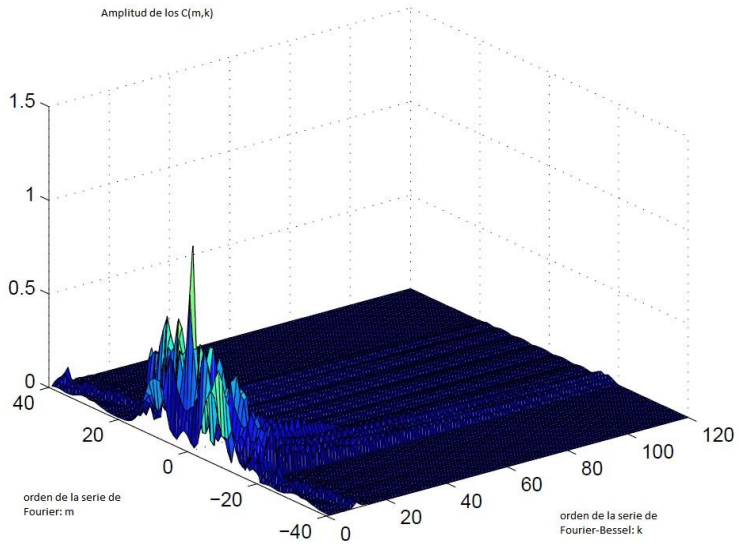
que permite una precisión de aproximación alta y reduce el número de parámetros.

Con todo, el modelo funcional de HRTFs se obtiene mediante

$$\begin{cases} \mathbf{HRTF(f,\theta)} = \sum_{\mathbf{m=-\infty}}^{\mathbf{M=+\infty}} \left(\sum_{\mathbf{k=1}}^{\mathbf{K=+\infty}} \mathbf{C_{mk}} \cdot \mathbf{J_{|m|}}\left(\boldsymbol{\beta}_{\mathbf{k}}^{(\mathbf{|m|})} \frac{\mathbf{f}}{\mathbf{f_{max}}}\right) \right) \cdot \mathbf{e^{jm\theta}} \\ \mathbf{C_{mk}} = \frac{\mathbf{1}}{\boldsymbol{\pi}\left(\mathbf{J_{|m+1|}}\left(\boldsymbol{\beta}_{\mathbf{k}}^{(\mathbf{|m|})}\right)\right)^{\mathbf{2}}} \int_{\mathbf{0}}^{\mathbf{f_{max}}} \int_{\mathbf{0}}^{\mathbf{2\pi}} \mathbf{f} \cdot \mathbf{HRTF(f,\theta)} \cdot \mathbf{J_{|m|}}\left(\boldsymbol{\beta}_{\mathbf{k}}^{(\mathbf{|m|})} \frac{\mathbf{f}}{\mathbf{f_{max}}}\right) \cdot \mathbf{e^{-jm\theta}} \cdot \mathbf{d\theta df} \end{cases}$$

donde $\{\boldsymbol{\beta_k}^{(|\mathbf{m}|)}\}$ son las raíces positivas de $\mathbf{J}_{|\mathbf{m}|}(\mathbf{x}) = \mathbf{0}$ y K es el orden de truncatura.
Este par de ecuaciones ilustra como calcular el modelo de parámetros a partir de una función HRTF continua.
Para HRTFs medidas experimentalmente, los coeficientes {C(m,k)} se calculan usando la suma de Riemann izquierda para aproximar la integral. Los parámetros fundamentales que influyen en una correcta adaptación del modelo son los *números de truncatura* M y K. La regla general es monitorizar la distribución de los **{C(m,k)}** sobre la serie de Fourier de orden m y la serie de Fourier-Bessel de orden K para cada conjunto específico de datos. Aquí se representa esto para un modelo analítico simple de HRTF (el mismo que el usado en la figura anterior). El criterio es que como mínimo el 90% de la energía total de los **{C(m,k)}** esté contenida en la aproximación. Los resultados demuestran que mínimos números de truncatura pueden ser perfectamente $M = 16$ y $K = 87$. Más información de cómo calcular los **{C(m,k)}** y K y M en [11].

Como resultado, por un lado el modelo propuesto puede conseguir una reconstrucción de HRTF para cualquier punto en frecuencia para un angulo azimutal θ arbitrario. Por otro lado, si todos los oyentes usan las mismas funciones base **{Bk(f)}**, para cada medición individualizada de HRTF, solamente se necesita guardar en memoria una matriz de coeficientes **[C(m,k)]**, que es muy inferior en tamaño a las de las medidas originales de la HRTF en cuestión. Con ello, la meta de compresión de datos para las bases de datos de HRTFs es alcanzada.



2.6.6. Modelos paramétricos de HRTFs mediante Análisis de Componentes Principales (PCA)

Siguiendo la discusión de esta sección, es importante volver a destacar que las HRTFs describen los cambios en la onda acústica conforme se propaga desde la fuente sonora al oído humano (específicamente, para las HRTFs este cambio de presión se mide en o cerca del tímpano). Otra posible representación, aparte de las anteriormente descritas, es usar *Análisis de Componentes Principales (Principal Component Analysis PCA)* que descompone los datos en componentes principales con sus correspondientes pesos como si de una combinación lineal se tratara [16].
Esta técnica se puede aplicar a bases de datos no individualizadas como la librería de HRTFs KEMAR del MIT Media Lab [12] u otras. Si las amplitudes lineales para elevación $\varphi = 0º$ se descomponen en cuatro componentes principales y cuatro pesos por amplitud [16], se puede encontrar mediante análisis una regularidad matemática en los pesos para diferentes azimutales θ. Esta regularidad es que la variación de cada peso se puede aproximar con una función matemática apropiada (senos o polinomios). Dicha representación como variaciones de pesos permite reconstruir las amplitudes de las HRTFs para posiciones que no fueron medidas y además reduce la información implícita.

PCA describe la base de datos original usando solamente unas cuantas componentes ortogonales con sus correspondientes pesos. Es una técnica no paramétrica, útil para reducir la dimensionalidad de bases de datos para compresión o para propósitos de comparación y coincidencia. El método extrae de una manera simple información relevante de conjuntos de datos difíciles de caracterizar. PCA compone un "mapa de carreteras" de manera que el ingeniero sepa cómo reducir un conjunto de datos complejo a dinámicas simplificadas que normalmente están ocultas en los datos.

Denotemos el conjunto de datos como **[S]**. Cada fila de [S] corresponde a todas las mediciones de un tipo particular Si:

$$[\mathbf{S}] = \begin{bmatrix} \mathbf{S1} \\ \dots \\ \mathbf{Sm} \end{bmatrix}$$

Cada columna de [S] corresponde a un conjunto de medidas de una medida en particular. De aquí se puede definir una *matriz de covariancia*:

$$[\mathbf{K_s}] = \frac{\mathbf{1}}{\mathbf{m-1}} \mathbf{S} \cdot \mathbf{S^T}$$

Para encontrar las direcciones donde la variación es máxima, se necesitan obtener los autovalores de la matriz [Ks]. Para extraer la mayor parte de la variación, se calcula una matriz de transformación lineal [PC] para los autovectores de mayor módulo.
Si se define ahora el vector:

$$\mathbf{PC} = \left[\mathbf{pc_1}, \mathbf{pc_2}, \dots, \mathbf{pc_j}\right]$$

los elementos de PC son las componentes principales de [S] y $j = 1, \dots, 4$ o mayor si se quieren contemplar mas componentes principales. La proyección del conjunto de datos originales [S] en el autoespacio ["Algebra y geometría", hernandez] es necesaria para obtener los j pesos apropiados {wj}:

$$\mathbf{w_j} = \mathbf{PC^T} \cdot \mathbf{S_j}$$

Para la reconstrucción, cada uno de los j pesos se multiplica por la componente principal correspondiente.
En el caso de considerar cuatro componentes, S vendría dado por:

$$\mathbf{S} = \sum_{\mathbf{m=1}}^{\mathbf{x}} \sum_{\mathbf{j=1}}^{\mathbf{4}} \mathbf{pc_j} \cdot \mathbf{w_{mj}}$$

De esta manera, la matriz [w(m,j)] especifica la contribución de cada pc a la reconstrucción de los datos. Interesa en este punto conseguir encontrar regularidades matemáticas implícitas en la base de datos de HRTFs que no se muestran de manera obvia usando la librería de HRTFs en su estado original.

Las componentes principales {pc1, pc2, pc3, pc4} son ortogonales y no tienen nada en común. Luego es interesante fijarse en las variaciones de los pesos {w1, w2, w3, w4}, que van a poder ser descritas usando

funciones matemáticas simples.
La variación del primer peso (el mas importante) w1 que se deriva del autovalor más grande, se puede describir correctamente con medio periodo de una función seno:

$$\mathbf{w_1(az)} = 5.1 \cdot \sin(\text{az}) + 0.212,\ 15 \leq \text{az} \leq 51$$

El valor de az varía entre $15 \leq \text{az} \leq 51$, donde $\text{az} = 15$ se corresponde con el azimutal -90º mientras que el valor $\text{az} = 51$ se corresponde con el azimutal $+90º$. La función seno adecuada puede determinarse usando un algoritmo especial basado en el criterio de mínimos cuadrados LMS entre los datos originales y los reconstruidos.

Las variaciones de los demás pesos {w2, w3, w4} no tienen una formulación tan simple. No obstante, pueden describirse con pocos coeficientes. Una aproximación polinómica de diferentes grados resulta adecuada. Cada variación sucesiva de los pesos {w2, w3, w4} necesita de un polinomio con grado cada vez mayor. Los grados se pueden calcular usando el criterio LMS. Las investigaciones realizadas en [] arrojan:

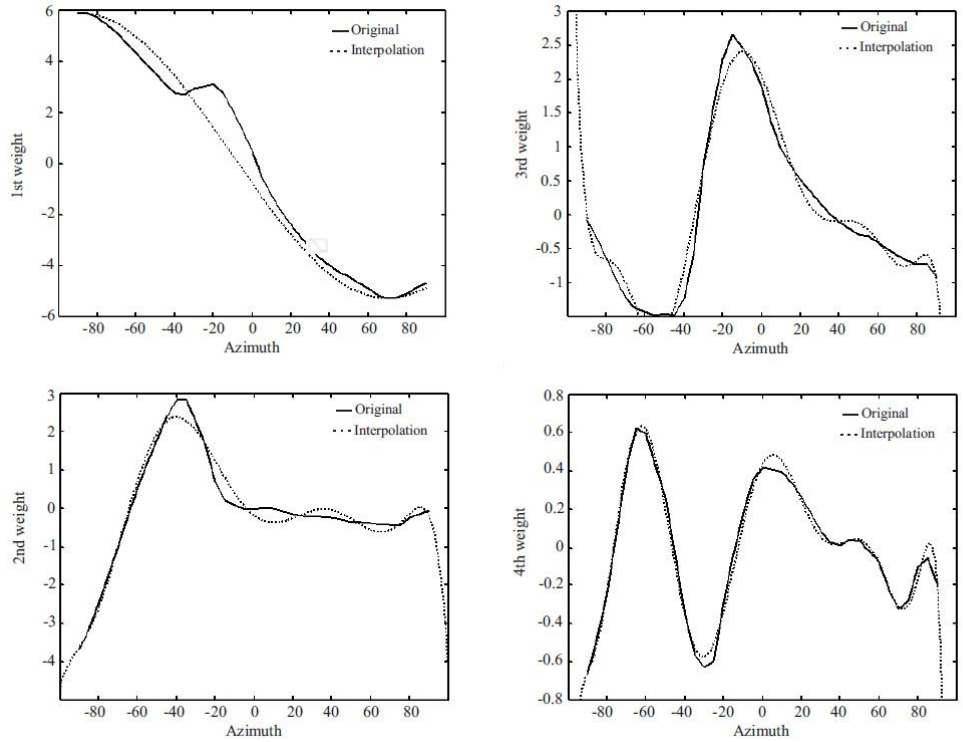
$$\mathbf{w_2(az)} = -2.7 \cdot 10^{-9} \cdot az^8 + 3.8 \cdot 10^{-7} \cdot az^7 - 2.2 \cdot 10^{-5} \cdot az^6 + 6.4 \cdot 10^{-4} \cdot az^5 + 0.0094 \cdot az^4 + 0.061 \cdot az^3 - 0.11 \cdot az^2 + 0.46 \cdot az - 4$$

$$\mathbf{w_3(az)} = -4.2 \cdot 10^{-10} \cdot az^9 + 7.3 \cdot 10^{-8} \cdot az^8 - 5.3 \cdot 10^{-6} \cdot az^7 + 2.1 \cdot 10^{-4} \cdot az^6 - 0.0048 \cdot az^5 + 0.063 \cdot az^4 + 1.9 \cdot az^2 - 3.8 \cdot az + 2.3$$

$$\mathbf{w_4(az)} = -10^{-11} \cdot az^{10} + 3.4 \cdot 10^{-9} \cdot az^9 - 2.7 \cdot 10^{-7} \cdot az^8 + 1.1 \cdot 10^{-5} \cdot az^7 - 2.9 \cdot 10^{-4} \cdot az^6 + 0.0045 \cdot az^5 - 0.04 \cdot az^4 + 0.18 \cdot az^3 - 0.38 \cdot az^2 + 0.48 \cdot az - 0.89$$

Todos los polinomios están en el intervalo $1 \leq \text{az} \leq 37$, donde $\text{az} = 1$ se corresponde con un azimutal -90º y $\text{az} = 37$ con $+90º$.

La figura siguiente muestra los cuatro pesos w1, w2, w3 y w4, para ángulos azimutales θ de -90º a $+90º$, para el conjunto de datos original y el interpolado. La reducción en los valores de los pesos, sobre todo para el primero w1 , salta a la vista.



2.6.7. Interpolación de HRTFs en distancia, azimutal y elevación

Aunque las bases de datos de HRTFs dependientes de la distancia proveen posibilidades interesantes, p.ej. el renderizado de fuentes en campo cercano, hay una falta de algoritmos y herramientas para manejarlas, en general.

En esta sección se quiere presentar la investigación desarrollada en [6] para integrar al software implementado la útil herramienta de interpolación desarrollada en [6], ya que parece prometedora cara a mejorar y actualizar el software de este Proyecto Fin de Carrera. Se propone un método para la interpolación de HRTFs en 3D (i.e. azimutal, elevación y distancia) usando interpolación tetraédrica con pesos baricéntricos. Para la interpolación, se genera una malla tetraédrica mediante un método de triangulación (triangulación de Delaunay) y se recorre la malla para búsquedas usando caminos de adyacencia [7], haciendo que esta herramienta sea robusta frente a HRTFs posicionadas irregularmente. Además, permite la interpolación directa de HRTFs mediadas a diferentes distancias. Y es computacionalmente eficiente.

El método de triangulación permite agrupar eficientemente HRTFs en conjuntos para interpolación, el uso de un algoritmo de búsqueda rápido para seleccionar una HRTF apropiada y el uso de interpolación tetraédrica con pesos baricéntricos para interpolar el conjunto de HRTFs en cuestión.

Dado un conjunto de HRTFs medidas, se puede obtener una HRTF interpolada a partir de cuatro puntos de medida formando un tetraedro que encierre la posicion de fuente deseada. La interpolación tetraédrica aquí descrita asume como punto de partida que una HRTF de estimación puede obtenerse para la posición de la fuente virtual deseada a partir de una interpolación de HRTFs adyacentes.

Triangulación de los puntos de medida

Un conjunto de puntos en 2D se pueden agrupar en triángulos que no se solapen usando una triangulación. Cuando se usan triángulos para interpolar, es deseable que sean lo más equiangulares posibles. La trangulacion de Delaunay es optima en este sentido y maximiza el angulo minimo de los triángulos generados. Para puntos en un plano, la triangulación de Delaunay genera triángulos tales que el circumcentro de cada trinagulo no contiene otros puntos.
Este concepto es aplicable a puntos espaciados irregularmente asi como a dimensiones superiores a 2D. En 3D, la triangulación de Delaunay genera tetraedros tales que la circumesfera de cada tetraedro no contiene otros puntos.

Cálculo de los pesos de interpolación

Una vez que se ha generado una malla tetraédrica de HRTFs mediante triangulación, un estimador de la HRTF deseada para cualquier punto $\mathbf{X}$ dentro de la malla puede ser obtenido interpolando los vértices del tetraedro que encierra a $\mathbf{X}$. Considerando un tetraedro de vértices $\mathbf{A, B, C}$ y $\mathbf{D}$, cualquier punto dentro de este tetraedro puede representarse mediante una combinación lineal de sus vértices:

$$\mathbf{X} = g_1 \cdot \mathbf{A} + g_2 \cdot \mathbf{B} + g_3 \cdot \mathbf{C} + g_4 \cdot \mathbf{D}$$

siendo {gi} pesos escalares. Con la condición adicional

$$\sum_{i=1}^{4} g_i = 1$$

los pesos {gi} son las *coordenadas baricéntricas* de $\mathbf{X}$. Estas coordenadas baricentricas pueden ser usadas directamente como los pesos de interpolación para estimar la HRTF deseada $\hat{\mathbf{H}}_{\mathbf{X}}$ para el punto $\mathbf{X}$ como la suma ponderada de HRTFs $\mathbf{H_i}$ medidas en $\mathbf{A, B, C}$ y $\mathbf{D}$:

$$\hat{\mathbf{H}}_{\mathbf{X}} = \sum_{i=1}^{4} g_i \cdot \mathbf{H_i}$$

Por otro lado, la ecuación $\mathbf{X} = g_1 \cdot \mathbf{A} + g_2 \cdot \mathbf{B} + g_3 \cdot \mathbf{C} + g_4 \cdot \mathbf{D}$ puede reformularse como:

$$\mathbf{X} - \mathbf{D} = [g_1 \ g_2 \ g_3] \cdot \mathbf{T}$$

con

$$\mathbf{T} = \begin{bmatrix} \mathbf{A} - \mathbf{D} \\ \mathbf{B} - \mathbf{D} \\ \mathbf{C} - \mathbf{D} \end{bmatrix}$$

Dada la posición deseada de la fuente virtual, **X**, los pesos de interpolación baricentricos se obtienen evaluando:

$$[g_1 \ g_2 \ g_3] = (\mathbf{X} - \mathbf{D}) \cdot \mathbf{T}^{-1}$$

Usando la condición de que la suma de los pesos sea 1, encontramos el ultimo peso, g4, que se obtiene mediante:

$$g_4 = 1 - g_1 - g_2 - g_3$$

Note el lector que T depende solamente de la geometría del tetaredro y es independiente de la posición de la fuente, X. Por ello, $\mathbf{T}^{-1}$ puede ser precalculado para todos los tetraedros y guardados en memoria en una fase de inicialización. Esto reduce considerablemente el numero de operaciones necesarias para el calculo de la HRTF interpolada.

Elección del tetraedro para la interpolación

Dadas la malla tetraédrica de HRTFs obtenida por triangulación y la posición deseada de la fuente virtual, **X**, un método de "fuerza bruta" para encontrar el tetraedro optimo para la interpolación es evaluando las coordenadas baricentricas de los tetraedros.
X esta dentro de un tetraedro si y solo si todas las coordenadas baricentricas son positivas; luego, una manera directa de encontrar el tetraedro apropiado para la interpolación es iterar por los tetraedros de la malla hasta que uno satisface la condición anterior (que todas las coordenadas baricentricas sean positivas).

Debido al gran numero de tetraedros generados para bases de datos de HRTFs densas y por las restricciones de tiempo de las aplicación de audio en tiempo real, es deseable acelerar el proceso de elección de un tetraedro para interpolar. Una manera mas eficiente que la anterior de localizar un punto en una triangulación es mediante un *camino de adyacencia.* Empezando por un tetraedro aleatorio, se evalúan las coordenadas baricentricas de este y se avanza al tetraedro adyacente por el triangulo formado por los vértices con las tres coordenadas baricentricas mayores. El paseo termina cuando todas las coordenadas baricentricas son positivas.

Una manera eficiente y simple de encontrar los vecinos mas cercanos a un punto en 3D es usando una representación de los puntos de medida de las HRTFs en forma de *octree.* Un cuboide conteniendo todos los puntos forma la raíz del octree. Empezando en el cuboide raíz, el octree es generado dividiendo recursivamente cada cuboide en ocho cuboides iguales en tamaño. La subdivisión de un cuboide acaba cuando contiene como mucho N puntos, haciendo de este una hoja del octree. N se elige para asegurar la resolución espacial deseada en el octree. Para encontrar el tetraedro mas cercano a la dirección de la fuente deseada, **X**, se busca por el octree para encontrar el cuboide hoja que encierre a **X**. Un tetraedro con un vértice contenido en este cuboide hoja esta cerca de **X** y por tanto puede elegirse como punto de partida para el camino de adyacencia, reduciendo asi el numero de iteraciones necesarias para terminar el paseo y el tiempo de computo del algoritmo.

Pre-procesado e incializacion

Para minimizar la carga computacional en tiempo de computo, el algoritmo pre-procesa ciertos parámetros en la inicialización. Primero, se realiza la triangulación de Delaunay y la malla tetraédrica resultante es guardada. Para cada tetraedro, se precalcula la inversa $\mathbf{T^{-1}}$ para acelerar el computo de los pesos baricentricos. Tambien se crea un mapa de adyacencia que lista los tetraedros adyacentes en la malla para permitir una selección rápida del tetraedro mediante el algoritmo de camino de adyacencia. Finalmente, se genera una representación en forma de octree de los puntos de medida de las HRTFs.

La complejidad teorica de este algoritmo es de $\mathbf{O(n^{1/3})}$, bastante menor que la de la fuerza bruta que es $\mathbf{O(n)}$.

2.7. Cancelación de Crosstalk. Audio transaural

2.7.1. Introducción a la cancelación de Crosstalk

Los sistemas de audio 3D deben permitir posicionar sonidos alrededor del oyente de manera que estos sean percibidos como prevenientes de puntos arbitrarios en el espacio. Esto no es posible con los sistemas estéreo clásicos.
Por ello, el audio 3D tiene el potencial de incrementar la sensación de realismo en música o películas. Es una útil herramienta para la realidad virtual, realidad aumentada, teleconferencias o entretenimiento para el hogar medio.
Como se ha visto, la percepción de sonido virtual es creada en el audio 3D sintetizando un par de señales binaurales a partir de una fuente monoaural con la información acústica 3D dada: la distancia (r) y la dirección (θ, φ) de la fuente sonora respecto al oyente. La percepción de dirección se puede conseguir usando las funciones de transferencia relacionadas con la cabeza (HRTFs) que se pueden obtener de bases de datos o modelos experimentales (ver sección 2.x.x.).

La cancelación de Crosstalk (de ahora en adelante cancelación de XT) fue desarrollada en los años 1960 por Atal y Schroeder [21] y Bauer [22]. Desde entonces, las técnicas han ido mejorando y los algoritmos se han sofisticado con el paso de los años. Para radiar las señales binaurales, lo más sencillo es usar altavoces. No obstante, en muchas aplicaciones p.ej. para el entretenimiento en el hogar o la teleconferencia para empresas, los oyentes pueden preferir no utilizar auriculares. Si se usan altavoces, la correcta transmisión de las señales binaurales de los altavoces a los oídos del oyente no es un proceso tan sencillo como para auriculares. Cada oído recibe algo de la llamada *componente de Crosstalk .* Las señales directas, además, son modificadas debido a la reverberación de la sala (que está en directa relación con la respuesta al impulso de la sala RIR). Para abordar estos problemas de diseño, se requiere de un filtro inverso antes de reproducir las señales binaurales con los altavoces.

La cancelación de XT se puede realizar directamente o adaptativamente. Suponiendo que se conocen de antemano los caminos de transferencia acústicos de altavoces a oídos, el metodo directo calcula el filtro de cancelación de XT invirtiendo directamente las funciones de transferencia acústicas. En el método adaptativo, el filtro de cancelación de XT es recalculado adaptativamente usando señales de realimentación obtenidas de unos micrófonos en miniatura posicionados en los oídos del oyente. No obstante, el método adaptativo sigue siendo más un tema de investigación mas que una solución real al problema del XT.

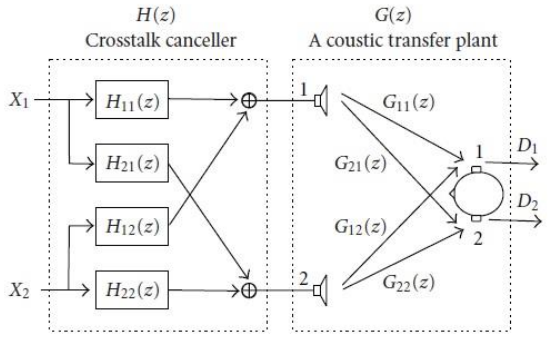
El audio que se obtiene al procesar el material binaural para que al escucharlo con altavoces no tenga problemas de componentes de XT se suele denominar *audio transaural.*

Una limitación fundamental en el sistema de cancelación de XT es el hecho de que cualquier movimiento que exceda los 75-100 mm destruye completamente la cancelación conseguida a la altura de los oídos del oyente. No obstante, este problema se puede resolver mediante “head-tracking” o seguimiento del movimiento de la cabeza del oyente usando un sistema de procesamiento de imágen/vídeo. La posición de la cabeza es grabada por una cámara, se calcula la posición del oyente y finalmente se recalculan los filtros HRTF y el filtro cancelador de XT en base a la nueva posición del oyente. Todo esto se ha de procesar y actualizar en tiempo real. Un ejemplo de este método de cancelación de XT se encuentra en la tesis doctoral de Gardner [23].

El software desarrollado para este Proyecto Fin de Carrera en concreto ha usado un sistema directo de cancelación de XT, suponiendo que el oyente se quedará en una posición fija escuchando el material de audio transaural y por tanto no hace uso del “head-tracking” (aunque se podría incorporar relativamente bien en una siguiente fase de diseño). En esta sección se expone el problema del cancelador de XT directo, con un oyente, fijo. Se proponen tres métodos diferentes, cada uno mejor (mayor atenuación de la componente de XT) y más eficiente (menor coste computacional) que el anterior.

2.7.2. Cancelador de Crosstalk convencional

Es común en la práctica usar sólo dos altavoces en un sistema estéreo [24]. Un diagrama de bloques de la implementación directa del cancelador de XT. Las señales binaurales yL(n) e yR(n) se denominan X1(z) y X2(z) y las señales de entrada a la altura de los tímpanos eL(n) y eR(n) se denominan D1(z) y D2(z) para el dominio que aquí nos interesa que es el de la transformada Z [Discrete-time signal Processing]. Por mera aclaración,1 = L y 2 = R.



El objetivo de la cancelación de XT es reproducir las señales binaurales a la altura de los tímpanos, esto es,

$$\mathbf{D(z) = z^{-d} \cdot X(z)}$$

donde d denota un retardo adicional debido a pasar el audio por el sistema de cancelación de XT. Este objetivo se consigue invirtiendo la función de transferencia del camino acústico seguido por la señal G(z) con el filtro de cancelación de XT H(z).

Para una mejor calidad, la respuesta inversa del altavoz también debería tenerse en cuenta y multiplicarse a la cadena cuando se diseñe el cancelador de XT. No obstante, esta parte se puede implementar por separado y no se ha considerado debido a la falta de medios para obtenerla en el momento de realizar el proyecto. Si se quisiera se podría medir la respuesta del altavoz y simplemente filtrar todo el audio (los canales X1 y X2) con la respuesta inversa.

G(z) y H(z) se escriben en forma matricial:

$$\mathbf{G(z)} = \begin{bmatrix} \mathbf{G11(z)} & \mathbf{G12(z)} \\ \mathbf{G21(z)} & \mathbf{G22(z)} \end{bmatrix}, \qquad \mathbf{H(z)} = \begin{bmatrix} \mathbf{H11(z)} & \mathbf{H12(z)} \\ \mathbf{H21(z)} & \mathbf{H22(z)} \end{bmatrix}$$

Aquí, Gij(z), (i, j) = 1, 2 es la función de transferencia acústica del altavoz j a la oreja i (luego siguiendo la nomenclatura usada hasta ahora, L = 1 y R = 2). Hij(z) es el filtro de cancelación de XT desde la señal Xj(z) al altavoz i.

Como se ha comentado, para asegurar que no exista XT, la funcion de transferencia global debe ser

$$\mathbf{D(z) = z^{-d} \cdot X(z) = G(z) \cdot H(z) \cdot X(z)}$$

Luego, se concluye que

$$\mathbf{z^{-d} \cdot I = G(z) \cdot H(z)}$$

$$\mathbf{H(z) = z^{-d} \cdot G^{-1}(z)}$$

El término de retardo es necesario para garantizar que H(z) sea *causal* y por tanto físicamente realizable.

Hay que notar que no se puede conseguir una cancelación perfecta. Esto se debe a que G(z) es normalmente una función de fase no mínima, en cuyo caso un *algoritmo LMS* aproxima el filtro inverso óptimo G-1(z).

El algoritmo LMS se realiza como sigue. Supóngase que los vectores
$\mathbf{g_{ij}} = [\mathbf{g_{ij,0}},\ \mathbf{g_{ij,1}},\ \dots,\ \mathbf{g_{ij,L_g-1}}]$

$\mathbf{h_{ij}} = [\mathbf{h_{ij,0}},\ \mathbf{h_{ij,1}},\ \dots,\ \mathbf{h_{ij,L_h-1}}]$

representan la respuesta al impulso de Gij(z) (de longitud Lg) y la respuesta al impulso de Hij(z) (de longitud Lh). Reescribiendo la ecuación para la cancelación se tiene

$$\begin{bmatrix} \widetilde{\mathbf{G11}} & \widetilde{\mathbf{G12}} \\ \widetilde{\mathbf{G21}} & \widetilde{\mathbf{G22}} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{h11} & \mathbf{h12} \\ \mathbf{h21} & \mathbf{h22} \end{bmatrix} = \begin{bmatrix} u_d & \mathbf{0} \\ \mathbf{0} & u_d \end{bmatrix}$$

O en forma compacta:

$$\boxed{\mathbf{G \cdot H = U}}$$

Aquí,

$$u_d = [0, \dots, 0, 1, 0, \dots, 0]^T$$

es un vector cuya d-ésima componente es 1, $\mathbf{0}$ es un vector de longitud L1, que contiene solo ceros, con $L1 = Lh + Lg - 1$ y $\widetilde{G_{ij}}$:

$$\widetilde{\mathbf{G_{ij}}} = \begin{bmatrix} \mathbf{g_{ij,0}} & \cdots & \mathbf{g_{ij,L_g-1}} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{g_{ij,0}} & \cdots & \mathbf{g_{ij,L_g-1}} & \dots & \mathbf{0} \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{g_{ij,0}} & \cdots & \mathbf{g_{ij,L_g-1}} \end{bmatrix}^T$$

$\widetilde{G_{ij}}$ es una matriz de convolucion de tamaño L1xLh, poniendo en cascada los vectores gij. La solución LMS a la ecuación $G \cdot H = U$ es:

$$\boxed{\mathbf{H_{LMS} = G^{+}U}}$$

donde G^+ es la *pseudoinversa* de G, definida como

$$\mathbf{G^{+} = (G^{T} \cdot G + \beta \cdot I)^{-1} \cdot G^{T}}$$

siendo β un parámetro de regularización para aumentar la robustez de la inversión. El filtro de cancelación de XT se obtiene de $\mathbf{H_{LMS} = G^{+}U}$ y es de longitud $Lh1 = Lh$.

La matriz de los caminos acústicos G(z) depende de la posición del oyente y más concretamente de la posición de la cabeza. Si el oyente se moviese, seria necesario actualizar G(z) y recalcular H(z) (en tiempo real). Esto se puede conseguir con un sistema de head-tracking como el que se presenta en el anexo "HeaDSPeaker". La carga de cómputo se vuelve muy grande cuando G(z) es de tamaño grande.

En la siguiente pagina se muestra cómo implementar un cancelador de XT sencillo en Matlab® usando la base de datos de HRTFs que se genera de la función "simpleHRIR.m" descrita en la sección 2.x.x. Este método funciona bien para altavoces cerca el uno del otro. Un ángulo muy grande hará que este sistema colorearee el espectro a bajas frecuencias, modificando el audio original. El área donde el efecto de cancelación de XT es audible es muy pequeña, pues si el oyente se mueve de la línea media entre los altavoces 1 o 2 cm el efecto se pierde.

Lo bueno de esta técnica es sobre todo la externalización que produce y que falta en la escucha con auriculares. No obstante, es muy complicado obtener un efecto de envolvimiento con esta técnica. Con un dipolo estéreo al frente, la escena sonora reproducida se percibe normalmente solo como si estuviera de frente.

Además, las reflexiones y reverberaciones de la sala de escucha también afectan al grado de eficiencia de este sistema de cancelación de XT. Funciona mejor en espacios sin reflexiones prominentes. Además, los mejores resultados dependerán también de si se conocen o simulan/modelan las HRTFs del oyente (no obstante se obtienen resultados bastante plausibles con HRTFs genéricas)

```
function result = simplified_XTC(theta_0,FILEin, Fs)
%Cancelador de Crosstalk simplificado
% Fuentes:
% DAFx: Digital Audio Effects, 2nd ed., Zoelzer, U.
%
% Usa las HRTFs de simpleHRIR.m

%theta_0 = 10;
%Fs = 44100;
b = 10^-5;
H(1,1,:) = simpleHRIR(theta_0/2,Fs)
H(1,2,:) = simpleHRIR(-theta_0/2,Fs)
H(2,1,:) = H(1,2,:);
H(2,2,:) = H(1,1,:);

hrir_length = length(H(1,1,:));

% Trasladar al dominio frecuencial
for i=1:2
        for j=1:2
                C_f(i,j,:) = fft(H(i,j,:),hrir_length);
        end
end

%Inversion de la matriz C  mediante regularización Moore-Penrose
H_f=zeros(2,2,hrir_length);

for k=1:hrir_length
        H_f(:,:,k)=inv((C_f(:,:,k)'*C_f(:,:,k)+eye(2)*b))*C_f(:,:,k)';
end

% Vuelta al dominio temporal
for k=1:2
        for m=1:2
                H_n(k,m,:)=real(ifft(H_f(k,m,:)));
                H_n(k,m,:)=fftshift(H_n(k,m,:));
        end
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Generar señales binaurales
input = audioread(FILEin, Fs);

% Convolucionar para obtener las señales de alimentacion de los altavoces
if( min(size(input)) == 2) % si estereo
        L = conv(reshape(H_n(1,1,:),hrir_length,1),input(:,1)) +
conv(reshape(H_n(2,1,:),hrir_length,1),input(:,1));
        R = conv(reshape(H_n(1,2,:),hrir_length,1),input(:,2)) +
conv(reshape(H_n(2,2,:),hrir_length,1),input(:,2));
        result(:,1) = L;
        result(:,2) = R;
else
        L = conv(squeeze(H_n(1,1,:)),input(:,1)) + conv(squeeze(H_n(2,1,:)),input(:,1));
        R = conv(squeeze(H_n(1,2,:)),input(:,1)) + conv(squeeze(H_n(2,2,:)),input(:,1));
        result(:,1) = L;
        result(:,2) = R;
end

end
```

2.7.3. Cancelador de Crosstalk usando un filtro de Wiener

Otra manera de definir un cancelador de XT es calcular la inversa de G mediante una estructura de filtro único, que hace que la cantidad de computo sea menor [24]. Para obtenerlo, reescribimos H(z):

$$\mathbf{H(z) = z^{-d} \cdot G^{-1}(z) \rightarrow H(z) = \frac{z^{-d}\begin{bmatrix} G22(z) & -G12(z) \\ -G21(z) & G11(z) \end{bmatrix}}{G11(z) \cdot G22(z) - G12(z) \cdot G21(z)}}$$

Definiendo:

$$\mathbf{Q(z) = det(G(z)) = |G(z)| = G11(z) \cdot G22(z) - G12(z) \cdot G21(z)}$$

y:

$$\mathbf{T(z) = \frac{z^{-d}}{Q(z)}}$$

el problema de invertir G(z) se convierte en

$$\boxed{\mathbf{Q(z) \cdot T(z) = z^{-d} \cdot I}}$$

Definiendo los vectores:

$\mathbf{q} = \left[\mathbf{q_0},\ \mathbf{q_1},\ \dots,\ \mathbf{q_{L_q-1}}\right]^{\mathbf{T}}$

$\mathbf{t} = \left[\mathbf{t_0},\ \mathbf{t_1},\ \dots,\ \mathbf{t_{L_t-1}}\right]^{\mathbf{T}}$

la respuesta al impulso de Q(z) (de longitud Lq = 2·Lg – 1) y siendo t la respuesta al impulso de T(z) (de longitud Lt,), reescribiendo la ecuación anterior en el dominio del tiempo tenemos:

$$\boxed{\mathbf{Q \cdot t = u_d}}$$

Aquí,

$$Q = \begin{bmatrix} q_0 & \dots & q_{L_q-1} & 0 & \dots & 0 \\ 0 & q_0 & \dots & q_{L_q-1} & \dots & 0 \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & q_0 & \dots & q_{L_q-1} \end{bmatrix}^T$$

es una matriz de convolucion de tamaño L2xLt, con L2 = Lt + Lq – 1.
La solución LMS a este nuevo problema es

$$\boxed{\mathbf{t_{LMS} = Q^{+} \cdot u_d}}$$

donde la pseudoinversa de Q vale:

$$\mathbf{Q^{+} = (Q^T \cdot Q + \beta \cdot I)^{-1} \cdot Q^T}$$

siendo β un parámetro de regularización para aumentar la robustez de la inversión. El filtro de cancelación de XT se obtiene del par:

$$\begin{cases} \mathbf{H(z) = z^{-d} \cdot G^{-1}(z) = \dfrac{z^{-d}\begin{bmatrix} G22(z) & -G12(z) \\ -G21(z) & G11(z) \end{bmatrix}}{G11(z) \cdot G22(z) - G12(z) \cdot G21(z)} \\ \mathbf{t_{LMS} = Q^{+} \cdot u_d} \end{cases}$$

con una longitud del filtro cancelador de XT Lh2 = Lt + Lg – 1.
Combinando G(z) y H(z), obtenemos la funcion de transferencia global:

$$\mathbf{F(z) = G(z) \cdot H(z) = T(z)\begin{bmatrix} G11(z) & G12(z) \\ G21(z) & G22(z) \end{bmatrix} \cdot \begin{bmatrix} G22(z) & -G12(z) \\ -G21(z) & G11(z) \end{bmatrix}}$$

$$\mathbf{F(z) = T(z)} \begin{bmatrix} G11(z)G22(z) & 0 \\ -G12(z)G21(z) & \\ 0 & G11(z)G22(z) \\ & -G12(z)G21(z) \end{bmatrix}$$

Los términos antidiagonales de F(z) son siempre 0 independientemente del valor de T(z). Esto delata que el XT se suprime casi completamente; no obstante, debido al efecto de filtrado de los términos diagonales de F(z), se introduce cierta coloración y distorsión al reproducir las señales. Este es la desventaja inherente de una estructura de filtro con filtro único.

Cancelaciones de XT basadas en otros modelos, p.ej. la basaba en modelos con *polos/ceros acústicos comunes CAPZ* o la cancelación óptima de XT desarrollada en la universidad de Princeton [] son más efectivas e introducen menor distorsión y coloración. A continuación se describe la cancelación de XT mediante modelos CAPZ.

2.7.4. Cancelador de Crosstalk basado en modelos CAPZ

La función de transferencia acústica es usualmente un modelo todo ceros, cuyos coeficientes son su respuesta al impulso. No obstante, cuando la duración de la respuesta al impulso es larga, se necesitan muchos parámetros para representar a la función. Esto aumenta el coste computacional para la cancelación de Crosstalk, a sumar al ya alto coste de la síntesis binaural. Los modelos polos-ceros pueden reducir el coste computacional, pero los polos y ceros de dichos sistemas cambian si la función de transferencia acústica varía. Esto es más inconveniente desde el punto de vista de la inversión del camino acústico [24].

Se ha probado que una HRTF contiene implícita el sistema de resonancia del canal auditivo cuyas frecuencias de resonancia y factores Q son independientes de las direcciones de la fuente.

Por ello, la HRTF se puede modelar eficientemente usando polos ($\vec{\mathbf{a}}$) independientes de las direcciones de la fuente, que representan las frecuencias de resonancia factores Q. Se usarán sin embargo ceros ($\vec{\mathbf{b}}$) dependientes de las direcciones de la fuente.
Este modelo se denomina *modelo de polos/ceros acústicos comunes CAPZ (common-acoustical pole/zero model CAPZ).* Los modelos CAPZ comparten un conjunto de polos y procesan los ceros por separado. Esto obviamente reduce el numero de parámetros con respecto al modelo convencional polo/cero. Cuando se quiere aproximar una función de transferencia acústica Hi(z) con un modelo CAPZ, se expresa como

$$\mathbf{H_I(z)} = \frac{\mathbf{B_i(z)}}{\mathbf{A_i(z)}} = \frac{\sum_{\mathbf{n=0}}^{\mathbf{Nq}} \mathbf{b_{n,i}} \cdot \mathbf{z^{-n}}}{\mathbf{1} + \sum_{\mathbf{n=1}}^{\mathbf{Np}} \mathbf{a_n} \cdot \mathbf{z^{-n}}}$$

donde Np es el numero de polos y Nq el número de ceros. Los vectores

$$\vec{\mathbf{a}} = \left[\mathbf{1},\ \mathbf{a_1},\ \dots,\ \mathbf{a_{Np}}\right]^{\mathbf{T}}$$
$$\vec{\mathbf{b}}_{\mathbf{i}} = \left[\mathbf{b_{1,i}},\ \mathbf{b_{2,i}},\ \dots,\ \mathbf{b_{Nq,i}}\right]^{\mathbf{T}}$$

son los vectores de coeficientes de los polos y los ceros, respectivamente.
Los parámetros para el modelo CAPZ pueden ser estimados usando un método de mínimos cuadrados LMS, pero también otras aproximaciones son útiles. Aquí se desarrolla el modelo CAPZ usando LMS.

Para obtener los parámetros CAPZ usando aproximación LMS, supóngase un conjunto de K funciones de transferencia. Entonces, el error de modelado total se define como:

$$\mathbf{E} = \sum_{\mathbf{i=1}}^{\mathbf{K}} \sum_{\mathbf{n=0}}^{\mathbf{N-1}} |\mathbf{e_i(n)}|^{\mathbf{2}} = \sum_{\mathbf{i=1}}^{\mathbf{K}} \sum_{\mathbf{n=0}}^{\mathbf{N-1}} \left| \mathbf{h_i(n)} + \sum_{\mathbf{j=1}}^{\mathbf{Np}} \mathbf{a_j} \cdot \mathbf{h_i(n-j)} - \sum_{\mathbf{j=0}}^{\mathbf{Nq}} \mathbf{b_{j,i}} \cdot \boldsymbol{\delta}\mathbf{(n)} \right|^{\mathbf{2}}$$

donde N es la longitud de e(n) y hi(n) es la respuesta al impulso de Hi(z) original.

Para hallar el vector $\vec{\mathbf{a}}$ y los vectores $\vec{\mathbf{b}}_\mathbf{i}$, i = 1, 2, …, K, minimizamos el error E y obtenemos el siguiente sistema de ecuaciones matricial:

$$\begin{bmatrix} \mathbf{I} & \mathbf{H_{o,1}} \\ \mathbf{0} & \mathbf{H_1} \end{bmatrix} \begin{bmatrix} \mathbf{b_1} \\ -\mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{r_{o,1}} \\ \mathbf{r_1} \end{bmatrix}$$

…

$$\begin{bmatrix} \mathbf{I} & \mathbf{H_{o,K}} \\ \mathbf{0} & \mathbf{H_K} \end{bmatrix} \begin{bmatrix} \mathbf{b_K} \\ -\mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{r_{o,K}} \\ \mathbf{r_K} \end{bmatrix}$$

Aquí, I es la matriz identidad, los vectores $\overrightarrow{\mathbf{r_{o,i}}}$ y $\vec{\mathbf{r}}_\mathbf{i}$ se definen con i = 1, 2, …, K:

$\overrightarrow{\mathbf{r_{o,i}}} = [\mathbf{h_i(0)}\ \mathbf{h_i(1)}\ \ldots\ \mathbf{h_i(Nq)}]^T$

$\vec{\mathbf{r}}_\mathbf{i} = [\mathbf{h_i(Nq+1)}\ \mathbf{h_i(Nq+2)}\ \ldots\ \mathbf{h_i(N-1)}]^T$.

Las matrices $\mathbf{H_{o,i}}$ y $\mathbf{H_i}$ son ambas matrices de convoluciones que se forman poniendo en cascada las funciones hi(n):

$$\mathbf{H_{o,i}} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \mathrm{h_i}(0) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{h_i}(Nq-1) & \mathrm{h_i}(Nq-2) & \cdots & \mathrm{h_i}(Nq-Np) \end{pmatrix}$$

$$\mathbf{H_i} = \begin{pmatrix} \mathrm{h_i}(Nq) & \cdots & \mathrm{h_i}(Nq-Np+1) \\ \vdots & \ddots & \vdots \\ \mathrm{h_i}(N-2) & \cdots & \mathrm{h_i}(N-1-Np) \end{pmatrix}$$

Usando el sistema de ecuaciones anterior, se obtienen el vector de polos y los vectores de ceros mediante:

$$\begin{cases} \vec{\mathbf{a}} = -\left(\widetilde{\mathbf{H}}^\mathbf{T} \cdot \widetilde{\mathbf{H}}\right)^{-1} \cdot \widetilde{\mathbf{H}}^\mathbf{T} \cdot \mathbf{R} \\ \overrightarrow{\mathbf{b_i}} = \mathbf{H_{o,i}} \cdot \vec{\mathbf{a}} + \mathbf{r_{o,i}}, \qquad \mathbf{i = 1, 2, \ldots, K} \end{cases}$$

siendo el vector $\mathbf{R} = [\mathbf{r_1}\ \mathbf{r_2}\ \ldots\ \mathbf{r_K}]^\mathbf{T}$ y la matriz $\widetilde{\mathbf{H}} = [\mathbf{H_1}\ \mathbf{H_2}\ \ldots\ \mathbf{H_K}]^\mathbf{T}$.

Suponiendo que el camino de transferencia acústico G(z) es conocido, los modelos CAPZ para las funciones de transferencia acústicas de los altavoces a los oídos son directamente:

$$\mathbf{G_{11}(z)} = \frac{\mathbf{B_{11}(z)}}{\mathbf{A(z)}} \mathbf{z^{-d_{11}}}, \qquad \mathbf{G_{12}(z)} = \frac{\mathbf{B_{12}(z)}}{\mathbf{A(z)}} \mathbf{z^{-d_{12}}}$$

$$\mathbf{G_{21}(z)} = \frac{\mathbf{B_{21}(z)}}{\mathbf{A(z)}} \mathbf{z^{-d_{21}}}, \qquad \mathbf{G_{22}(z)} = \frac{\mathbf{B_{22}(z)}}{\mathbf{A(z)}} \mathbf{z^{-d_{22}}}$$

Donde d11, d12, d21 y d22 son los retardos de transmisión de los altavoces a los oídos. Sustituyendo este conjunto de ecuaciones en la ecuación obtenida

$$\mathbf{H(z)} = \mathbf{z^{-d}} \cdot \mathbf{G^{-1}(z)} = \frac{\mathbf{z^{-d}} \begin{bmatrix} \mathbf{G22(z)} & \mathbf{-G12(z)} \\ \mathbf{-G21(z)} & \mathbf{G11(z)} \end{bmatrix}}{\mathbf{G11(z) \cdot G22(z) - G12(z) \cdot G21(z)}}$$

se ve que:

$$\mathbf{H(z)} = \frac{\mathbf{z^{-d}}}{\mathbf{B_{11}(z) \cdot B_{22}(z) \cdot z^{-(d_{11}+d_{22})} - B_{12}(z) \cdot B_{21}(z) \cdot z^{-(d_{12}+d_{21})}}} \times$$
$$\times \begin{bmatrix} \mathbf{B_{22}(z) \cdot A(z) \cdot z^{-d_{22}}} & \mathbf{-B_{12}(z) \cdot A(z) \cdot z^{-d_{12}}} \\ \mathbf{-B_{21}(z) \cdot A(z) \cdot z^{-d_{21}}} & \mathbf{B_{11}(z) \cdot A(z) \cdot z^{-d_{11}}} \end{bmatrix}$$

Sin perdida de generalidad se asume $\mathbf{d_{11} + d_{22} < d_{12} + d_{21}}$. Definiendo:

$$\Delta = \mathbf{(d_{11} + d_{22}) - (d_{12} + d_{21})}$$

Sustituyendo en la expresión para H(z), se obtiene:

$$\mathbf{H(z)} = \frac{\mathbf{z^{-(d-d_{11}-d_{22})}}}{\mathbf{B_{11}(z)\cdot B_{22}(z) - B_{12}(z)\cdot B_{21}(z)\cdot z^{-\Delta}}} \times$$

$$\times \begin{bmatrix} \mathbf{B_{22}(z)\cdot A(z)\cdot z^{-d_{22}}} & \mathbf{-B_{12}(z)\cdot A(z)\cdot z^{-d_{12}}} \\ \mathbf{-B_{21}(z)\cdot A(z)\cdot z^{-d_{21}}} & \mathbf{B_{11}(z)\cdot A(z)\cdot z^{-d_{11}}} \end{bmatrix}$$

Y de forma más compacta:

$$\mathbf{H(z)} = \frac{\mathbf{z^{-\delta}}}{\mathbf{B(z)}} \times \begin{bmatrix} \mathbf{B_{22}(z)\cdot A(z)\cdot z^{-d_{22}}} & \mathbf{-B_{12}(z)\cdot A(z)\cdot z^{-d_{12}}} \\ \mathbf{-B_{21}(z)\cdot A(z)\cdot z^{-d_{21}}} & \mathbf{B_{11}(z)\cdot A(z)\cdot z^{-d_{11}}} \end{bmatrix}$$

con $\mathbf{B(z) = B_{11}(z)\cdot B_{22}(z) - B_{12}(z)\cdot B_{21}(z)\cdot z^{-\Delta}}$, $\boldsymbol{\delta = d - d_{11} - d_{22}}$. Si por último se escribe el primer término para H(z) como $\mathbf{C(z)} = \frac{\mathbf{z^{-\delta}}}{\mathbf{B(z)}}$ entonces:

$$\boxed{\mathbf{H(z) = C(z)} \begin{bmatrix} \mathbf{B_{22}(z)\cdot A(z)\cdot z^{-d_{22}}} & \mathbf{-B_{12}(z)\cdot A(z)\cdot z^{-d_{12}}} \\ \mathbf{-B_{21}(z)\cdot A(z)\cdot z^{-d_{21}}} & \mathbf{B_{11}(z)\cdot A(z)\cdot z^{-d_{11}}} \end{bmatrix}}$$

y el problema de invertir el camino acústico G(z) se convierte en

$$\boxed{\mathbf{B(z)\cdot C(z) = z^{-\delta}\cdot I}}$$

Sea $\vec{\mathbf{b}} = [\mathbf{b_1},\ \mathbf{b_2},\ \dots,\ \mathbf{b_{Lb-1}}]^{\mathbf{T}}$ la respuesta al impulso de B(z), con
$\mathbf{Lb = 2\cdot(Nq+1) + \Delta - 1}$ y sea $\vec{\mathbf{c}} = [\mathbf{c_1},\ \mathbf{c_2},\ \dots,\ \mathbf{c_{Lc-1}}]^{\mathbf{T}}$ la respuesta al impulso de C(z). En el dominio del tiempo y con estas definiciones, la ecuación anterior toma la forma:

$$\boxed{\mathbf{B}\cdot\vec{\mathbf{c}} = \boldsymbol{u_\delta}}$$

Aquí, $\mathbf{B}$ es una matriz de convolución de tamaño L3 x Lc, con L3 = Lb + Lc – 1,

$$\mathbf{u_\delta} = [\mathbf{0}, \dots, \mathbf{0}, \mathbf{1}, \mathbf{0}, \dots, \mathbf{0}]^{\mathbf{T}}$$

es un vector columna de longitud L3 cuya δ-ésima componente es igual a 1 y la matriz se describe mediante**:**

$$\mathbf{B} = \begin{bmatrix} \mathbf{b_0} & \dots & \mathbf{b_{L_b-1}} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{b_0} & \dots & \mathbf{b_{L_b-1}} & \dots & \mathbf{0} \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{b_0} & \dots & \mathbf{b_{L_b-1}} \end{bmatrix}^{\mathbf{T}}$$

Ya que B(z) es normalmente de fase no mínima, la solución LMS a la ecuación

$$\mathbf{B}\cdot\vec{\mathbf{c}} = \boldsymbol{u_\delta}$$

es:

$$\boxed{\mathbf{c_{LMS} = B^{+}\cdot u_\delta}}$$

donde la pseudoinversa de B vale

$$\mathbf{B^{+} = (B^{T}\cdot B + \beta\cdot I)^{-1}\cdot B^{T}}$$

siendo β el parámetro de regularización. Finalmente, el cancelador de XT se obtiene del par:

$$\begin{cases} \mathbf{H(z) = C(z)} \begin{bmatrix} \mathbf{B_{22}(z)\cdot A(z)\cdot z^{-d_{22}}} & \mathbf{-B_{12}(z)\cdot A(z)\cdot z^{-d_{12}}} \\ \mathbf{-B_{21}(z)\cdot A(z)\cdot z^{-d_{21}}} & \mathbf{B_{11}(z)\cdot A(z)\cdot z^{-d_{11}}} \end{bmatrix} \\ \mathbf{c_{LMS} = B^{+}\cdot u_\delta} \end{cases}$$

La longitud del filtro cancelador de XT es:

$$\mathbf{L_{h3} = Lc + (Nq+1) + (Np+1) + \max\{d11, d12, d21, d22\} - 1}$$

$$\mathbf{L_{h3} = Lc + Nq + Np + d_{max} + 1}$$

con $\mathbf{d_{max} = \max\{d11, d12, d21, d22\}}$.

Es útil especificar la selección del numero de polos y ceros (Np, Nq). Cuantos mas polos y ceros sean usados en el modelo, mejor aproximación resultara de el. Por otro lado, mas parámetros requerirán mas tiempo de computo. Por ello, se tiene que encontrar un balance entre el tiempo de computo y el orden de la aproximación.

2.8. Síntesis de Campos de Onda (WFS)

Wavefield Synthesis (WFS) es una técnica para la reproducción de un campo sonoro en el espacio. Utiliza un gran numero de altavoces para crear una escena auditiva virtual sobre un área amplia de la sala. Una ventaja a destacar sobre el audio 3D es que salva los problemas de reproducción estereofónica, entre ellos cabe destacar el sweet spot que en WFS es un área mucho más amplia.

El primer esbozo de lo que ahora se conoce como WFS fue presentado en [2] a mediados de los años 1950. No obstante el sistema no fue puesto en marcha debido a las limitaciones tecnológicas de la época en el uso de un gran número de altavoces. Los autores del artículo usaron por ello solo unos cuantos altavoces, fundando no obstante las bases teóricas de las técnicas de reproducción estereofónicas.

La base teórica moderna de lo que hoy se conoce como WFS fue formulada en la universidad de Delft, Holanda, a finales de los años 1980 en [35]. No obstante el término parece ser acuñado por primera vez en 1993 [36]. También en esos años se llevan a cabo los primeros experimentos de laboratorio. Varios proyectos de investigación asi como tesis doctorales se han escrito en el contexto de la WFS [38].

2.8.1. Sistemas de coordenadas

Para una correcta descripción de la propagación del sonido en el espacia se requiere de una formulación de los campos y procesos acústicos involucrados en tres dimensiones (3D). No obstante, en muchos casos los altavoces o fuentes o transmisores y los oyentes o receptores (nótese la pluralidad de receptores; para WFS no se considera solamente un receptor) están situados en un plano. En estos casos, una descripción bidimensional (2D) es más apropiada [38].

Las coordenadas a usar serán las cartesianas y/o las esféricas. En el caso bidimensional, se definen como:

$$\vec{\mathbf{x}} = \begin{bmatrix} \mathrm{x} \\ \mathrm{y} \end{bmatrix}, \qquad \vec{\mathbf{r}} = \begin{bmatrix} \mathrm{r} \\ \theta \end{bmatrix}$$

Las componentes están relacionadas según:

$$\begin{bmatrix} \mathrm{x} \\ \mathrm{y} \end{bmatrix} = \mathrm{r} \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}, \qquad \begin{bmatrix} \mathrm{r} \\ \theta \end{bmatrix} = \begin{bmatrix} \sqrt{\mathrm{x}^2 + \mathrm{y}^2} \\ \tan^{-1}(\mathrm{x}/\mathrm{y}) \end{bmatrix}$$

y los elementos infinitesimales para integración son:

$$\mathrm{d}\vec{\mathbf{x}} = \mathrm{dx}\,\mathrm{dy}, \qquad \mathrm{d}\vec{\mathbf{r}} = \mathrm{r}\,\mathrm{dr}\,\mathrm{d}\theta$$

En el caso tridimensional, el vector de posición en coordenadas cartesianas, denotado por la letra z, se define:

$$\vec{\mathbf{z}} = \begin{bmatrix} \vec{\mathbf{x}} \\ \mathrm{z} \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

y el elemento infinitesimal de volumen para integración es:

$$\mathrm{d}\vec{\mathbf{z}} = \mathrm{dx}\,\mathrm{dy}\,\mathrm{dz}$$

El operador $(\cdot)$ denota la multiplicación de escalares ordinaria.

2.8.2. Ecuación de Onda y solución de onda plana

La ecuación de onda está dada por [37, 38]:

$$\boxed{\Delta p(t,\vec{z}) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} p(t,\vec{z}) = 0}$$

Siendo $p(t,\vec{z})$ la presión sonora en el tiempo t y en el punto $\vec{z}$. El operador $(\Delta = \nabla^2)$ es el operador de Laplace (derivada segunda con respecto al espacio) y c es la velocidad de propagación de la onda, que para el caso del sonido en el aire es aprox. 334 m/s. Las soluciones posibles a la ecuación de onda están restringidas a señales con las derivadas parciales de segundo orden iguales, tanto temporales como espaciales. Las soluciones a la ecuación de onda se suelen llamar *campos de onda o campos sonoros.* Como toda ecuación en derivadas parciales, estará sujeta a alguna condición de contorno.
Al aplicar el operador de Fourier en el tiempo para ecuación de onda, se obtiene la *ecuación de Helmholtz:*

$$\boxed{\Delta P(\omega,\vec{z}) + \left(\frac{\omega}{c}\right)^2 \cdot P(\omega,\vec{z}) = 0}$$

Para esta expresión se ha usado el teorema de diferenciación de la transformada de Fourier para sustituir la derivada segunda temporal en la ecuación de onda por $(j\omega)^2$. La relación entre frecuencia y velocidad de propagación se denomina *número de onda* , se denota por la letra k y vale:

$$\mathbf{k = \omega / c}$$

Una onda plana es una solución especial de la ecuación de onda; tiene una forma sencilla si se expresa en coordenadas cartesianas. Queda determinada por su forma de onda y por la dirección desde la cual se generó. La forma de onda se caracteriza por una funcion temporal $f(t,\zeta)$ y la dirección por el vector unitario $\overrightarrow{\mathbf{n}_\zeta}$. Cuando se revisen las ondas planas con componente z igual a cero, dicho vector estará contenido en el plano (x,y) y quedará unívocamente determinado por sus componentes (x,y):

$$\overrightarrow{\mathbf{n}_\zeta} = \begin{bmatrix}\cos(\zeta)\\ \sin(\zeta)\end{bmatrix}$$

La solución de onda plana a la ecuación de Helmholtz es la señal tridimensional

$$p(t,\vec{\mathbf{x}}) = f\left(t + \frac{1}{c}\left\langle \vec{\mathbf{x}}, \overrightarrow{\mathbf{n}_\zeta}\right\rangle, \zeta\right)$$

donde $\left\langle \vec{x}, \overrightarrow{\mathbf{n}_\zeta}\right\rangle$ denota el producto escalar de ambos vectores. La función $p(t,\vec{x})$ describe un frente de onda plano que se propaga por el espacio desde la dirección $\overrightarrow{\mathbf{n}_\zeta}$ con velocidad c. Se comprueba que:

$$p(t,\vec{\mathbf{x}} = \mathbf{0}) = f(t,\zeta)$$

luego en el origen se puede observar directamente la forma de onda original. La transformada de Fourier respecto al tiempo de la solución de onda plana a la ecuación de Helmholtz es:

$$P(\omega,\vec{\mathbf{x}}) = F(\omega,\zeta)\cdot e^{j\cdot\frac{\omega}{c}\cdot\left\langle \vec{\mathbf{x}},\overrightarrow{\mathbf{n}_\zeta}\right\rangle}$$

Es posible obtener un campo de onda P más genérico mediante superposición de soluciones en θ para todas las direcciones posibles ζ:

$$P(\omega,\vec{\mathbf{r}}) = \int_0^{2\pi} F(\omega,\zeta)\cdot e^{j\cdot\frac{\omega}{c}\cdot r\cdot\cos(\zeta-\theta)}\cdot d\zeta$$

Aquí, se ha expresado el producto escalar en coordenadas polares: $\left\langle \vec{\mathbf{x}}, \overrightarrow{\mathbf{n}_\zeta}\right\rangle = r\cdot\cos(\zeta-\theta)$
Esta formulación integral esta estrechamente relacionada con la descomposición en ondas planas de un campo de onda.

2.8.3. Funciones de Green

Función de Green para una fuente puntual

Para describir soluciones arbitrarias de la ecuación de onda con condiciones de contorno homogéneas, lo más efectivo es usar funciones de Green. Estas funciones se pueden interpretar como la respuesta de un campo sonoro a un impulso en el espacio-tiempo. Hay impulsos de varios tipos: aquí se consideran funciones de Green para fuentes puntuales y también para fuentes lineales.

Una fuente puntual en el espacio se define por la función impulso de Dirac tridimensional:

$$\delta_{3D}(\vec{z}) = \delta(x) \cdot \delta(y) \cdot \delta(z)$$

siendo δ() la funcion impulso de Dirac unidimensional. Una fuente puntual situada en $\overrightarrow{z'}$ con fuerza de fuente variable en el tiempo se describe mediante

$$q_0(t,\vec{z}) = q_0\left(t,\overrightarrow{z'}\right)\delta_{3D}\left(\vec{z}-\overrightarrow{z'}\right), \qquad Q_0(\omega,\vec{z}) = Q_0\left(\omega,\overrightarrow{z'}\right)\delta_{3D}\left(\vec{z}-\overrightarrow{z'}\right)$$

El subíndice 0 indica que las fuentes puntuales tienen dimensión cero. La distribución espacial arbitraria de fuentes puntuales queda determinada por la ecuación

$$Q_0(\omega,\vec{z}) = \iiint_V Q_0\left(\omega,\overrightarrow{z'}\right)\delta_{3D}\left(\vec{z}-\overrightarrow{z'}\right)\cdot d\overrightarrow{z'}$$

El campo sonoro P generado por una distribución de fuentes $Q_0(\omega,\vec{z})$ en el espacio es;

$$P_0(\omega,\vec{z}) = \iiint_V G_0\left(\omega,\vec{z}\middle|\overrightarrow{z'}\right)\cdot Q_0\left(\omega,\overrightarrow{z'}\right)\cdot d\overrightarrow{z'}$$

La *funcion de Green* $G_0\left(\omega,\vec{z}\middle|\overrightarrow{z'}\right)$ describe la contribución al campo sonoro en un punto z (llamado *posición del oyente*) de una fuente puntual en la posición z'. La integración se lleva a cabo en el volumen V donde se considera la solución de la ecuación de onda.
La forma de la funcion de Green depende del volumen V y del tipo de condiciones de contorno en su superficie. *En campo libre, $V = R^3$, la funcion de Green 3D para cualquier tipo de condición de contorno* es:

$$G_0{}^f\left(\omega,\vec{z}\middle|\overrightarrow{z'}\right) = \frac{1}{4\pi}\frac{e^{-j\cdot\frac{\omega}{c}|\vec{z}-\overrightarrow{z'}|}}{\left|\vec{z}-\overrightarrow{z'}\right|}$$

Esta funcion describe una onda esférica; el denominador modela la atenuación de la amplitud debida a la distancia y el termino exponencial modela el retardo temporal de la onda esférica que se propaga.

Función de Green para una fuente lineal

Una fuente lineal consiste en una superposición de fuentes puntuales iguales a lo largo de una recta. Cuando la recta se orienta paralela al eje z, todas las fuentes puntuales con mismas coordenadas (x, y) tienen la misma fuerza $Q_0(\omega,\vec{z})$. Luego la fuerza no depende de la coordenada z y la integral para la fuerza se simplifica a una integración en el plano x-y:

$$Q_0(\omega,\vec{z}) = \iiint_V Q_0\left(\omega,\overrightarrow{z'}\right)\cdot\delta_{3D}\left(\vec{z}-\overrightarrow{z'}\right)\cdot d\overrightarrow{z'} = \iint_L Q_1\left(\omega,\overrightarrow{x'}\right)\cdot\delta_{2D}\left(\vec{x}-\overrightarrow{x'}\right)\int_{-\infty}^{+\infty}\delta(z-z')\cdot dz'd\overrightarrow{x'} \equiv Q_1(\omega,\vec{x})$$

donde L es un corte horizontal a lo largo del volumen V.

Este resultado se puede interpretar de dos maneras:

· Como antes, $Q_0(\omega,\vec{z})$ describe un conjunto de fuentes puntuales (subíndice cero). Las componentes (x, y, z) denotan la posición de cada fuente puntual en el espacio.
Eso sí, la fuerza de fuente es constante en la dirección z con lo que el resultado no depende de la coordenada z.

· $Q_1(\omega, \vec{x})$ describe un conjunto de fuentes lineales paralelas al eje z. El subíndice 1 denota el carácter unidimensional de las fuentes lineales. Las dos componentes (x, y) denotan las coordenadas de los puntos de partida de cada línea en el plano x-y.

$Q_0(\omega, \vec{z})$ es una funcion de tres variables espaciales (x, y, z) que denota la posición de una entidad cero-dimensional (una fuente puntual) en el espacio. Por otro lado, $Q_1(\omega, \vec{x})$ es una función de dos variables espaciales (x, y) que denota la posición de una entidad 1-dimensional (una fuente paralela al eje z). Luego tanto $Q_0(\omega, \vec{z})$ como $Q_1(\omega, \vec{x})$ describen un campo sonoro 3D especial, que no depende de la coordenada z.

El campo sonoro generado por un conjunto de fuentes lineales $Q_1(\omega, \vec{x})$ puede obtenerse de la ecuación del campo sonoro P generado por una distribución de fuentes obtenida anteriormente:

$$P_0(\omega, \vec{z}) = \iiint_V G_0\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) \cdot Q_0\left(\omega, \overrightarrow{z'}\right) \cdot d\overrightarrow{z'}$$

Para el ecaso de que $Q_0(\omega, \vec{z})$ no depende de la coordenada z. La integración entonces sólo aplica a la función de Green y el resultado es la *función de Green de una fuente lineal* paralela al eje z*:*

$$G_1\left(\omega, \vec{x}\middle|\overrightarrow{x'}\right) = \int_{-\infty}^{+\infty} G_1\left(\omega, \vec{x}\middle|\overrightarrow{z'}\right) dz'$$

Al evaluar la integral para el caso de la función de Green de campo libre con dependencia exponencial, se obtiene el útil resultado:

$$G_1{}^f\left(\omega, \vec{x}\middle|\overrightarrow{x'}\right) = \frac{-j}{4} H_0{}^{(2)}\left(\frac{\omega}{c}\rho\right)$$

siendo $H_0{}^{(2)}\left(\frac{\omega}{c}\rho\right)$ la función de Hankel de segunda especie y orden cero:

$$H_0{}^{(2)}(a) = J_0(a) - j \cdot N_0(a)$$

con Jo() y No() funciones de Bessel y Neumann de primera especie y orden cero. La letra ρ denota la distancia entre oyente y fuente lineal:

$$\rho = \left|\vec{\mathbf{x}} - \overrightarrow{\mathbf{x'}}\right| = \sqrt{(x - x')^2 + (y - y')^2}$$

Debido a la simetría circular, esta función depende solo de la distancia entre la posición del oyente $\vec{\mathbf{x}}$ y la de la fuente lineal $\overrightarrow{\mathbf{x'}}$.

La notación puede acortarse un poco si se desea:

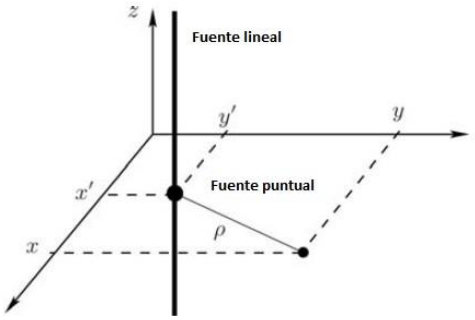
$$G_1{}^f\left(\omega, \vec{x}\middle|\overrightarrow{x'}\right) = \widetilde{G_1{}^f}(\omega, \rho) = \frac{-j}{4} H_0{}^{(2)}\left(\frac{\omega}{c}\rho\right)$$

El campo resultante es constante a lo largo de z y queda descrito por

$$P_1(\omega, \vec{x}) = \iint_L G_1\left(\omega, \vec{x}\middle|\overrightarrow{x'}\right) \cdot Q_1\left(\omega, \overrightarrow{x'}\right) \cdot d\overrightarrow{x'}$$

2.8.4. Relación entre las funciones de Green para fuentes puntuales y lineales en el caso de campo libre

Llegados a este punto interesa presentar la relación entre una fuente lineal paralela al eje z y una fuente puntual situada en el punto de partida de la fuente lineal. La orientación junto con el sistema de coordenadas se muestra en la siguiente figura. El efecto de ambas fuentes sobre el campo sonoro en el plano x-y merece ser comparado.



El efecto de la fuente lineal se ha demostrado queda descrito por:

$$\widetilde{G_1}^{f}(\omega,\rho) = \frac{-j}{4} H_0^{(2)}\left(\frac{\omega}{c}\rho\right)$$

La funcion de Green 3D para una fuente puntual en campo libre estaba dada por:

$$G_0{}^{f}\left(\omega,\vec{z}\middle|\vec{z'}\right) = \frac{1}{4\pi}\frac{e^{-j\cdot\frac{\omega}{c}|\vec{z}-\vec{z'}|}}{\left|\vec{z}-\vec{z'}\right|}$$

Y en este caso el efecto de la fuente puntual queda descrito por esta ecuación para ($z = 0, z' = 0$):

$$G_0{}^{f}\left(\omega,\vec{x}\middle|\vec{x'}\right) = \frac{1}{4\pi}\frac{e^{-j\cdot\frac{\omega}{c}|\vec{x}-\vec{x'}|}}{\left|\vec{x}-\vec{x'}\right|} = \frac{1}{4\pi}\frac{e^{-j\cdot\frac{\omega}{c}\rho}}{\rho} \triangleq \widetilde{G_0}^{f}(\omega,\rho)$$

La relación entre el campo sonoro para una fuente puntual en el plano x-y y una fuente lineal paralela al eje z queda establecida por una aproximación, que se puede llevar a cabo de dos maneras: mediante el *método de fase estacionaria* o el *método de aproximación de campo lejano* .
El desarrollo de estos métodos muestra que la fuente lineal puede aproximarse por una fuente puntual; en concreto, puede aproximarse mediante [38]

$$\boxed{\widetilde{G_1}^{f}(\omega,\rho) = H(\omega)A(\rho)\widetilde{G_0}^{f}(\omega,\rho)}$$

con

$$H(\omega) = \sqrt{\frac{c}{j\omega}}, \qquad A(\rho) = \sqrt{2\pi\rho}$$

2.8.5. La integral de Kirchhoff-Helmholtz para un volumen 3D general

La integral de Kirchhoff-Helmholtz es el elemento clave y fundamental de la síntesis de campos de onda. Esta integral provee de la relación necesaria entre el campo sonoro dentro de un volumen con forma arbitraria y el contorno que lo encierra.

La *ecuación integral de Helmholtz o integral de Kirchhoff-Helmholtz* da los valores del campo sonoro $P_0(\omega, \vec{z})$ en un volumen V mediante una integral de la superficie de V (denotada ∂V):

$$-\oiint_{\partial V}\left(\mathbf{G_0}\left(\boldsymbol{\omega}, \vec{\mathbf{z}}\middle|\overrightarrow{\mathbf{z'}}\right)\cdot\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\vec{\mathbf{n}}}\mathbf{P_0}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{z'}}\right)-\mathbf{P_0}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{z'}}\right)\cdot\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\vec{\mathbf{n}}}\mathbf{G_0}\left(\boldsymbol{\omega}, \vec{\mathbf{z}}\middle|\overrightarrow{\mathbf{z'}}\right)\right)\cdot \mathbf{d}\overrightarrow{\mathbf{z'}} = \begin{cases}\mathbf{P_0}(\boldsymbol{\omega}, \vec{\mathbf{z}}), & \vec{\mathbf{z}} \in \mathbf{V} \\ \mathbf{0}, & \mathbf{resto}\end{cases}$$

$\mathbf{G_0}\left(\boldsymbol{\omega}, \vec{\mathbf{z}}\middle|\overrightarrow{\mathbf{z'}}\right)$ es una funcion de Green que satisfaga las condiciones de contorno pertinentes en ∂V.

Esta integral expresa el hecho de que, en cualquier punto dentro de la región V (libre de fuentes), la presión sonora $\mathbf{P_0}(\boldsymbol{\omega}, \vec{\mathbf{z}})$ pueda obtenerse si ambas cantidades $\mathbf{P_0}(\boldsymbol{\omega}, \vec{\mathbf{z}})$ y su gradiente direccional

$$\frac{\partial}{\partial \vec{n}} P_0(\omega, \vec{z}) = \langle\, \nabla P_0(\omega, \vec{z}), \vec{n}\,\rangle$$

fueran conocidas en el contorno ∂V que encierra al volumen V. El contorno no tiene por qué ser necesariamente una superficie física real.

Esta integral tiene su aplicación, en su mayoría, en tres áreas:
· el cálculo de un campo sonoro radiado por una superficie vibrante hacia una cierta región
· el cálculo de un campo sonoro dentro de una región finita, producido por una fuente fuera del volumen, mediante mediciones de la superficie
· el control acústico sobre un campo sonoro dentro de un volumen.

La tercera aplicación esta en relación directa con el uso de WFS para acústica de salas y reproducción de sonido.

2.8.6. La integral de Kirchhoff-Helmholtz para un prisma

A continuación se particulariza la integral para campos sonoros que no dependen de la coordenada z. La forma del volumen de integración ∂V se convierte en un prisma orientado paralelamente al eje z. Ya que el campo sonoro, se asume, es independiente de z, se deduce que $\mathbf{P_0}(\boldsymbol{\omega}, \vec{\mathbf{z}})$ depende solo de las coordenadas (x, y). Además, cualquier vector normal al contorno tiene componente z igual a cero y también su gradiente direccional es independiente de z. Es por esto que la integral con respecto a $\overrightarrow{\mathbf{z'}}$ se puede dividir en una integración de un contorno ∂L respecto a $\overrightarrow{\mathbf{x'}}$ y en una integración con respecto a z'.
El contorno ∂L se define por la intersección del prisma con el plano x-y.
El proceder así transforma el primer término de la integral de Kirchhoff-Helmholtz en:

$$\oiint_{\partial V} G_0\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right)\cdot\frac{\partial}{\partial \vec{n}}P_0\left(\omega, \overrightarrow{z'}\right)\cdot \mathrm{d}\overrightarrow{z'}$$
$$= \oint_{\partial L}\int_{-\infty}^{+\infty} G_0\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right)\cdot\frac{\partial}{\partial \vec{n}}P_0\left(\omega, \overrightarrow{z'}\right)\cdot \mathrm{d}z'\mathrm{d}\overrightarrow{x'} = \oint_{\partial L}\left(\int_{-\infty}^{+\infty} G_0\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right)\mathrm{d}z'\right)\frac{\partial}{\partial \vec{n}}P_0\left(\omega, \overrightarrow{z'}\right)\mathrm{d}\overrightarrow{x'}$$

Usando la relación obtenida:

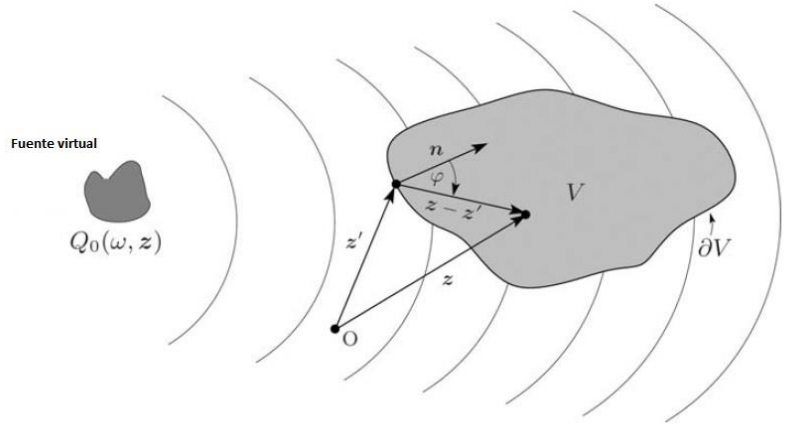
$$G_1(\omega, \vec{x}|\overrightarrow{x'}) = \int_{-\infty}^{+\infty} G_0(\omega, \vec{x}|\overrightarrow{z'}) \cdot dz'$$

se consigue una versión 2D de la integral de Kirchhoff-Helmholtz [38]:

$$\mathbf{P_1}(\boldsymbol{\omega}, \vec{\mathbf{z}}) = -\oint_{\partial L} \mathbf{G_1}\left(\boldsymbol{\omega}, \vec{\mathbf{x}}\middle|\overrightarrow{\mathbf{x'}}\right) \cdot \frac{\partial}{\partial \vec{\mathbf{n}}} \mathbf{P_1}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{x'}}\right) - \mathbf{P_1}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{x'}}\right) \cdot \frac{\partial}{\partial \vec{\mathbf{n}}} \mathbf{G_1}\left(\boldsymbol{\omega}, \vec{\mathbf{x}}\middle|\overrightarrow{\mathbf{x'}}\right) d\overrightarrow{\mathbf{x'}}$$

2.8.7. Reproducción de sonido basada en la integral de Kirchhoff-Helmholtz. El principio de Huygens

En esta sección se describe la reproducción del sonido para el escenario de la siguiente figura. Se desea reproducir el campo de onda emitido por una fuente virtual arbitraria $Q_0(\omega, \vec{z})$ en la región acotada V (*región de escucha*) donde se encuentran los oyentes. El uso de una sola fuente virtual no restringe el campo de onda a reproducir, ya que la fuente puede tener características cualesquiera tanto de forma como de respuesta en frecuencia. Además, si se quieren reproducir múltiples fuentes, se puede aplicar el principio de superposición y sumar así las contribuciones de cada fuente para obtener el campo de onda final.



El *principio de Huygens* es el punto de partida para describir la reproducción del sonido en esta situación. Huygens formuló que cualquier punto de un frente de onda propagándose para cualquier instante de tiempo se puede describir como la envolvente de ondas esféricas provenientes de cada punto del frente de onda en el instante anterior.
Este teorema puede usarse para sintetizar frentes de onda acústicos con forma arbitraria. La base matemática de esta descripción para la reproducción de sonido la brinda la integral de Kirchhoff-Helmholtz. En las subsecciones siguientes se usará para presentar una teoría de los sistemas de reproducción de WFS.

La integral de Kirchhoff-Helmholtz modela una multitud de problemas físicos caracterizados por sus condiciones de contorno específicas (y por tanto por sus correspondientes funciones de Green).
Para el escenario presentado en la figura anterior, la funcion de Green y su gradiente direccional se relacionan directamente con el campo emitido por las fuentes posicionadas en el contorno ∂V. Estas fuentes que aparecen en el borde del volumen debido al principio de Hyugens se suelen llamar *fuentes secundarias.* La fuerza de estas fuentes está determinada por la presión $P_0\left(\omega, \overrightarrow{z'}\right)$ del campo virtual $Q_0\left(\omega, \overrightarrow{x'}\right)$ en el borde ∂V y su gradiente direccional $\frac{\partial}{\partial \vec{n}} P_0\left(\omega, \overrightarrow{z'}\right)$.

Por todo esto, la integral de Kirchhoff-Helmholtz puede interpretarse como sigue. Si se eligen las fuentes secundarias adecuadamente y se alimentan por la presión sonora y el gradiente de presión direccional del campo emitido por la fuente virtual $Q_0\left(\omega, \overrightarrow{x'}\right)$, el campo de onda dentro del volumen V es equivalente al campo de onda que hubiera sido generado por la misma fuente virtual situada dentro del volumen V.

Este fenómeno pone de manifiesto que las bases teóricas para la reproducción de sonido están descritas por la ecuación de Kirchhoff-Helmholtz. En el contexto de la reproducción de sonido, la funcion de Green 3D en campo libre puede interpretarse como el campo de una distribución de fuentes puntuales monopolo sobre la superficie ∂V.
La integral también usa el gradiente direccional de la funcion de Green 3D. En el mismo contexto, puede interpretarse como el campo de una fuente dipolar con el eje principal en la direccion del vector normal $\vec{\mathbf{n}}$.
La integral de Kirchhoff-Helmholtz dice en estos términos que la presión acústica dentro de V puede controlarse mediante una distribución de fuentes puntuales monopolo y dipolo sobre la superficie ∂V.

Esta interpretación de la integral de Kirchhoff-Helmholtz bosqueja una primera aproximación para un sistema para reproducción de audio 3D. En resumen, dicho sistema consistiría en aproximaciones de monopolos y dipolos acústicos mediante altavoces adecuados. Estos altavoces cubrirían la superficie de un volumen elegido adecuadamente para que encierre las posibles posiciones de escucha. Los altavoces serían alimentados por las funciones adecuadas con el fin de reproducir el campo sonoro dentro del volumen.

No obstante, hay algunas cuestiones técnicas fundamentales a resolver para implementar exitosamente el sistema de reproducción de audio:

- · Fuentes monopolo y dipolo
- · Reduccion a dos dimensiones espaciales
- · Muestreo espacial
- · Señales de alimentación

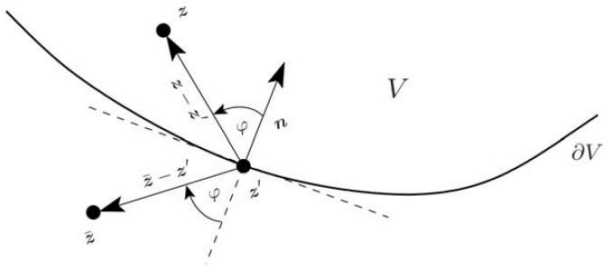
Estas se revisan individualmente a continuación.

2.8.8. Fuentes monopolo y dipolo

El uso de fuentes tipo monopolo y dipolo en la integral de Kirchhoff-helmholtz permite una reproducción bastante precisa del campo de onda deseado; es $P_0(\omega, \vec{z})$ para todo punto dentro del volumen V y cero fuera de él. No obstante, tal restricción no suele ser necesaria para la reproducción espacial de sonido. Con que la reproducción sea correcta dentro de V, lo que pasa en su exterior no suele ser relevante (siempre que el volumen de sonido no sea muy alto y moleste a alguien). Por ello se puede establecer el siguiente compromiso: usar solamente uno y solo un tipo de fuentes y tolerar algo de sonido fuera del volumen V.

Para conseguir este compromiso, se construye una funcion de Green $G_0\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right)$ que satisfaga las condiciones de contorno (de primera o segunda especie) en el borde ∂V. Ya que es deseable usar monopolos en vez de dipolos, se elige la funcion de Green 3D de una fuente puntual $G_0{}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right)$. Para cada posición z' en el borde, se superponen la funcion de Green para una posición $\mathbf{z}$ dentro de V, $G_0{}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right)$ y la funcion de Green para una posición $\bar{\mathbf{z}}(\vec{\mathbf{z}})$ fuera de V, $G_0{}^{f}\left(\omega, \bar{z}(\vec{z})\middle|\overrightarrow{z'}\right)$:

$$\mathbf{G_0}\left(\boldsymbol{\omega}, \vec{\mathbf{z}}\middle|\overrightarrow{\mathbf{z'}}\right) = \mathbf{G_0}^{\mathbf{f}}\left(\boldsymbol{\omega}, \vec{\mathbf{z}}\middle|\overrightarrow{\mathbf{z'}}\right) + \mathbf{G_0}^{\mathbf{f}}\left(\boldsymbol{\omega}, \bar{\boldsymbol{z}}(\vec{\mathbf{z}})\middle|\overrightarrow{\mathbf{z'}}\right)$$



Como muestra la figura, $\bar{\mathbf{z}}(\vec{\mathbf{z}})$ se elige como la imagen simétrica de $\vec{\mathbf{z}}$ respecto al plano tangente en $\overrightarrow{\mathbf{z}'}$ en ∂V. El plano tangente se caracteriza por el vector unitario $\vec{\mathbf{n}}$. La construcción de una funcion de Green de acuerdo a este criterio induce condiciones de borde de segunda especie (tipo Neumann) en la superficie ∂V. Los efectos de estas condiciones de contorno son tenidos en cuenta para el calculo de las señales de alimentación a los altavoces (ver subsección sobre este tema mas adelante).

Para aproximar monopolos y dipolos acústicos se usan altavoces de diferentes tipos. La restricción a un solo tipo de altavoz es ventajosa para una implementación técnica.
La notación $\bar{\mathbf{z}}(\vec{\mathbf{z}})$ denota que $\bar{\mathbf{z}}$ depende de $\vec{\mathbf{z}}$. Debido a la simetría,

$$|\vec{\mathbf{z}} - \overrightarrow{\mathbf{z}'}| = |\bar{\mathbf{z}}(\vec{\mathbf{z}}) - \overrightarrow{\mathbf{z}'}| = \rho_z$$

y por ello la dependencia con $\vec{\mathbf{z}}$ y $\overrightarrow{\mathbf{z}'}$ es idéntica para ${G_0}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right)$ y ${G_0}^{f}\left(\omega, \bar{z}(\vec{z})\middle|\overrightarrow{z'}\right)$:

$${G_0}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) = {G_0}^{f}\left(\omega, \bar{z}(\vec{z})\middle|\overrightarrow{z'}\right)$$

No obstante se debe distinguir entre ambas funciones para posiciones dentro y fuera de V. Esta diferencia se vuelve evidente al calcular los gradientes:

$$\boldsymbol{\nabla {G_0}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) = -\frac{1 + jk\rho_z}{\rho_z} {G_0}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) \overrightarrow{n_z}, \qquad \overrightarrow{n_z} = \frac{\vec{z} - \overrightarrow{z'}}{|\vec{z} - \overrightarrow{z'}|}}$$

$$\boldsymbol{\nabla {G_0}^{f}\left(\omega, \bar{z}\middle|\overrightarrow{z'}\right) = -\frac{1 + jk\rho_z}{\rho_z} {G_0}^{f}\left(\omega, \bar{z}\middle|\overrightarrow{z'}\right) \overrightarrow{n_{\bar{z}}}, \qquad \overrightarrow{n_{\bar{z}}} = \frac{\bar{z}(\vec{z}) - \overrightarrow{z'}}{|\bar{z}(\vec{z}) - \overrightarrow{z'}|}}$$

La derivada con respecto a $\vec{\mathbf{n}}$ está dada por:

$$\boldsymbol{\frac{\partial}{\partial \vec{n}} {G_0}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) = \left\langle \nabla {G_0}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right), \vec{n} \right\rangle = -\frac{1 + jk\rho_z}{\rho_z} \cdot {G_0}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) \cdot \left\langle \overrightarrow{n_z}, \vec{n} \right\rangle}$$

y de manera similar para ${G_0}^{f}\left(\omega, \bar{z}(\vec{z})\middle|\overrightarrow{z'}\right)$. Usando la ecuación obtenida ${G_0}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) = {G_0}^{f}\left(\omega, \bar{z}(\vec{z})\middle|\overrightarrow{z'}\right)$ junto con (ver figura):

$$\boldsymbol{\left\langle \overrightarrow{n_z}, \vec{n} \right\rangle + \left\langle \overrightarrow{n_{\bar{z}}}, \vec{n} \right\rangle = 0}$$

encontramos que

$$\boldsymbol{\frac{\partial}{\partial \vec{n}} G_0\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) = \frac{\partial}{\partial \vec{n}} {G_0}^{f}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) + \frac{\partial}{\partial \vec{n}} {G_0}^{f}\left(\omega, \bar{z}(\vec{z})\middle|\overrightarrow{z'}\right) = 0, \qquad \overrightarrow{z'} \in V}$$

Como conclusión, la funcion de Green $\mathbf{G_0}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right)$ induce un campo sonoro no solo dentro de V pero también en su exterior. Por otro lado, la derivada normal de la funcion de Green es cero para todos los puntos del borde $\overrightarrow{z'} \in \partial$V. Por ello, el insertar esta funcion de Green $\mathbf{G_0}\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right)$ en la integral de Kirchhoff-Helmholtz lleva al resultado:

$$\boldsymbol{P_0(\omega, \vec{z}) = -\iint_{\partial V} G_0\left(\omega, \vec{z}\middle|\overrightarrow{z'}\right) \cdot \frac{\partial}{\partial \vec{n}} P_0\left(\omega, \overrightarrow{z'}\right) d\overrightarrow{z'}, \qquad \vec{z} \in V}$$

Este resultado se conoce como *integral de Rayleigh de tipo I*. Esta ecuación produce el campo deseado dentro de V. Fuera del volumen, el campo de onda consiste en una imagen simétrica del campo dentro del volumen.

La integral de Rayleigh de tipo I dicta que el campo sonoro dentro de un volumen puede ser reproducido mediante una distribución de fuentes puntuales si se tolera una versión simétrica de este campo fuera del volumen.

Para tener en cuenta solo tipos de fuente monopolo para esta funcion de Green, podemos expresar la ecuación $G_1(\omega, \vec{x}|\vec{x'}) = \int_{-\infty}^{+\infty} G_0(\omega, \vec{x}|\vec{z'}) \cdot dz'$ usando la condición $G_0{}^f\left(\omega, \vec{z}\middle|\vec{z'}\right) = G_0{}^f\left(\omega, \bar{z}(\vec{z})\middle|\vec{z'}\right)$, con lo que:

$$G_0\left(\omega, \vec{z}\middle|\vec{z'}\right) = 2 \cdot G_0{}^f\left(\omega, \vec{z}\middle|\vec{z'}\right)$$

i.e. representar las fuentes en el borde como fuentes puntuales en campo libre pero con doble intensidad. El factor 2 proviene de que estas fuentes puntuales describen la contribución dentro y fuera de V de igual manera. Así, la integral de Rayleigh tipo I se puede escribir como

$$\mathbf{P_0}(\boldsymbol{\omega}, \vec{\mathbf{z}}) = -\iint_{\partial \mathbf{V}} \mathbf{2} \cdot \mathbf{G_0}^{\mathbf{f}}\left(\boldsymbol{\omega}, \vec{\mathbf{z}}\middle|\overrightarrow{\mathbf{z'}}\right) \cdot \frac{\boldsymbol{\partial}}{\boldsymbol{\partial \vec{n}}} \mathbf{P_0}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{z'}}\right) \mathbf{d}\overrightarrow{\mathbf{z'}}, \qquad \vec{\mathbf{z}} \in \mathbf{V}$$

2.8.9. Reduccion a dos dimensiones espaciales

El volumen V tiene que ser bastante grande para englobar tan solo a una pequeña audiencia o para dar la posibilidad a un oyente de moverse por una sala con el campo sonoro. Cubrir toda la superficie ∂V con fuentes sonoras apropiadas parece un reto tanto tecnológico como económico. Mas aún, puede que no fuera necesario reproducir el campo sonoro en todo el volumen (por ejemplo en el suelo y en el techo, fuera del alcance de los oyentes); una reproducción cuidadosa a la altura de los oídos del oyente en un plano horizontal puede ser suficiente. Tal simplificación requiere reducir el problema de 3D a 2D.

Para reducir la distribución de fuentes de una a lo ancho de la superficie a una que sea una curva cerrada en un plano horizontal (preferiblemente a la altura de los oídos de los oyentes), se procede como sigue. Por facilitar el desarrollo se supone que esta altura es $z = 0$.

Las herramientas matemáticas para la reducción de una distribución de fuentes iban implícitas en el desarrollo presentado para la integral de Kirchhoff-Helmholtz para un prisma y las relaciones entre fuentes lineales y puntuales dada por:

$$\widetilde{G_1}^f(\omega, \rho) = H(\omega)\, A(\rho)\, \widetilde{G_0}^f(\omega, \rho)$$

Estas dos consideraciónes se aplican en dos pasos a la representación de un campo sonoro en la ecuación deducida del apartado anterior:

$$P_0(\omega, \vec{z}) = -\iint_{\partial V} 2G_0{}^f(\omega, \vec{z}|\vec{z'}) \cdot \frac{\partial}{\partial \vec{n}} P_0(\omega, \vec{z'}) \cdot d\vec{z'}, \qquad \vec{z} \in V$$

El primer paso es convertir la superficie ∂V en un prisma. Desarrollando se obtiene []:

$$P_1(\omega, \vec{x}) = -\oint_{\partial L} 2G_1{}^f(\omega, \vec{x}|\vec{x'}) \cdot \frac{\partial}{\partial \vec{n}} P_1(\omega, \vec{x'}) \cdot d\vec{x'}, \qquad \vec{x} \in L$$

En un segundo paso se sustituyen las contribuciones de las fuentes lineales por fuentes puntuales, obteniéndose:

$$P_1(\omega,\vec{x}) = -\oint_{\partial L} 2G_0{}^{f}\left(\omega,\vec{x}\middle|\overrightarrow{x'}\right)\cdot D\left(\omega,\vec{x}\middle|\overrightarrow{x'}\right)\cdot d\overrightarrow{x'}, \qquad \vec{x}\in L$$

con

$$D\left(\omega,\vec{x}\middle|\overrightarrow{x'}\right) = 2\cdot A\left(\left|\vec{x}-\overrightarrow{x'}\right|\right)\cdot H(\omega)\cdot\frac{\partial}{\partial\vec{n}}P_1\left(\omega,\overrightarrow{x'}\right)$$

y la variable ρ desarrollada para reflejar la dependencia con $\vec{\mathbf{x}}$ y $\overrightarrow{\mathbf{x'}}$. En la ecuación para $P_1(\omega,\vec{x})$, el termino $G_0{}^{f}\left(\omega,\vec{x}\middle|\overrightarrow{x'}\right)$ denota la funcion de Green de fuentes tipo monopolo en el contorno ∂L en el plano x-y. Describe la propagación de las ondas en el espacio, no obstante, las posiciones de los oyentes se asume residen también en el plano x-y. $D\left(\omega,\vec{x}\middle|\overrightarrow{x'}\right)$ denota la señal fuente para los monopolos.

2.8.10. Muestreo espacial

La integral de Kirchhoff-Helmholtz es valida para una distribución continua de fuentes sobre el contorno ∂V. Una aproximación de las fuentes sonoras usando altavoces resulta en una distribución espacial discreta de fuentes. Los efectos de discretizacion resultantes deben ser descritos en términos de muestreo espacial:

$$P_1(\omega,\vec{x}) \cong -\sum_n G_0{}^{f}\left(\omega,\vec{x}\middle|\overrightarrow{x'}_n\right)\cdot D\left(\omega,\vec{x}\middle|\overrightarrow{x'}_n\right)\cdot \Delta x'_n$$

siendo $\Delta x'_n$ la longitud del incremento espacial $\Delta\overrightarrow{x'}_n$ entre muestras. No es necesario que sea equidistante.

Desafortunadamente, el efecto del aliasing espacial para campos de ondas no es un tema muy comprendido por ahora. Un resumen de unos pocos pocos aspectos técnicos habrán de bastarle al lector.

El muestreo de funciones multidimensionales sí que es bien conocido en el procesado de imágenes y video. No obstante, derivar un espaciamento apropiado a partir del requerimiento del teorema de muestreo de Nyquist requiere posicionar dos altavoces por la longitud de onda minima permisible. Para el rango de audio usual hasta los 20 kHz, esto requeriría posiciones altavoces a distancias menores de 1 cm. Tal array de altavoces no es técnicamente realizable, considerando el tamaño disponible de los altavoces actuales.

Parece que hay dos factores de influencia que permiten reducir el numero de altavoces significativamente.
Primero, el campo de onda no es una señal arbitraria restringida solo por un limite superior en su rango de frecuencia. En vez de eso, todas las señales en el ámbito de la Acustica son soluciones de la ecuación de onda. Este propiedad particular restringe las posibilidades de representación para campos de onda en el dominio de la frecuencia [39].

El segundo factor es la percepción humana de efectos de aliasing espacial. Los experimentos llevados a cabo hasta ahora demuestran que estos efectos de aliasing pueden ser enmascarados efectivamente por otras componentes del sonido. No obstante, poco se conoce hasta ahora sobre el aliasing espacial en la percepción auditiva humana. Para mas información sobre aliasing espacial se recomienda revisar [40].

Para concluir, decir que el muestreo espacial parece ser una útil herramienta para la reproducción de audio 3D aunque los efectos sobre el sistema de percepción humano todavía no han sido explorados.

2.8.11. Señales de alimentación

Una vez que la distribución de fuentes ha sido aproximada por una malla de altavoces lo suficientemente densa, las señales de alimentación para los altavoces han de generarse mediante hardware DSP y conversores AD/DA. Estas señales se derivan de un análisis que tenga en cuenta la naturaleza del campo de onda deseado a la ecuación

$$D\left(\omega, \vec{x}\middle|\overrightarrow{x'}\right) = 2 \cdot A\left(\left|\vec{x} - \overrightarrow{x'}\right|\right) \cdot H(\omega) \cdot \frac{\partial}{\partial \vec{n}} P_1\left(\omega, \overrightarrow{x'}\right)$$

Los campos de onda se pueden modear mediante configuraciones de fuentes de distinto tipo, p.ej. monopolos y dipolos, y mediante ondas planas.
La obtención de las señales de alimentación mediante un modelo del campo de onda se llama *renderizado basado en el modelo.* Si por el contrario se graba un campo de onda en un entorno real como una sala de conciertos o una iglesia, la obtención de las señales de alimentación por medio de un campo de onda sonoro ya grabado se llama *renderizado basado en datos.*

Aquí se analiza el caso del renderizado basado en un modelo bastante general, donde el campo de onda deseada esta dado por una descomposición en ondas planas de acuerdo a la ecuación del apartado 2.10.1.

$$P(\omega, \vec{\mathbf{r}}) = \int_0^{2\pi} F(\omega, \zeta) \cdot e^{j \cdot \frac{\omega}{c} \cdot r \cdot \cos(\zeta - \theta)} \cdot d\zeta$$

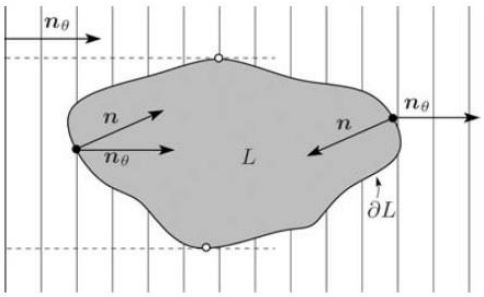
Hay tres puntos a tener en cuenta.

Condiciones de contorno

El eliminar las fuentes dipolares (ver subsección "Fuentes monopolo y dipolo") se basó en la elección de una funcion de Green con condiciones de contorno homogéneas de segunda especie (tipo Neumann) i.e. una derivada normal que se desvanece en el borde:

$$\frac{\partial}{\partial \vec{\mathbf{n}}} \mathbf{G_0}\left(\boldsymbol{\omega}, \vec{\mathbf{z}}\middle|\overrightarrow{\mathbf{z'}}\right) = \frac{\partial}{\partial \vec{\mathbf{n}}} \mathbf{G_0^f}\left(\boldsymbol{\omega}, \vec{\mathbf{z}}\middle|\overrightarrow{\mathbf{z'}}\right) + \frac{\partial}{\partial \vec{\mathbf{n}}} \mathbf{G_0^f}\left(\boldsymbol{\omega}, \bar{\mathbf{z}}(\vec{\mathbf{z}})\middle|\overrightarrow{\mathbf{z'}}\right) = \mathbf{0}, \qquad \overrightarrow{\mathbf{z'}} \in \mathbf{V}$$

Las condiciones de contorno de este tipo se sabe producen reglexiones en los bordes. A primera vista no esta claro como van a ocurrir estas reflexiones ya que el borde ∂L es solamente un contorno arbitrario para el posicionamiento de fuentes puntuales y no es una pared sólida real.
Para entender mejor esta situación, considere la siguiente figura.



Se muestra un contorno ∂L en un plano con dos posiciones diferentes arbitrarias de dos monopolos. Estas posiciones son muestras espaciales $\overrightarrow{x'}_n$ de la coordenada $\overrightarrow{x'}$ en ∂L. Debido a la arbitrariedad de las posiciones y para simplificar la notación, se omitirá el subíndice n. Notese que las posiciones de los

monopolos no se describen solo por sus coordenadas $\overrightarrow{x'}$ sino que también influye el vector normal $\vec{\mathbf{n}}$ en el contorno ∂L en la posición $\overrightarrow{x'}$ y el ángulo:

$$\vec{\mathbf{n}} = \begin{bmatrix} \mathbf{cos}(\boldsymbol{\gamma}) \\ \mathbf{sin}(\boldsymbol{\gamma}) \end{bmatrix}$$

Los monopolos deben alimentarse de manera que produzcan una onda plana en la direccion $\overrightarrow{\mathbf{n}_\zeta}$**.** En la figura, el contorno se ha separado en dos mediante círculos. Los monopolos a la izquierda emanan ondas al dominio L con una componente en la direccion $\overrightarrow{n_\zeta}$ de la onda plana. Ya que producen ondas esféricas, los monopolos también radian una componente en direccion opuesta a $\overrightarrow{n_\zeta}$ (que no afecta al dominio L). No obstante un monopolo en la parte derecha radia hacia L una componente opuesta a $\overrightarrow{n_\zeta}$. La superposición de los monpolos y dipolos requerida por la integral de Kirchhoff-Helmholtz crea una directividad que previene componentes de onda en direcciones falsas. Pero si solo se usan monopolos, no queda mas remedio que aceptar radiación de sonido omnidireccional (igual en todas direcciones).

Por otro lado, las radiaciones de los monopolos de la parte derecha de la figura al dominio L igualan las reflexiones que se hubieran producido por una superficie rigida en el borde ∂L. Aunque dicha pared no este presente realmente, sus reflexiones son producidas por los monopolos de la parte derecha.
No obstante estas componentes que viajan en la direccion opuesta a $\overrightarrow{n_\zeta}$ no pertenecen al campo de onda deseado. Y ya que hemos eliminado los dipolos del conjunto de soluciones del problema, las reflexiones en la parte derecha han de ser contrarrestadas mediante un adecuado tratamiento de las señales de alimentación. Una manera simple pero efectiva de prevenir la radiación de sonido en direcciones falsas es cancelar las señales de alimentación de los altavoces (monopolos) de la parte derecha. Esto se hace mediante una funcion ventana rectangular que determina la actividad de la fuente para cada posición $\overrightarrow{x'}$ en el contorno y para cada posible direccion ζ de una onda plana. Los valores de la ventana dependen del producto escalar $\langle \vec{\mathbf{n}}, \overrightarrow{\mathbf{n}_\zeta} \rangle$ entre la direccion de propagación de la onda plana $\overrightarrow{n_\zeta}$ y el vector normal en cada posición $\overrightarrow{x'}$ de ∂L. Ya que

$$\langle \vec{\mathbf{n}}, \overrightarrow{\mathbf{n}_\zeta} \rangle = \cos(\zeta - \gamma)$$

la funcion ventana se define:

$$\mathbf{v}(\overrightarrow{\mathbf{x}'}, \boldsymbol{\zeta}) = \begin{cases} \mathbf{1}, & \mathbf{si}\ \langle \vec{\mathbf{n}}, \overrightarrow{\mathbf{n}_\zeta} \rangle > \mathbf{0}\ \ \text{ó}\ \ |\boldsymbol{\zeta} - \boldsymbol{\gamma}| < \dfrac{\boldsymbol{\pi}}{\mathbf{2}} \\ \mathbf{0}, & \mathbf{resto} \end{cases}$$

Con esta definición, la señal de alimentación para los monopolos de la ecuación $\mathrm{D}(\omega, \vec{x}|\overrightarrow{x'})$ se puede reescribir como:

$$\mathbf{D}(\boldsymbol{\omega}, \vec{\mathbf{x}}|\overrightarrow{\mathbf{x}'}) = \mathbf{2}\ \mathbf{v}(\overrightarrow{\mathbf{x}'}, \boldsymbol{\zeta})\ \mathbf{A}(|\vec{\mathbf{x}} - \overrightarrow{\mathbf{x}'}|)\ \mathbf{H}(\boldsymbol{\omega})\ \frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\vec{\mathbf{n}}}\mathbf{P_1}(\boldsymbol{\omega}, \overrightarrow{\mathbf{x}'})$$

La funcion ventana elimina así las reflexiones introducidas por las condiciones de borde tipo Neumannn. El campo de onda se reproduce fielmente dentro de la distribución de fuentes secundarias monopolo. El campo fuera de esa región no es cero (que seria el caso si se hubieran usado ambas fuentes monopolo y dipolo) pero no es relevante.

Calculo de la derivada normal

El calculo de la derivada normal en la ecuación para $\mathrm{D}(\omega, \vec{x}|\overrightarrow{x'})$ requiere expresar la derivada normal $\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\vec{\mathbf{n}}}\mathbf{P_1}(\boldsymbol{\omega}, \overrightarrow{\mathbf{x}'})$ mediante una caracterización adecuada del campo de onda. Se consideran campos de onda con una representación en forma de expansión de ondas planas. Así, la derivada normal puede ser expresada por los coeficientes de las ondas planas i.e. por las formas de onda de las compononentes de onda plana individuales.

Considere primero solo una única onda plana. Como se ha visto, está descrita en el dominio frecuencial mediante la ecuación

$$\mathbf{P_1}\left(\omega, \overrightarrow{\mathbf{x'}}\right) = \mathrm{F}(\omega, \zeta) \cdot \mathrm{e}^{\mathrm{j}\cdot\frac{\omega}{\mathrm{c}}\cdot\langle \overrightarrow{\mathbf{x'}}, \overrightarrow{\mathbf{n_\zeta}}\rangle}$$

El gradiente se calcula directamente:

$$\boldsymbol{\nabla \mathbf{P_1}}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{x'}}\right) = \mathbf{j} \cdot \frac{\boldsymbol{\omega}}{\mathbf{c}} \cdot \mathbf{P_1}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{x'}}\right) \overrightarrow{\mathbf{n_\zeta}}$$

Y la derivada normal es, según se ha visto:

$$\frac{\partial}{\partial \vec{\mathrm{n}}} \mathrm{P}_1(\omega, \overrightarrow{\mathrm{x'}}) = \left\langle \nabla \mathrm{P}_1\left(\omega, \overrightarrow{\mathrm{x'}}\right), \vec{\mathrm{n}} \right\rangle = \mathrm{j} \cdot \frac{\omega}{\mathrm{c}} \cdot \mathrm{P}_1\left(\omega, \overrightarrow{\mathrm{x'}}\right) \cdot \langle \overrightarrow{\mathrm{n_\zeta}}, \vec{\mathrm{n}} \rangle = \mathrm{j} \cdot \frac{\omega}{\mathrm{c}} \cdot \mathrm{P}_1\left(\omega, \overrightarrow{\mathrm{x'}}\right) \cdot \cos(\zeta - \gamma)$$

Al insertar esta ecuación en la formula para $\mathbf{D}\left(\boldsymbol{\omega}, \vec{\mathbf{x}} \middle| \overrightarrow{\mathbf{x'}}\right)$ y después de algunas manipulaciones algebraicas se llega al siguiente resultado:

$$\mathbf{D_\zeta}\left(\boldsymbol{\omega}, \vec{\mathbf{x}} \middle| \overrightarrow{\mathbf{x'}}\right) = \mathbf{2}\, \mathbf{w}\left(\overrightarrow{\mathbf{x'}}, \boldsymbol{\zeta}\right) \mathbf{A}\left(\left|\vec{\mathbf{x}} - \overrightarrow{\mathbf{x'}}\right|\right) \mathbf{K}(\boldsymbol{\omega}) \cdot \mathbf{e}^{\mathbf{j}\cdot\frac{\boldsymbol{\omega}}{\mathbf{c}}\cdot\langle \overrightarrow{\mathbf{x'}}, \overrightarrow{\mathbf{n_\zeta}}\rangle} \mathbf{F}(\boldsymbol{\omega}, \boldsymbol{\zeta})$$

donde:

· La funcion w() combina los efectos de la ventana rectangular y el termino coseno de la derivada normal

· el espectro F() es la transformada de Fourier de la forma de onda f()

· el termino exponencial describe el retardo de la onda plana del origen a la posición x' del monopolo. Se puede implementar mediante un retardo temporal de la forma de onda f().

· La respuesta en frecuencia K() combina el termino H() de la aproximación de una fuente lineal mediante una fuente puntual y el efecto de la diferenciación de la derivada normal:
Se puede implementar filtrando la forma de onda f()

· La modificación de amplitud viene de la aproximación de una fuente lineal mediante una fuente puntual.

Finalmente, las señales de alimentación para un campo de onda que este compuesto por ondas planas puede obtenerse mediante la superposición de las señales de alimentación para las ondas planas individuales:

$$\mathbf{D}\left(\boldsymbol{\omega}, \vec{\mathbf{x}} \middle| \overrightarrow{\mathbf{x'}}\right) = \int_{\mathbf{0}}^{\mathbf{2\pi}} \mathbf{D_\zeta}\left(\boldsymbol{\omega}, \vec{\mathbf{x}} \middle| \overrightarrow{\mathbf{x'}}\right) \mathbf{d\zeta}$$

2.8.12. Sistema para el tratamiento de señales de audio

Una vez que se han determinado las señales de alimentación para los altavoces, se puede investigar la estructura para el procesado de señal. En esta sección solo se usan señales de alimentación independientes del oyente y denotadas por $D_{0,\zeta}\left(\omega, \overrightarrow{\mathrm{x'}}\right)$. Son la salida de un sistema con formas de onda de entrada $\mathrm{F}(\omega, \zeta)$. Esta cadena de procesamiento se puede describir mediante:

$$D_{0,\zeta}\left(\omega, \overrightarrow{\mathrm{x'}}\right) = \mathrm{M}\left(\omega, \overrightarrow{\mathrm{x'}}, \zeta\right) \cdot \mathrm{F}(\omega, \zeta)$$

con

$$\mathrm{M}\left(\omega, \overrightarrow{\mathrm{x'}}, \zeta\right) = 2 \cdot \mathrm{w}\left(\overrightarrow{\mathrm{x'}}, \zeta\right) \cdot \mathrm{A}\left(\left|\overrightarrow{x_0} - \overrightarrow{\mathrm{x'}}\right|\right) \cdot \mathrm{K}(\omega) \cdot \mathrm{e}^{\mathrm{j}\frac{\omega}{\mathrm{c}}\langle\, \overrightarrow{\mathrm{x'}}, \overrightarrow{\mathrm{n_\zeta}}\,\rangle}$$

Se pueden construir campos de onda complejos mediante una superposición de ondas planas según

$$\mathbf{D}\left(\boldsymbol{\omega}, \vec{\mathbf{x}} \middle| \overrightarrow{\mathbf{x'}}\right) = \int_{\mathbf{0}}^{\mathbf{2\pi}} \mathbf{D_\zeta}\left(\boldsymbol{\omega}, \vec{\mathbf{x}} \middle| \overrightarrow{\mathbf{x'}}\right) \mathbf{d\zeta}$$

No obstante, las componentes suelen ser limitadas, por lo que las señales de alimentación que contengan contribuciones de un numero finito de ondas planas de un conjunto de angulos discreto $\{\zeta_m\}$ es

$$\mathbf{D_0}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{x}'}\right) = \sum_{\mathbf{m}} \mathbf{D}_{\mathbf{0},\,\boldsymbol{\zeta}_\mathbf{m}}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{x}'}\right) = \sum_{\mathbf{m}} \mathbf{M}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{x}'}, \boldsymbol{\zeta}_\mathbf{m}\right) \cdot \mathbf{F}(\boldsymbol{\omega}, \boldsymbol{\zeta}_\mathbf{m})$$

A partir de este numero finito de componentes de las ondas planas, se tienen que generar las señales de alimentación para cada posición discreta $\overrightarrow{\mathbf{x}_\mathbf{n}'}$. la estructura resultante se maneja mejor usando notación vectorial:

$$\overrightarrow{\mathbf{D_0}}(\boldsymbol{\omega}) = \begin{bmatrix} \vdots \\ \mathbf{D_0}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{x}_\mathbf{n}'}\right) \\ \vdots \end{bmatrix}, \qquad \vec{\mathbf{F}}(\boldsymbol{\omega}) = \begin{bmatrix} \vdots \\ \mathbf{F}(\boldsymbol{\omega}, \boldsymbol{\zeta}_\mathbf{m}) \\ \vdots \end{bmatrix},$$

$$\vec{\mathbf{M}}(\boldsymbol{\omega}) = \begin{bmatrix} \cdots & \mathbf{M}\left(\boldsymbol{\omega}, \overrightarrow{\mathbf{x}_\mathbf{n}'}, \boldsymbol{\zeta}_\mathbf{m}\right) & \cdots \end{bmatrix}$$

Así descritas, las señales de alimentación para cada altavoz serán calculadas a partir de las formas de onda de las componentes de cada onda plana mediante

$$\overrightarrow{\mathbf{D_0}}(\boldsymbol{\omega}) = \vec{\mathbf{M}}(\boldsymbol{\omega})\, \vec{\mathbf{F}}(\boldsymbol{\omega})$$

En el dominio temporal, las señales de alimentación son el resultado de una **convolución multicanal**

$$\overrightarrow{\mathbf{d_0}}(\mathbf{t}) = \vec{\mathbf{m}}(\mathbf{t}) * \vec{\mathbf{f}}(\mathbf{t})$$

Resumeniendo, la estructura para el procesamiento de señal es un *sistema de entradas multiples y salidas multiples (multiple-input, multiple-output MIMO)* que realiza una convolución multicanal con las formas de onda de las componentes de las ondas planas.

La síntesis de campos de onda también se ha usado para extrapolar HRTFs, ver ["Efficient range extrapolation of head-related impulse responses by wavefield synthesis techniques", Ahrens y Spors, iEEE, 2011]

2.8.13. Implementación de un sistema WFS

La universidad TU Berlin en Berlín, Alemania, entre otras universidades del mundo, dispone de un sistema WFS instalado en uno de sus auditorios más grandes. El sistema se puede ver en la siguiente figura, donde los bloques grises en las paredes son altavoces diseñados para el renderizado del campo sonoro 3D. Alrededor de toda la sala, con capacidad para unas 640 personas, se halla una banda de 2700 altavoces separados 10 cm, situados a la altura de las cabezas de los oyentes. Los altavoces son controlados por un cluster de computadores con capacidad para 832 canales de audio.

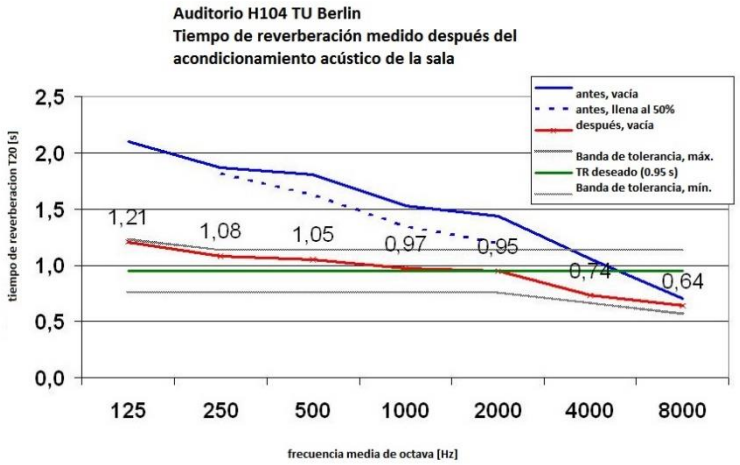
En relación al diseño acústico de la sala, los requisitos para una sala para reproducción de WFS tienen que evitar interferencias acústicas (i.e. eliminar ecos de flutter o de reflexiones de las paredes de atrás), tener el tiempo de reverberación adecuado y controlar las reflexiones. La capacidad de sonorización del espacio no debería interferir con el campo sonoro sintetizado. Para esta sala se quiso que la reverberación y las

reflexiones no afectaran negativamente a la síntesis de campos sonoros. Se concretaron dos objetivos para su remodelación arquitectónica:

1) un tiempo de reverberación deseado de 0.95s (en un estado de ocupación de la sala del 50%) con una respuesta en frecuencia estable.

2) Prevenir reflexiones tempranas de la sala, ya que afectan negativamente al campo generado por WFS. Las reflexiones laterales se previnieron mediante superficies absorbentes y las reflexiones no favorables de suelo y el techo se eliminaron mediante las características direccionales de los altavoces para WFS, que son sobre todo horizontales.

El resultado de arquitectos e ingenieros en la sala se muestra en la siguiente figura.



El autor de este Proyecto Fin de Carrera estuvo estudiando en esta universidad en 2014 y asistió personalmente a una demostración abierta al público del sistema de WFS. Las sensaciones que este sistema de WFS puede transmitir son, cuanto menos, impresionantes. Es una lástima que no se puedan expresar exactamente mediante palabras, pero desde luego se puede asegurar una sensación total de inmersión y una extraordinaria precisión en los campos sonoros que este sistema genera.

Más información sobre el hardware de este sistema de WFS y otros aspectos en:
https://www.ak.tu-berlin.de/menue/research/wave_field_synthesis/parameter/en/

BIBLIOGRAFÍA

[1] "Improvements in and relating to sound transmission, sound recording and sound reproducing systems", Blumlein. 1931, UK Patent.

[2] "basic principles of stereophonic sound", Snow. SMPTE, 1953.

[3] "Discrete-matrix multichannel stereo", Cooper and Shinga. AES, 1972.

[4] "Virtual Sound Source Positioning Using Vector Base Amplitude Panning", Pulkki, V. AES, 1997.

[5] "Localization of Amplitude-Panned Virtual Sources I: Stereophonic Panning", Pulkki, V. and Karjalainen, M. AES, 2001.

[6] "Localization of Amplitude-Panned Virtual Sources II: Two- and Three Dimensional Panning", Pulkki, V. AES, 2001.

[7] "On the function of the two ears in the perception of space", Thompson. 1882.

[8] "On our perception of sound direction", Lord Rayleigh. Philosophical Magazine Series 6 , vol. 13, no. 74, pp. 214-232, 1907.

[9] "3-D Spatialization and Localization, and Simulated Surround Sound with Headphones", O'Neil, L. and Cassidy, B. University of Victoria, Canada, 2006.

[10] "Dolby Surround Pro Logic Decoder. Principles of Operation", Dressler, R. Technical Report, Dolby Laboratories, 2001. www.dolby.com/tech.

[11] "Dolby Surround Pro Logic II Decoder. Principles of Operation", Dressler, R. Technical Report, Dolby Laboratories, 2001. www.dolby.com/tech.

[12] "HRTF Measurements of a KEMAR Dummy-Head Microphone", Gardner, W et al. MIT Media Lab, 1994. http://sound.media.mit.edu/resources/KEMAR.html.

[13] "Distance Dependent Head-related Transfer Function Database of KEMAR", Qu, T. et al. Peking University, 2008.

[14] "Range dependence of the response of a spherical head model", Duda, R. O and Martens, W. L. Journal of Acoustical Society of America, 1998.

[15] "Horizontal plane HRTF reproduction using continuous Fourier-Bessel functions", Zhang, W. et al. AES, 2007.

[16] "Representation of Head Related Transfer Functions with Principal Component Analysis", Sodnik, J. et al. AES, 2004.

[17] "An efficient wavelet-based HRTF model for auralization", Torres et al. EAA,

[18] "HRTF modeling for efficient auralization", Torres, J. C. B. and Petraglia, M. R. IEEE, 2003.

[19] "A new approach to HRTF Audio Spatialization", Marolt, University of Ljubljana.

[20] "HRTF-based Systems", CIPIC Laboratories, UC Davis. http://interface.cipic.ucdavis.edu/sound/tutorial/hrtfsys.html

[21] "Apparent Sound Source Translator", Schroeder, M.R. and Atal, B.S. U.S. Patent 3236949, 1966.

[22] "Stereophonic Earphones and Binaural Loudspeakers", Bauer, B. B.
JAES Volume 9 Number 2 pp. 148-151. AES, 1961.

[23] "3-D Audio using Loudspeakers", Gardner, W. 1998.

[24] "A stereo Crosstalk cancellation system based on the common-acoustical pole/zero model", Wang, L. , Yin, F. and Chen, Z. EURASIP, 2010.

[25] "IRIS. Medición de la respuesta al impulso en 3D", Alava Ingenieros.
http://www.alava-ing.es/ingenieros/productos/acustica-y-vibraciones/software-simulacion-acustica/software-de-medicion-de-respuesta-al-impulso-en-3d/

[26] "Room Acoustics", 5th ed. Kuttruff, H. CRC Press, 2009.

[27] "Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality", Vorländer, M. Springer, 2008.

[28] "A model for Room Acoustics", McGovern, S. 2003.
http://www.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator/content/rir.m

[29] "Fast Convolution", McGovern, S. 2003.
http://www.mathworks.com/matlabcentral/fileexchange/5110-fast-convolution/content/fconv.m

[30] "Efficient Convolution without Input-Output Delay", Gardner, W. AES, 1995.

[31] "DAFx", chap. 6, Zolzer, U. John Wiley & Sons, 2002.

[32] "space: un programma di spazializzazione per il Live Electronics",
Belladonna, A. and Vidolin, A.
Proc. Second Int. Conf. on Acoustics and Music Research, pp 113-118, 1995.

[33] "3-D Sound Spatialization using Ambisonc Techniques",
Malham, D. G, and Myatt, A.
Computer Music Journal, Vol. 19, No. 4, pp. 58-70. MIT Press, 1995.

[34] "3D audio technologies: applications to sound capture, post-production and listener perception", Cengarle, G. Universidad Pompeu Fabra, 2012.

[35] "A Holographic Approach to Acoustic Control", Berkhout, A. J.,
JAES Vol. 36 Issue 12 pp. 977-995. AES, 1988.

[36] "Acoustic control by wave field synthesis", Berkhout, A. J., de Vries, D. and Vogel, P.
Journal of the Acoust. Soc. Of America 93, 2764, 1993.

[37] "Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography",
Williams, E. G. Academic Press, 1999.

[38] "Topics in Acoustic Echo and Noise Control. Selected Methods for the Cancellation of Acoustic Echoes, the Reduction fo Background Noise, and Speech Processing", chap. 13, Hänsler, E. and Schmidt, G. Springer, 2006.

[39] "Representation of two-dimensional wave fields by multidimensional signals",
Rabenstein, R, Steffen, P. and Spors, S.
ELSEVIER Signal Processing Vol. 86, Issue 6, pp. 1341–1351, 2006.

[40] "Audio Signal Processing for Next-Generation multimedia communication systems",
cap. "Sound Field Synthesis", Spors, S. et al. Kluwer, 2004.

[41] ”Wave Field Synthesis at TU Berlin”, Berlin, Alemania.
https://www.ak.tu-berlin.de/menue/research/wave_field_synthesis/parameter/en/

[42] Altiverb 7 convolution reverberation plug-in by Audio Ease
http://www.audioease.com/Pages/Altiverb/

[43] “Auralization using headphones”, Eide, I. N.
Norwegian University of Science and Technology, Norway, 2012.

[44] "On the acoustic shadow of a sphere", Strutt, J.W.(Lord Rayleigh). Phil. Trans. R. Soc. London, vol. 203A, pp. 87-97 1904; The theory of sound, 2nd ed. New York: Dower, 1945

[45] “A Structural Model for Binaural Sound Synthesis”, Brown, C. P and Duda, R. O.
IEEE Transactions on Speech and Audio Proc., Vol. 6, No. 5, 1998.

[46] “Psychoacoustical aspects of synthesized vertical locale cues”, Watkins, A. J.
J. Acoust. Soc. America, Vol. 63, pp. 1152-1165, 1978.

[47] “Sound localization by human listeners”, Middlebrooks, J. C. and Green, D. M.
Ann. Rev. Psych. Vol. 42, pp. 135-159, 1991.

[48] “Spatial Audio Reproduction with the SoundScape Renderer (SSR)”, Geier, M. and Spors, S.
27th Tonmeistertagung – VDT International Convention, 2012.
http://spatialaudio.net/ssr/

[49] “Pro Tools ©”, Avid Technology, Inc.
http://www.avid.com/us/products/family/pro-tools, 1989-2015.

3. APLICACIÓN DE TÉCNICAS DE AUDIO 3D EN MATLAB. DESARROLLO DE LA APLICACIÓN

3.1. Introducción. Motivación y desarrollo

La idea de este Proyecto Fin de Carrera nace de una charla con el director del departamento de tecnología de la Escuela Politécnica Superior en la Universidad Autónoma de Madrid, España. La creciente expectación por las tecnologías futuras y los numerosos avances desarrollados hasta ahora conllevan cierta intriga sobre cómo se implementan este tipo de dispositivos para la reproducción y mejora de material de audio en general.

Con el fin de perseguir un sistema de audio 3D diseñado por y para estudiantes de ciencias pero también con cierta facilidad de uso para estudiantes de artes escénicas, se creó este software. La aplicación quiere permitir posicionar una fuente de sonido en el espacio 3D y poder moverla por el espacio y que suene en los oídos del oyente como si éste se sentara en esa misma sala con esa misma fuente moviéndose con la trayectoria que se ha establecido. El software se ha diseñado en módulos y de manera ampliable. El autor espera poder ver una mejora de este software, p.ej. añadiendo más fuentes, haciendo la GUI más interactiva y amigable o añadiendo RIRs para poder situar al oyente en distintos entornos virtuales.

La aplicación recibe:

· dimensiones de la sala y coeficientes de absorción de paredes, suelo y techo
· posición fija del oyente (receptor RX) en la sala
· posición dinámica o trayectoria 3D en forma paramétrica de la fuente de sonido (transmisor TX):

$$(t) = (x(t), y(t), z(t))$$

· número de puntos de la trayectoria = nº bloques para el procesado por tramas
· audio (monofónico o estereofónico) a auralizar

La aplicación genera:

· audio binaural o audio 3D para auriculares
· audio binaural o audio 3D para auriculares "stereo-widened"
· audio transaural o audio 3D para altavoces
· gráficas de las señales más relevantes

La aplicación usa:

· base de datos de filtros relacionados con la cabeza (HRIRs)
· respuesta al impulso de la sala (RIR) implementada mediante la función "rir.m"
· *convolución por bloques mediante el método de Solapamiento y Suma (Overlap-and-Add)*. Cada bloque de audio original se convoluciona mediante FFT con la respuesta al impulso binaural de la sala (BRIR), generada mediante la convolución de la HRIR con la RIR. Estas son recalculadas para cada bloque (punto en la trayectoria) en base a la posición entre fuente (TX) y oyente (RX) usando la distancia y los ángulos azimutal y de elevación
· cancelador de Crosstalk para transformar el audio binaural generado a audio transaural
· Widening estéreo para mejorar en cierta medida el audio binaural para auriculares

3.2. Pasos básicos del programa "auralization.m"

3.2.1. Pre-procesado y definiciones

a) Definición de la sala, posición del oyente (RX) y síntesis de características acústicas de la sala mediante "room_parameters0.m" y "room_parameters.m"
b) Definición del tamaño de bloque para el procesado por tramas. Definición de los J puntos de la trayectoria de la fuente sonora (TX): $\sigma(n) = \{\sigma(ni) = (x(ni), y(ni), z(ni))\}$, $ni = 1, 2, \ldots, J$ mediante "traj_computing.m".
c) Definición del audio o tono s(n) a emitir por la fuente sonora mediante "load_audioinput.m"

3.2.2. Procesado por bloques: convolución variante en el tiempo

a) Obtención de las J tramas de entrada xi(n) de la señal de audio continua s(t)
b) Obtención de los J ángulos de elevación y azimutal (θ, φ) mediante "get_angles.m"
c) Obtención de las 2· J HRTFs para L y R, extraídas de la base de datos en función de los ángulos anteriores mediante "get_HRTF.m"
d) Obtención de las 2· J HRTFs para L y R dependientes de la distancia, extraídas de la base de datos en función de los ángulos anteriores mediante "get_HRTF_d.m"
e) Obtención de las J respuestas al impulso de la sala (RIR) en función de las características de la sala y de cada punto σ(ti) de la trayectoria mediante "rir.m"
f) Obtención de las 2· J respuestas binaurales (BRIR) L y R mediante convolución(*) con las anteriores:

$$\mathbf{BRIR_x = RIR * HRIR_x}, \qquad \mathbf{x = L, R}$$

g) Obtención de las 2· J tramas de salida yi_L e yi_R mediante convolución de cada trama xi(n) con las BRIRs L y R:

$$\mathbf{y_{i_x} = x_i * BRIR_x}, \qquad \mathbf{x = L, R}, \qquad \mathbf{i = 1, 2, \ldots, J}$$

h) *Solapamiento y suma* de las J tramas yi_L y las J tramas yi_R para obtener las señales binaurales yL(n) e yR(n) (señales de entrada a a los oídos izquierdo y derecho) finales.

3.2.3. Post-procesado y gráficas

a) Grabación de la señal de audio 3D (audio binaural) a un archivo de audio con extensión .wav
b) Widening estéreo. Post-procesado de las señales binaurales yL(n) e yR(n) para obtener las señales yLp(n) e yRp(n)
c) Cancelador de Crosstalk. Obtención del audio transaural yLc(n) e yRc(n)
d) GRÁFICOS de señales y características binaurales
e) Reproducción de las señales de audio 3D
f) EQ de auriculares (para auriculares HDJ1500, optativo) para la reproducción

(*): Se contemplan 3 formas de convolucionar: fconv(), filter() con historia y fftfilt().

3.3. Audio binaural usando auriculares

Si toda la información evaluada durante el proceso de escucha binaural fuera inherente a las señales de entrada a los tímpanos, se podría simular el efecto de 3D grabando y reproduciendo estas señales en los oídos. Esto fue demostrado por [1] y de hecho se pueden usar bien auriculares o altavoces para la reproducción de grabaciones binaurales.

3.3.1. Señales binaurales

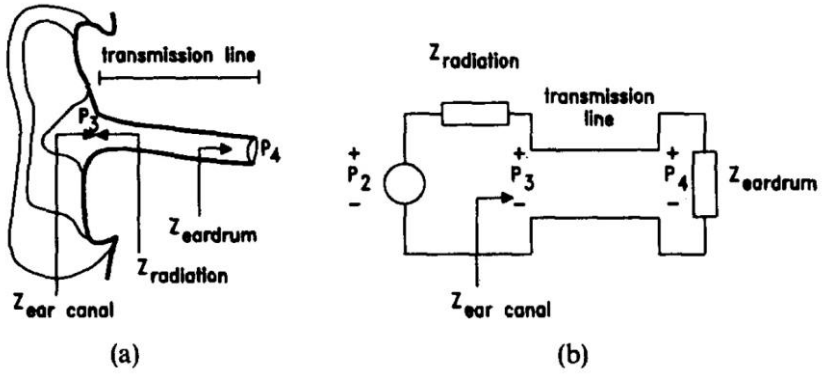
La transmisión del sonido en campo libre de un punto del espacio a los oídos se describe usando las funciones de transferencia HRTFs (o su equivalente temporal HRIR). Estas funciones deberían contener toda la información espacial inherente a la fuente. Consecuentemente, las HRTFs dependen de (r, θ, φ) así como de la complexión del propio oyente [11].
Se ha demostrado que la transmisión de sonido desde un punto situado solo unos milímetros lejos del tímpano es independiente de la dirección. Por ello, la función de transferencia desde un punto en el espacio a un punto dentro o a la entrada del pabellón auricular (bloqueado o abierto) también se entiende como HRTF.
Siguiendo la nomenclatura de [1] y hablando de presión sonora simplemente como presión:

$$\mathbf{HRTF_{ed}(r,\theta,\varphi,\omega,sujeto)} = \frac{\mathbf{P_4(\omega)}}{\mathbf{P_1(\omega)}} = \frac{\mathbf{presión\ en\ el\ tímpano}}{\mathbf{presion\ en\ el\ centro\ de\ la\ cabeza\ (oyente\ ausente)}}$$

$$\mathbf{HRTF_{oe}(r,\theta,\varphi,\omega,sujeto)} = \frac{\mathbf{P_3(\omega)}}{\mathbf{P_1(\omega)}} = \frac{\mathbf{presion\ a\ la\ entrada\ del\ pabellón\ (abierto)}}{\mathbf{presion\ en\ el\ centro\ de\ la\ cabeza\ (oyente\ ausente)}}$$

$$\mathbf{HRTF_{be}(r,\theta,\varphi,\omega,sujeto)} = \frac{\mathbf{P_2(\omega)}}{\mathbf{P_1(\omega)}} = \frac{\mathbf{presion\ a\ la\ entrada\ del\ pabellón\ (bloqueado)}}{\mathbf{presion\ en\ el\ centro\ de\ la\ cabeza\ (oyente\ ausente)}}$$

donde *ed, oe* y *be* denotan *ear drum, open ear-canal* y *closed ear-canal*, respectivamente. Estas magnitudes asociadas a la posición en el sistema auditivo se aprecian en la siguiente figura.



En el caso de escucha binaural en salas, la transmisión del sonido se describe mediante la función de transferencia binaural de la sala BRTF (o su transformada inversa la respuesta al impulso binaural BRIR). La diferencia con la situación de campo libre es que ahora ya no hay sólo sonido proveniente de la dirección directa, sino también reflexiones indirectas con los bordes de la sala que llegan en muy distintos ángulos. Asumiendo la transmisión acústica como un sistema lineal e invariante en el tiempo (LTI), una de las múltiples maneras de formar una BRTF es:

$$\mathbf{BRTF(\omega)} = \sum_{\mathbf{k=1}}^{\mathbf{N}} \mathbf{e^{-j\omega t_k} \cdot w_k \cdot HRTF_k\ (\omega)}$$

donde wi(·) es un coeficiente de ponderación que depende de: el material de las paredes de la sala, la longitud total de la fuente al oyente, la absorción del aire y la directividad Γ de la fuente.

El factor exponencial viene del retardo introducido por recorrer el camino TX-RX. Luego la BRTF depende de la posición del TX respecto al RX, la posición y orientación del RX en la sala, las HRTFs del RX y, por supuesto, la sala misma.
En este trabajo se ha presentado una versión de BRTF más adecuada a los recursos disponibles:

$$\mathbf{BRIR(n) = RIR * HRIR} = \sum_{\mathbf{k=-\infty}}^{+\infty} \mathbf{RIR(k) \cdot HRIR(n-k)}$$
$$\mathbf{BRTF(\omega) = RTF \cdot HRTF}$$

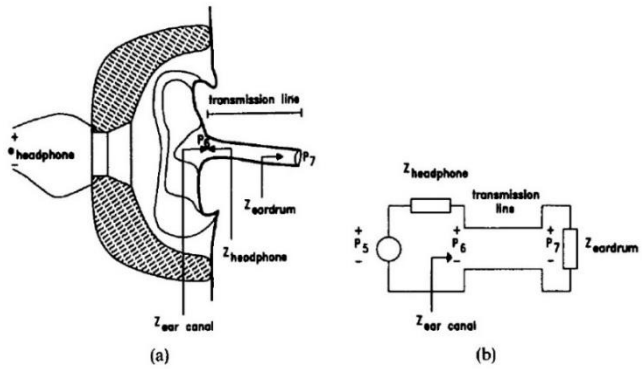
¡Esta convolución es la más problemática pues, aunque la longitud de la HRTF es fija y de valor 512 taps, la longitud de la RIR es en general de 3 segundos, que muestreada a 48kHz da un numero de muestras mayor que 10000! Además hemos de computar un filtro por trama y puede haber muchas tramas (y por tanto muchas BRIR) ya que la señal de audio a auralizar se supone de duración muy larga.
Una implementación mediante convolución directa es, por este motivo, imposible si queremos trabajar en tiempo real.

Una implementación *high-speed* usando procesado por tramas (p.ej. el algoritmo OLA) puede funcionar bien para calcular estas convoluciónes (ver sección 3.2.).
También se podría mejorar el núcleo del programa implementándolo mediante el DSP Toolbox de MatLab®, migrar el programa a mejores ordenadores que usaran GPGPUs o también usar DSPs dedicados, consiguiendo tiempo real verdadero.

La transmisión de sonido desde el auricular al tímpano (o en términos mas general a la entrada al canal auditivo, ya esté este bloqueado o abierto) se describe mediante la *función de transferencia de los auriculares (HpTF):*

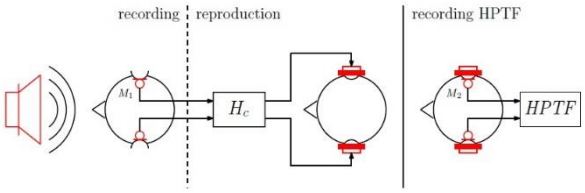
$$\mathbf{HpTF_{ed}(\omega, auricular, sujeto)} = \frac{\mathbf{P_7}}{\mathbf{E_{hp}}}$$
$$\mathbf{HpTF_{oe}(\omega, auricular, sujeto)} = \frac{\mathbf{P_6}}{\mathbf{E_{hp}}}$$
$$\mathbf{HpTF_{be}(\omega, auricular, sujeto)} = \frac{\mathbf{P_5}}{\mathbf{E_{hp}}}$$

donde Ehp es el voltaje en los terminales del auricular. Debido a que las señales binaurales se graban usando micrófonos, las presiones Pi se pueden sustituir por el cociente Emic / MTF del voltahe entre la *función de transferencia del micrófono (MTF*) en caso de disponer de esta información. En la figura siguiente se describen las presiones involucradas.



3.3.2. Un filtro de compensación/ecualización para la síntesis binaural

Los transductores involucrados en grabar y reproducir las señales binaurales (específicamente la tarjeta de audio, los altavoces y los auriculares y los micrófonos) introducen una coloración espectral no deseada así como distorsión de fase.
Esto puede arruinar todo el proceso de codificación del espacio en el sonido. Para compensar esta coloración no deseada, un filtro digital como el de la figura siguiente puede ser empleado [11].



Empezando por los auriculares primero, estos se pueden compensar filtrando las grabaciones con la inversa de la HpTF:

$$\mathbf{H_{c,HpTF,ed}(\omega)} = \frac{\mathbf{1}}{\mathbf{HpTF_{ed}}}$$

$$\mathbf{H_{c,HpTF,oe}(\omega)} = \frac{\mathbf{1}}{\mathbf{HpTF_{oe}}}$$

$$\mathbf{H_{c,HpTF,be}(\omega)} = \frac{\mathbf{1}}{\mathbf{HpTF_{be}}} \cdot \frac{\mathbf{Z_{canal} + Z_{hp}}}{\mathbf{Z_{canal} + Z_{radiation}}}$$

Lo más sencillo es pre-/post-procesar el audio con este filtro: se filtra la señal de audio con el filtro inverso de la respuesta en frecuencia de los auriculares, antes o después del proceso de auralización. Esto puede hacerse de estas dos formas porque como se sabe tratamos con sistemas LTI y la operación de convolución se demuestra es conmutitativa [Señales y Sistemas, Oppenheim]. Hay que notar que este filtro de compensación sólo tiene en cuenta los auriculares; no tiene en cuenta las respuestas en frecuencia de los micrófonos usados. Tampoco se ha incluido la influencia del altavoz usado como fuente sonora en las grabaciones, ya que estos datos no están disponibles en ningún caso. De hecho, la influencia del altavoz en entornos semi-difusos sólo se puede compensar modelando la directividad de la fuente sonora a reproducir. Para medir y modelar directividades de fuentes, se usan micrófonos esféricos y arrays de altavoces.

Para concluir, el efecto de auralización usando *síntesis binaural dinámica* se consigue convolucionando material de audio anecoico con HRTFs (o BRTFs) y escuchándolo mediante auriculares.

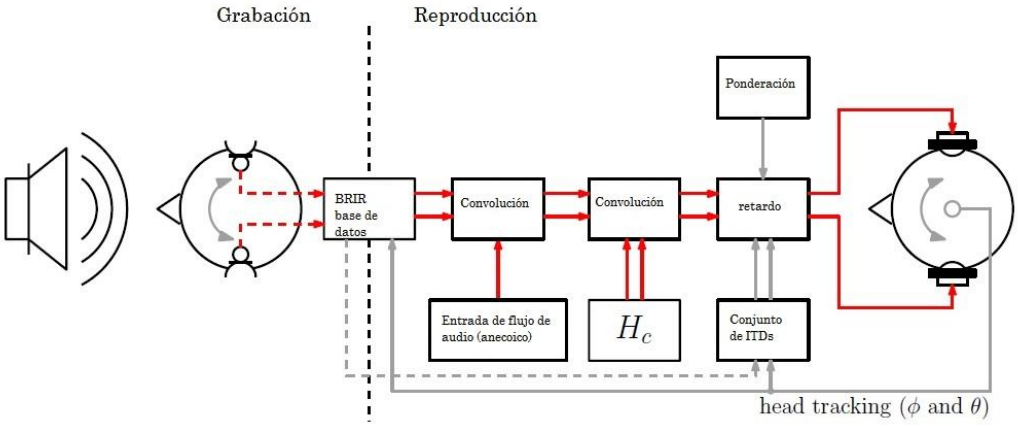
Para la adquisición de HRTFs, se usan las bases de datos y modelos descritos [ver sección 2.5]. La adquisición de las dos BRTFs izquierda y derecha se consigue computando la convolución de las HRTFs izquierda y derecha con una base de datos de RIRs. Las RIR han sido implementadas con una función rir.m que calcula para cada posición de la trayectoria σ(t) = (x(t), y(t), z(t)) una RIR diferente, pero también se pueden usar RIR de otras fuentes, modificando para ello el programa principal para que no calcule para cada punto de la trayectoria una RIR nueva si no que use la misma RIR durante todo el proceso.

Un flujo de audio anecoico es convolucionado con un par de BRIRs por punto de la trayectoria σ(t). La convolución se realiza en el dominio frecuencial mediante el algoritmo de solapamiento y suma (algoritmo OLA), descrito mas adelante.

Antes o después de la convolución, se aplica un filtro de compensación Hc = 1 / HpTF. Por consiguiente, hay que invertir la HpTF, un procedimiento algo laborioso ya que esta inversión es un proceso no lineal. La inversión de la HpTF se puede realizar mediante varios métodos, siendo los preferidos la inversión

LMS y la de fase mínima. Se ha optado como solución contactar con una empresa de calibración de auriculares para obtener la respuesta en frecuencia inversa del dispositivo de reproducción preferido para este proyecto.

Como mejora adicional, se pueden extraer las ITDs de la base de datos de BRIRs antes de la auralización y reinsertarlas de nuevo en el audio usando un retardo variable [11].



Si se usan auriculares, idealmente las señales de entrada a los oídos tienen que ser iguales a las suministradas a los transductores; esto será así siempre que se aplique una calibración o EQ de los auriculares mediante técnicas de filtrado inverso de la respuesta en frecuencia de los auriculares. Si no, la respuesta en frecuencia no plana coloreará el audio, posiblemente destruyendo o atenuando algunos efectos espaciales.
Es por ello que esta calibración ha sido llevada a cabo durante el proyecto para los auriculares Pioneer modelo HDJ-1500. Para ello, se contactó con la empresa Sonarworks® que proporcionó al autor de este proyecto un archivo de texto plano con la respuesta en frecuencia inversa para los canales L y R.

3.3.3. Sistema de coordenadas utilizado

En la escucha mediante auriculares, una señal monofónica se puede posicionar virtualmente como proveniente de cualquier dirección si las HRTFs de ambos oídos son conocidas para la dirección deseada de la fuente virtual. La señal de audio se filtra usando filtros digitales que modelan las HRTFs medidas. Este método simula las señales a la entrada de los canales auditivos (a la altura del timpano) que resultarían de una fuente que existiera y estuviera posicionada en la dirección deseada. Si el oyente se mueve (mueve la cabeza), entonces sus movimientos también deberían ser tenidos en cuenta para el procesamiento. No obstante esto requiere de un sistema de video para head-tracking que no se ha investigado ni implementado en este proyecto debido al trabajo adicional que requeriría implementar dicho sistema adicionalmente a un software de auralizacion.
La falta de un sistema de seguimiento de movimientos tiene la desventaja que el oyente tiene que estar fijo para la escucha con altavoces. Los movimientos destruirán el efecto conseguido. Además, los auriculares producirán efectos no deseados ya que los movimientos del oyente no ayudan a la mejor localización del sonido, como ocurre en la realidad (el ser humano se gira cuando se produce un error en la localización para permitir un reajuste del sistema sensorial y perceptual).

La sala también juega un papel importante. Un entorno anecoico previene reflexiones acústicas que nada tienen que ver con las impresas en el audio y que deterioran a aquel. Las características del audio que pueden verse afectadas por una sala no acondicionada acústicamente son, entre otras: el timbre, la coloración, el brillo y calidez, la reverberación y la disminución en el efecto de cancelación de Crosstalk debido a esta y la sensación envolvente de sentirse presente en el espacio acústico simulado. Ciertamente, la posesión o el uso de una sala de escucha optimizada es ciertamente un lujo al alcance de pocos.

Usando un sistema de coordenadas esférico, la posición de una fuente sonora queda determinada por el ángulo azimutal θ (izquierda-derecha) de 0º a 360º, el de elevación o latitud φ (arriba-abajo) de -90º a +90º y la distancia r entre el receptor RX y la fuente. El origen se ha definido como el centro del eje interaural (que pasa por las entradas a los oídos). Así un punto queda definido por la terna *(r, θ, φ)*. Algunos autores utilizan otras convenciones. φ podría variar también de -180º a +180º y la medida del ángulo azimutal (según se mida el ángulo en sentido reloj o contrarreloj y de 0º a 360º o de -180º a +180º) también puede ser distinta. Por consiguiente, se debe definir explícitamente un sistema de coordenadas y usarlo sin cambios a lo largo de la investigación.

A lo largo de este proyecto, el sistema de coordenadas utilizado se ha definido como:

$$\mathbf{r} = \sqrt{\mathbf{x^2 + y^2 + z^2}}, \qquad \mathbf{0 \le r < \infty}$$

$$\boldsymbol{\theta} = \begin{cases} \mathbf{arctg}\left(\dfrac{\mathbf{y}}{\mathbf{x}}\right), & \mathbf{x > 0}, \quad \mathbf{y > 0}, \quad (\mathbf{1º\ cuadrante}) \\ \mathbf{2\pi + arctg}\left(\dfrac{\mathbf{y}}{\mathbf{x}}\right), & \mathbf{x > 0}, \quad \mathbf{y < 0}, \quad (\mathbf{4º\ cuadrante}) \\ \dfrac{\boldsymbol{\pi}}{\mathbf{2}} \cdot \mathbf{sign(y)}, & \mathbf{x = 0} \\ \boldsymbol{\pi} + \mathbf{arctg}\left(\dfrac{\mathbf{y}}{\mathbf{x}}\right), & \mathbf{x < 0}, \quad (\mathbf{2º\ y\ 3º\ cuadrantes}) \end{cases}$$

$$\boldsymbol{\varphi} = \begin{cases} -\dfrac{\boldsymbol{\pi}}{\mathbf{2}} + \mathbf{arctg}\left(\dfrac{\sqrt{\mathbf{x^2 + y^2}}}{\mathbf{z}}\right), & \mathbf{z > 0} \\ \mathbf{0}, & \mathbf{z = 0} \\ \dfrac{\boldsymbol{\pi}}{\mathbf{2}} + \mathbf{arctg}\left(\dfrac{\sqrt{\mathbf{x^2 + y^2}}}{\mathbf{z}}\right), & \mathbf{z < 0} \end{cases}$$

con:

$$0 \le r \le \infty$$
$$0º \le \theta \le 360º$$
$$-90º \le \varphi \le +90º.$$

y:

$$\mathbf{x = (x\_TX - x\_RX)} \qquad \mathbf{y = (y\_TX - y\_RX)} \qquad \mathbf{z = (z\_TX - z\_RX)}\ .$$

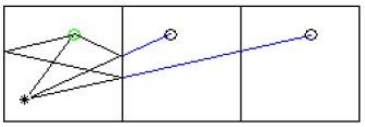
Los apéndices "_TX" y "_RX" detrás de las señales denotan las posiciones del transmisor TX (la fuente de sonido) y el receptor RX (el oyente o micrófono) respectivamente, para cada una de las coordenadas (x, y, z).

En un primer acercamiento a la teoría psicoacústica del audio 3D, es importante relacionar los atributos físicos de las fuentes sonoras con las características de localización. El escenario más simple es considerar una sola fuente en campo libre *(free-field).*

3.4. Modelo de respuesta al impulso de la sala (RIR)

En esta sección se quiere presentar la publicación [2], que ha sido el modelo usado para crear las RIRs del software desarrollado. El modelo usa el Método de Imágenes descrito anteriormente.

Para ilustrar el proceso de propagación del sonido en la sala, considere una fuente sonora representada



por un circulo verde y un receptor, micrófono u oyente, dentro de la sala representado por una estrella negra, como muestra la figura.

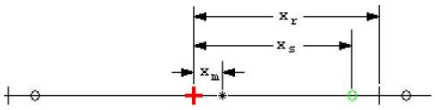
La línea que une ambos puntos es el camino acústico que toma la onda sonora. Este es el sonido directo. Por ser un recinto cerrado y la fuente suponerse como radiando omnidireccionalmente ondas esféricos, otra parte del sonido rebota con una pared y llega al micrófono. La onda reflejada es lo que escuchamos y conocemos como un eco. Al analizar este fenómeno desde el punto de vista de la Acústica de Rayos, se deduce que el oyente percibe este eco como si lo radiara una fuente que en realidad no existe, desde una posición detrás de la pared, como muestra la siguiente figura. El establecer una imagen de la sala y situarla adyacente a la original, si estuviéramos en la posición del micrófono percibiríamos algo como estas dos fuentes radiando a la vez. En este contexto, esta fuente imagen se denota también por *fuente virtual*. No en el sentido de posicionar una fuente real en un espacio 3D de sonido y denominar a la posición modificada con el nombre de fuente virtual, sino en el sentido de aumentar el número de fuentes (la original, en verdad siempre está fija). En la figura, las líneas azules representan el camino percibido de la onda sonora, mientras que las líneas negras muestran el camino que realmente toma la onda al dispersarse por la habitación, perdiendo energía en cada rebote.

Este proceso se puede extender a dos y tres dimensiones y repetirlo constantemente haciendo una imagen de la imagen para cada dirección, y así sucesivamente. La figura siguiente muestra un diagrama de las fuentes virtuales generadas para una superficie bidimensional.

En este modelo se considera que las fuentes virtuales no emiten ningún eco.

3.4.1 Posicionamiento de las fuentes imagen virtuales

En un espacio tridimensional, queremos encontrar las posiciones de las fuentes virtuales mas cercanas al oyente. Para empezar, considere el caso de una dimensión con ayuda del siguiente esquema. No obstante, se denotan el origen de coordenadas (0, 0, 0) en rojo, la fuente de sonido real en verde con coordenadas (xs, ys, zs), las dimensiones de la sal (xr, yr, zr), la fuente virtual en negro y el oyente como una estrella negra con coordenadas (xm, ym, zm).



Analizando detenidamente la situación, se puede demostrar que la coordenada x de cada fuente virtual se puede expresar como:

$$\mathbf{x} = (-\mathbf{1})^{\mathbf{i}}\,\mathbf{x_s} + \left[\mathbf{i} + \frac{\mathbf{1} - (-\mathbf{1})^{\mathbf{i}}}{\mathbf{2}}\right]\mathbf{x_r}$$

La posición de la fuente virtual i-ésima queda determinada al sustituir i por un entero. Si i se elige negativo, entonces la fuente virtual está a la izquierda del eje x. Al considerar un espacio tridimensional, las ecuaciones para calcular las posiciones relativas (xi, yj, zk) de las fuentes virtuales al micrófono son:

$$\mathbf{x_i} = (-\mathbf{1})^{\mathbf{i}}\,\mathbf{x_s} + \left[\mathbf{i} + \frac{\mathbf{1} - (-\mathbf{1})^{\mathbf{i}}}{\mathbf{2}}\right]\mathbf{x_r} - \mathbf{x_m}$$

$$\mathbf{y_j} = (-\mathbf{1})^{\mathbf{j}}\,\mathbf{y_s} + \left[\mathbf{j} + \frac{\mathbf{1} - (-\mathbf{1})^{\mathbf{j}}}{\mathbf{2}}\right]\mathbf{y_r} - \mathbf{y_m}$$

$$\mathbf{z_k} = (-\mathbf{1})^{\mathbf{k}}\,\mathbf{z_s} + \left[\mathbf{k} + \frac{\mathbf{1} - (-\mathbf{1})^{\mathbf{k}}}{\mathbf{2}}\right]\mathbf{z_r} - \mathbf{z_m}$$

La distancia a cada fuente virtual se calcula usando el teorema de Pitágoras para tres dimensiones:

$$\mathbf{d_{ijk}} = \sqrt{\mathbf{x_i^2} + \mathbf{y_j^2} + \mathbf{z_k^2}}$$

3.4.2. Calculo de la respuesta al impulso unidad para cada fuente imagen virtual

Definiendo una ecuación para la diferencia de tiempos

$$\mathbf{u_{ijk}(t)} = \mathbf{t} - \frac{\mathbf{d_{ijk}}}{\mathbf{c}}$$

siendo t el tiempo actual, la función respuesta al impulso unidad es:

$$\mathbf{a_{ijk}}\left(\mathbf{u_{ijk}(t)}\right) = \begin{cases} \mathbf{1}, & \mathbf{si}\ \ \mathbf{u_{ijk}} = \mathbf{0} \\ \mathbf{0}, & \mathbf{resto} \end{cases}$$

Con la definición anterior, cada respuesta al impulso unidad vale uno cuando u(i, j, k) = 0. Sería deseable corregir esto debido a que hay dos factores que afectan a la magnitud de los ecos: el primera es la distancia que recorre la onda que introduce cierta atenuación, el segundo es el numero de reflexiones que la onda sonora hará mientras exista. Denotando la magnitud total del eco por e(i, j, k) podemos decir que

$$\mathbf{e_{ijk}} = \mathbf{b_{ijk}}\,\mathbf{r_{ijk}}$$

siendo b(i, j, k) la contribución debido al primer factor y r(i, j, k) la contribución debida al segundo factor. Podemos deducir directamente que b(i, j, k) es proporcional al inverso de la distancia:

$$\mathbf{b_{ijk}} \propto \frac{\mathbf{1}}{\mathbf{d_{ijk}}}$$

Para deducir conclusiones acerca de cómo afecta el número de reflexiones, digamos primero que todos los coeficientes de reflexiones de las paredes son iguales y valen rw. Podemos hacer una estimación como:

$$\mathbf{r_{ijk}} = (\mathbf{r_w})^{\boldsymbol{n}}\,, \qquad \boldsymbol{n} = |\boldsymbol{i}| + |\boldsymbol{j}| + |\boldsymbol{k}|$$

Con n el numero total de reflexiones que ha sufrido la onda. Si cada pared tiene un coeficiente de reflexión distinto, entonces el problema es un poco mas complicado de modelar.

Siendo $\mathbf{r_{x=0}}$ el coeficiente de reflexión para la pared perpendicular al eje x mas cercana al origen, y $\mathbf{r_{x=x_r}}$ el coeficiente de reflexión para la pared opuesta, entonces el coeficiente de reflexión para todas las reflexiones de la fuente i-ésima a lo largo del eje x se puede calcular como:

$$\mathbf{r_{x_i}} = \left(\mathbf{r_{x=0}}\right)^{\left|\frac{1}{2}i - \frac{1}{4} + \frac{1}{4}(-1)^i\right|} \cdot \left(\mathbf{r_{x=x_r}}\right)^{\left|\frac{1}{2}i + \frac{1}{4} - \frac{1}{4}(-1)^i\right|}$$

De igual manera, el coeficiente de reflexión para las fuentes j-ésima y k-esima a lo largo de los ejes y, z

$$\mathbf{r_{y_j}} = \left(\mathbf{r_{y=0}}\right)^{\left|\frac{1}{2}j - \frac{1}{4} + \frac{1}{4}(-1)^j\right|} \cdot \left(\mathbf{r_{y=y_r}}\right)^{\left|\frac{1}{2}j + \frac{1}{4} - \frac{1}{4}(-1)^j\right|}$$

$$\mathbf{r_{z_k}} = \left(\mathbf{r_{z=0}}\right)^{\left|\frac{1}{2}k - \frac{1}{4} + \frac{1}{4}(-1)^k\right|} \cdot \left(\mathbf{r_{z=z_r}}\right)^{\left|\frac{1}{2}k + \frac{1}{4} - \frac{1}{4}(-1)^k\right|}$$

El coeficiente de reflexión total se calcula multiplicando estas tres relaciones:

$$\mathbf{r_{ijk}} = \mathbf{r_{x_i}}\ \mathbf{r_{y_j}}\ \mathbf{r_{z_k}}$$

Con todo, tenemos el factor de atenuación total:

$$\mathbf{e_{ijk}} = \mathbf{b_{ijk}}\ \mathbf{r_{ijk}} = \frac{\mathbf{1}}{\mathbf{d_{ijk}}}\ \mathbf{r_{x_i}}\ \mathbf{r_{y_j}}\ \mathbf{r_{z_k}}$$

3.4.3. Construcción de la RIR

Para obtener la respuesta al impulso de la sala, se multiplican los factores de atenuación con las funciones respuesta al impulso unidad y se suman los tres índices (i, j, k). Para entender esto mediante un ejemplo físico real, podemos suponer que la suma expresa el "fluir" (radiación) de todas las fuentes virtuales de sonido desde todas las direcciones. El resultado es:

$$\mathbf{h(t)} = \sum_{\mathbf{i=-n}}^{\mathbf{n}} \sum_{\mathbf{j=-n}}^{\mathbf{n}} \sum_{\mathbf{k=-n}}^{\mathbf{n}} \left(\mathbf{a_{ijk}} \cdot \mathbf{e_{ijk}}\right)$$

3.4.4. Función Matlab® rir.m

La funcion que usa el proyecto tiene dos diferencias con respecto a este modelo. Primero, usa tiempo discreto en vez de tiempo continuo. Segundo, sólo calcula a(i, j, k) y e(i, j, k) cuando a(i, j, k) = 1 para no usar tanta memoria. El filtro generado es muy largo (más de 20000 taps) y una convolución ordinaria es demasiado lenta. Por ello, se recomiendan algoritmos de convolución rápida para usarla.

3.4.5. Factores que no se han tenido en cuenta

Este modelo hace ciertas suposiciones que se sabe son fuentes de error: que los coeficientes de reflexión son independientes del angulo de incidencia y de la frecuencia, que no hay cambio de fase en una onda que se refleja, que el gas aire no tiene efecto en la magnitud de la onda, además de otros.

3.4.6. Convolución del audio con la RIR

Si la RIR de una sala está disponible en algún formato para ser procesada por el software, la reverberación más realista se consigue convolucionando la señal de audio con dicha respuesta. Se puede hacer una convolución directa convencional guardando cada muestra de la RIR como un coeficiente de un filtro FIR a cuya entrada se alimenta con la señal de audio original (sin efectos digitales de ningún tipo). No obstante, esta convolución directa se convierte fácilmente en algo poco práctico en cuanto la longitud del audio es mayor que aprox. 1 s, debido a la larga longitud de la RIR. Miles de taps son necesarios para que un filtro FIR modele una RIR, como ejemplo ilustrativo una RIR con 20000 taps no es un tamaño en absoluto grande.

Para auditorios, estadios y gimnasios los tamaños de los archivos que contienen estas respuestas al impulso pueden alcanzar tamaños de 30 GB o más.

La solución a este problema es una convolución lo más rápida y eficiente posible [3, 4]. Se lleva a cabo bloque a bloque en el dominio de la frecuencia usando las transformadas de Fourier de la RIR y de la señal de audio mediante algoritmos FFT. Ya que este procesamiento se realiza en bloques sucesivos de la señal de entrada, la señal de salida se obtiene solapando y sumando los resultados parciales (ver capítulo 3, sección 3.4.) Gracias a la FFT esta convolución rápida se puede hacer relativamente deprisa.
Una deficiencia es que, para querer ser usada en tiempo real, un bloque de L muestras tiene que ser leído y procesado al mismo tiempo que el siguiente bloque es leído. Por ello, la latencia entrada-salida es 2·L, que es un valor inaceptable (demasiado alto) en implementaciones practicas para tiempo real.

Un tercer tipo de convolución, que supera a los dos métodos de convolución anteriores, se basa en una descomposición de la RIR en bloques de longitud variable, cada bloque dos veces mayor que el anterior. Esto permite que la latencia de la computación previa sea usada para los cálculos de la siguiente trama de la RIR [3, 4].

3.5. Caracterización de altavoces monitores de estudio para auralización

3.5.1. Introducción

En esta sección se desea presentar un método para caracterizar en groso modo altavoces, obtener una respuesta en frecuencia y en fase aproximada, su patrón de directividad y algunas de las características de distorsion no lineal más relevantes. Todo lo expuesto aquí se ha expuesto anteriormente también en [5]. En concreto, los altavoces son monitores de estudio activos. Estos son unos altavoces con una respuesta en frecuencia casi plana en el rango de audición y que son adecuados para la reproducción de material binaural y/o transaural para propósitos de escucha detallada. Además, incorporan dentro de la misma caja también las etapas de amplificación, por lo que no se requiere de un amplificador dedicado externo. Unos altavoces de gama baja no dejarán apreciar al ingeniero de sonido los matices importantes del audio binaural necesarios para avanzar y mejorar en la síntesis de audio 3D que se quiera desarrollar. En resumen: es importante equipar al sistema de reproducción con unos altavoces y auriculares adecuados a la magnitud del proyecto que se quiera implementar.

Además, la obtención de las características arriba mencionadas permite la ecualización del material de audio, lo que mejora la calidad de este ya que no se introducen coloración y/o distorsiones excesivas debidas al dispositivo de radiación y su respuesta en frecuencia.

La Auralización es una técnica que, aparte de permitir generar audio 3D, posibilita la simulación de las características frecuenciales de un altavoz para la escucha sólo con auriculares. Al filtrar el audio (estéreo o 3D) con la respuesta en frecuencia del altavoz, se consigue en los auriculares un sonido muy parecido al que se escucharía por esos altavoces determinados por esa respuesta en frecuencia.

Usando una señal test especial, la respuesta al impulso y el patrón de directividad pueden ser medidos en una sala de grabación pequeña. Mediciones de campo cercano son también posibles.

Se presenta también un procedimiento para combinar las respuestas de campo cercano y de campo lejano para deducir de estas las reflexiones tempranas de la sala. El resultado es un conjunto de respuestas al impulso y respuestas direccionales que son suficientemente detalladas para una auralización convincente. El procedimiento se compone de cuatro partes: 1) modelo de radiación de sonido, 2) simulador de reverberación, 3) modelo de características direccionales relacionadas con la cabeza y 4) imprimación de las características binaurales en el audio en tiempo real [5].

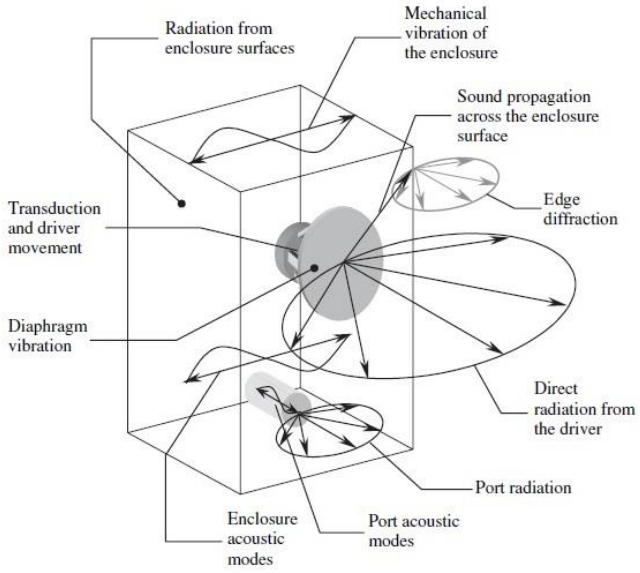
Es sabido que hay muchas sutilezas en cuanto a cómo afectan los sistemas de altavoces al sonido irradiado, tanto on-axis como al campo reverberante. Esto hace que el modelado de altavoces mediante análisis teórico sea poco menos que imposible, con lo que las simulaciones se basan en respuestas al impulso grabadas

No se han querido dejar reflejadas mediciones de ningún tipo llevadas a cabo con los altavoces (HARMAN-JBL® LSR 305) ya que esto excedía los límites de este proyecto. El único fin de este apartado es proporcionar una guía o manual para llevar a cabo estas mediciones en caso de ser necesarias para los altavoces.

3.5.2. Propiedades individuales de un altavoz

Hay mucha literatura disponible que discute por qué altavoces diferentes suenan diferentes. Asumiendo que las unidades fundamentales en el sistema altavoz han sido correctamente diseñadas y son lineales en la medida de lo normal, lo que sigue es un pequeño resumen.

En la siguiente imagen se muestra una descomposición de todos los efectos acústicos que puede sufrir un altavoz desde el punto de vista del modelado. La descripción de cada uno de estos efectos es demasiado extensa para ser expuesta aquí, pero una visión general permite hacerse una idea de la dificultad que significa esta tarea.



Factores que afectan a la respuesta on-axis

· *Tamaño del encapsulado que encierra al altavoz.*

El volumen físico que ocupa el encapsulado es la variable mas significativa que determina la respuesta en baja frecuencia del altavoz. Cuanto mas pequeña sea la caja, mas alta será la frecuencia resonante de la unidad de baja frecuencia, imposibilitando radiar por debajo de esa frecuencia resonante.

Cuanto mas grande sea la caja, normalmente usara bafles mayores que moverán mas cantidad de aire. Aunque un filtro activo puede incrementar la potencia entregada a la unidad de baja frecuencia y por tanto extender la respuesta en baja frecuencia, hay un limite practico. Esto es asi porque la cantidad de desplazamiento del bafle requerida se incrementa en un factor 4 para cada octava de baja frecuencia requerida.

El altavoz de este proyecto tiene unas dimensiones de 298 mm x 185 mm x 231 mm. El tamaño del driver de baja frecuencia es de 127 mm y el de alta frecuencia de 25 mm. Ambos usan amplificadores clase D de 41 W. Más detalles sobre este tipo de ampliicador en [10].

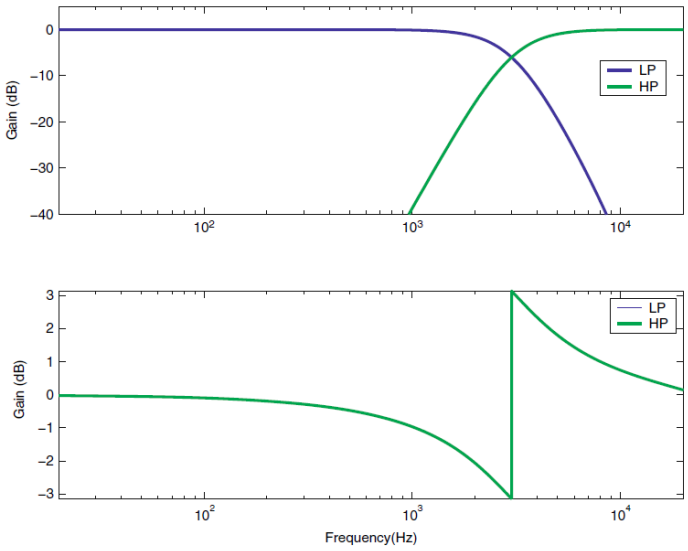
· *Puertos bass-reflex, radiadores pasivos y líneas de transmisión*. Un bass-reflex o puerto de reflexión es un resonador de Helmholtz añadido a la caja del altavoz y atenuado usando materiales porosos tales como espumas o fibras minerales. En general, se ajusta a una frecuencia una octava por debajo de la frecuencia resonante del woofer. Esto hace que la respuesta en baja frecuencia se extienda un poco. La ventaja de esta extensión del rango del altavoz es contrarrestada por el roll-off de baja frecuencia debido al puerto de reflexión (24dB/octava en vez de 12dB/octava). Esto afecta a la respuesta en fase de baja frecuencia y produce un pico en la respuesta al impulso del altavoz que va a afectar a señales transitorias rápidas.

Los radiadores pasivos y las líneas de transmisión son otra de las maneras puramente mecánicas de afectar a la carga de la unidad de graves y por tanto de extender la respuesta en baja frecuencia. El efecto es similar a los puertos de reflexión pero menos pronunciado (añaden 6dB/octava al roll-off de baja frecuencia).

· *Diseño del crossover.*

El crossover electrónico que distribute la señal de audio a los dos bafles de baja y alta frecuencia influye en la respuesta on-axis del altavoz. Un crossover cuidadosamente diseñado será plano on-axis; no obstante los crossovers pasivos típicos crean cambios de fase entre los bafles. En otras palabras, hay una hendidura o pico pronunciado en la respuesta en frecuencia a la frecuencia de crossover.

El altavoz elegido para este proyecto usa un crossover tipo Linkwitz-Riley de orden 4 con frecuencia de crossover a 1.725 kHz. Un crossover tipo LR de orden par se consigue poniendo en cascada dos filtros Butterworth idénticos. Por completitud, la figura siguiente muestra la respuesta en frecuencial tipica de un crossover LR. Se puede encontrar información muy útil sobre crossovers para altavoces en [12].



· *Difracción debida a la caja o encapsulado.*

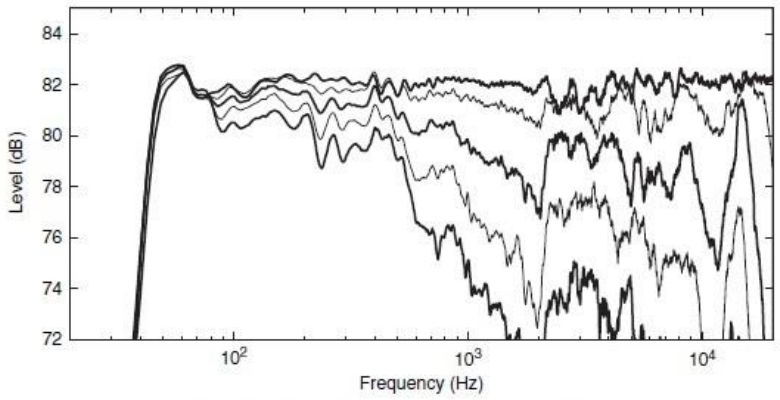
Las ondas sonoras se difractan alrededor de un obstáculo cuyo tamaño sea menor que la longitud de onda del sonido; para estas longitudes de onda el obstáculo es esencialmente ignorado. Por encima de un cierto umbral, los bordes o cantos del obstáculo presentan un cambio brusco a la onda de presión en cuanto a impedancia acústica y se generan y propagan ondas secundarias. El resultado es una respuesta tipo filtro peine (comb filter) en la posición de escucha. Altavoces con cantos afilados sufren mas este efecto que altavoces mas redondeados, para los que el cambio de impedancia acústica es mas gradual.

Se recomienda no usar altavoces en forma de cubo con todos los cantos a la misma distancia de un bafle, ya que sufrirán este efecto de la peor manera.

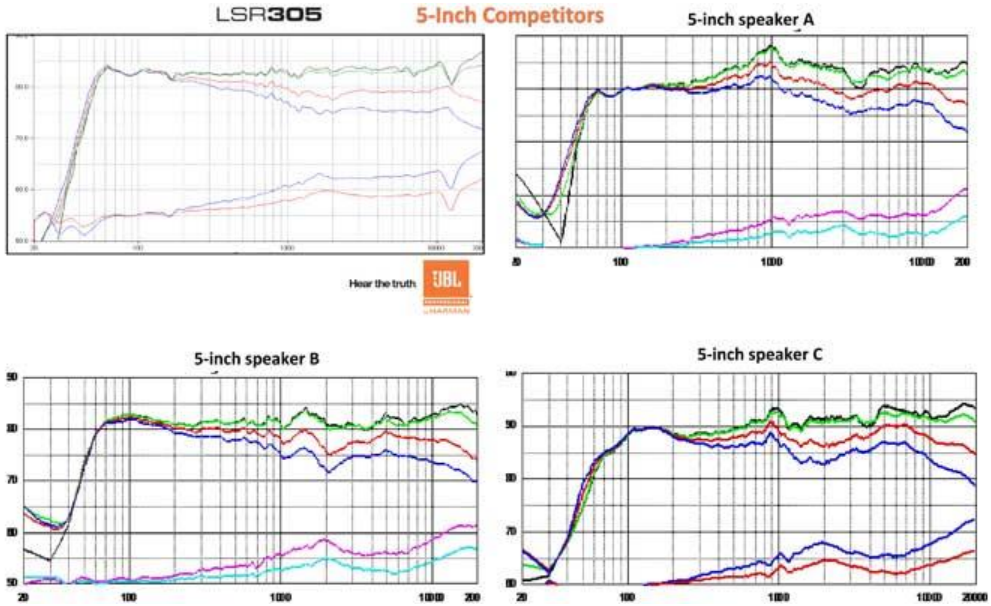
· *Reflexiones de la parte de atrás del encapsulado.*

En general, para la mayoría de altavoces la onda que emana del bafle de alta frecuencia (el tweeter) es reflejada por la parte trasera de la caja del altavoz. Esta reflexión impacta con la parte trasera del tweeter unos cuantos cientos de microsegundos después y el tweeter radia un eco. Este eco filtra tipo peine la respuesta en alta frecuencia (la respuesta del tweeter)

Como ilustración del cambio en la respuesta en frecuencia del altavoz conforme el oyente se mueve horizontalmente lejos del eje, la siguiente figura es clara. De arriba abajo, la primera curva es para el oyente posicionado on-axis, luego 15º off-axis, 30º, 45º y 60º off-axis (en el plano horizontal).



En la siguiente figura se muestra la respuesta en frecuencia de los monitores de 5'' HARMAN-JBL® LSR 305 en un rango de 50 dB a 90 dB, en comparación con otros tres altavoces también de 5''.
Las líneas verdes muestran la respuesta en frecuencia para la ventana de escucha. Las líneas rojas muestran la respuesta a las primeras reflexiones. Las líneas azules reflejan la potencia sonora total.
Los datos de las líneas inferiores de cada figura no se han podido obtener y se consideran irrelevantes para la discusión. Se deben ignorar.



Se puede apreciar la mayor regularidad en todas las respuestas del altavoz LSR 305, lo que hacen de este altavoz una herramienta útil para la escucha detallada de material binaural, con una respuesta en frecuencia bastante plana.

Además, hay un beneficio adicional en estos altavoces y es que incorporan la tecnología de guía de onda "Image Control Waveguide" para el crossover y ampliación del sweet spot. Pensando en el caso de escucha para más de un oyente, estos altavoces son útiles al aumentar sweet spot (disminuyen levemente la variación tan brusca de la respuesta en frecuencia debida a una posición no simétrica, off-axis).

Factores que afectan a la respuesta off-axis

· *Directividad vs. Frecuencia.*

Los altavoces con forma de cono se aproximan a la descripción de pistón. Debido a que tienen un ancho físico, no pueden radiar omnidireccionalmente a todas las frecuencias. Las cancelación de fase producidas por diferencias en el tiempo de llegada entre los cantos mas cercanos y lejanos causaran ceros locales an algunas posiciones de escucha. Cuando la longitud de onda sea lo suficientemente grande, este efecto será muy pronunciado.
Por encima de una cierta frecuencia se formaran ondas estacionarias en el cono del altavoz y se convierte en un *phased array,* enfocanfo el sonido hacia adelante a costa de una peor respuesta omnidireccional. No obstante, crossovers bien diseñados previenen este problema.

· *Resonancia de los paneles.*

La cubierta de los altavoces sirve para contener la onda de presión que se crea detrás de los drivers. Si no existiera, el desplazamiento de un bafle simplemente haría circular aire entre la parte frontal y la trasera y se radiaría demasiado poca energía hacia adelante. En la practica, ninguna caja es perfectamente rigida y los paneles de la caja resonaran a diferentes frecuencias. El resultado es una respuesta direccional no ideal, sobre todo a bajas frecuencias. En casos extremos esto puede efectar también a la respuesta on-axis.

· *Apantallamiento acústico*

Cuando la longitud de onda correspondiente a la onda acústica radiada es menor o aproximadamente igual al tamaño de la caja del altavoz, la caja hace de pantalla acústica que icnrementa la directividad a altas frecuencias.

· *Alineamiento de fase de los bafles*

Las diferencias de fase causadas por las señales de los diferentes bafles de un altavoz que llegan en tiempos ligeramente diferentes afectaran a la respuesta en frecuencia resultanto cuando la posición de escucha esta mas cerca de un altavoz que del otro. No obstante, cuando los altavoces están correctamente alineados este fenómeno no afecta a la respuesta on-axis; tampoco afecta al SPL total radiado en una sala. Es también algo inusual que este fenómeno influya al campo reverberante.

Stereo Matching

Los altavoces de clase media y baja no suelen usar componentes de alta tolerancia en sus crossovers. El control de calidad de los materiales usados en este tipo de altavoz también puede dejar que desear. Resultado: discrepancias de fase y frecuencia entre los canales L y R. Esto daña la capacidad de imagen estéreo del sistema a las frecuencias medias y altas, las mas importantes en cuanto a transmisión de voz. Para una auralización de calidad usando altavoces de grado medio es fundamental por tanto recabar información de todos los altavoces del sistema para obtener y compensar estas discrepancias.
No obstante, se puede asumir que para el mismo tipo de altavoces se tienen casi las mismas propiedades de direccionalidad, incluso cuando su respuesta on-axis difiera, ya que la direccionalidad tiene como factor dominante las dimensiones de la caja y de los baffles.
Por ello, solo habrá que modificar la respuesta on-axis de cada uno de los altavoces.

Distorsión no lineal

Hay un cierto numero de distorsiones lineales asociadas a altavoces. La compresión de potencia ocurre en el periodo de tiempo cuando la bobina de voz del bafle se va calentando, aumentando su resistencia. Debido a ello, el mismo voltaje ya no desplazara el bafle la misma distancia. Esto resulta en una disminución en el nivel de presión sonora de salida. No obstante, debido a que la compresión de potencia es un efecto de memoria (requiere conocer señales de larga duración pasadas), es raro incluirlo.

También seria incómodo para un usuario tener que "dejar reposar" una simulación de la misma manera que un altavoz tiene que reposar para contrarrestar la compresión de potencia. Es por ello que no se incluye en el algoritmo de auralización.

La distorsión armónica, de intermodulación y otras no lineales son causadas normalmente por los siguientes cuatro mecanismos:
· componentes sueltos o mal aislados/montados que resuenan
· el mecanismo de suspensión o la bobina son empujados más allá de su operación lineal normal cuando el nivel de presión sonora es alto
· Cambio en el espectro de audio de alta frecuencia debido a efecto Doppler por audio de baja frecuencia irradiado por el mismo bafle.

De estos cuatro efectos, solo el cuarto interesa para el algoritmo de auralización ya que los dos primeros pertenecen sobre todo al ámbito de montaje y calidad de componentes del altavoz y el último se da solo en sistemas donde se requieren niveles de presión muy altos como en espacios públicos o en altavoces bocina.

Por tanto, el efecto Doppler requiere una explicación. Depende fuertemente del nivel de presión sonora absoluto, pero una primera aproximación de su efecto puede derivarse de la siguiente formula que relaciona el desplazamiento del bafle con las propiedades físicas del altavoz y la onda de sonido:

$$s = \frac{\rho \cdot p}{f^2 \cdot A}$$

siendo s el desplazamiento máximo del baffle en metros, ρ la densidad del aire, p es la presión sonora en Pascales, f es la frecuencia de la señal de entrada y A es la superficie en movimiento del baffle.

La distorsión debida al efecto Doppler en el bafle de bajas frecuencias es la mas acusada en comparación con cualquier otro bafle del altavoz. El audio que pasa por la unidad de bajas frecuencias es re-muestreado con el retardo proporcional al desplazamiento absoluto. Esto puede añadirse al audio de entrada de manera relativamente eficiente usando un algoritmo de resampling bicúbico.

Se puede derivar un procedimiento para transformar la señal de audio de entrada en desplazamiento instantáneo para un altavoz en particular; el procedimiento consta de cuatro pasos:

1) Integrar la señal de audio dos veces:

$$\int \left(\int s(t)\, dt \right) dt$$

Así se consigue una señal cuyo valor para cualquier frecuencia es proporcional a $1/f^2$.

2) usar un filtro paso-bajo simple a una frecuencia muy baja (menor que 1 Hz) para eliminar la componente de DC que resulta de la integración doble.

3) multiplicar la señal resultante por un factor de escala apropiado para calcular el numero de muestras de la desviación requerido para resampling del audio. El factor de escala que convierte la señal integrada dos veces a un desplazamiento en muestras es

$$\mathbf{k_d = \frac{\pi^2 \cdot \rho \cdot p_0}{F_s \cdot A \cdot c}}$$

siendo p0 la presión deseada para un nivel de entrada igual a uno, c es la velocidad del sonido a la temperatura de la sala (aprox. 345 m / s), A la superficie del bafle y Fs la frecuencia de muestreo del audio.

La gran ventaja de conocer esta formula es que los efectos de distorsión debidos al efecto Doppler se pueden reproducir mediante auriculares basándose en las dimensiones del bafle de baja frecuencia, la frecuencia del crossover y el SPL pico deseado.

No obstante, pruebas de audición informales revelan que el efecto de la distorsión Doppler es casi imperceptible en sistemas simulados con bafles de baja frecuencia de 6''. Por ello, se puede obviar en una simulación para auralización de altavoces.

Para más información sobre técnicas de medida y obtención de respuestas en frecuencia para altavoces, se recomienda acudir a [13-16].

3.6. Método de Solapamiento y Suma. Realización de sistemas lineales e invariantes en el tiempo mediante la DFT y algoritmos FFT

3.6.1. Introducción

En esta sección, dedicada al tratamiento digital de señales, se presentan las herramientas matemáticas utilizadas por aplicaciones para computar y operar sobre señales discretas obtenidas del muestreo de señales continuas mediante el teorema de muestreo o teorema de Nyquist. La transformada de Fourier es la pieza fundamental de todo este este proceso.
Se presentan la transformada discreta de Fourier (DFT) así como el método o algoritmo de Solapamiento y Suma (OLA), que es una herramienta útil para hallar la convolución lineal de una señal de larga duración.

3.6.2. Convolución de dos señales de longitud finita

Es computacionalmente eficiente realizar la convolución de dos señales

$$y(t) = s(t) * h(t) = \int s(\tau) \cdot h(t-\tau)\, d\tau$$

en tiempo discreto mediante el siguiente procedimiento [17]:

1) Calcular las transformadas discretas de Fourier (DFTs) de N puntos de las secuencias s(n) y h(n):

$$\mathbf{H(k)} = \sum_{\mathbf{n=0}}^{\mathbf{N-1}} \mathbf{h(n)} \cdot \mathbf{e}^{-\mathbf{j}\frac{\mathbf{2\pi}}{\mathbf{N}}\mathbf{k\,n}}$$

2) Calcular el producto $Y(k) = S(k) \cdot H(k)$, para $0 \leq k \leq N-1$

3) Obtener la secuencia y(n) como la IDFT de S(k):

$$\mathbf{h(n)} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{n=0}}^{\mathbf{N-1}} \mathbf{H(k)} \cdot \mathbf{e}^{\mathbf{j}\frac{\mathbf{2\pi}}{\mathbf{N}}\mathbf{k\,n}}$$

La operación de DFT y de IDFT se realiza, en un computador, en general usando algoritmos de computo eficientes denominados *algoritmos FFT* [17]. En la mayoría de las aplicaciones, nuestro interés es realizar una *convolución lineal* de dos secuencias. Es decir, deseamos realizar un *sistema lineal e invariante con el tiempo LTI* [18]. Esto ocurre al filtrar la señal de audio o al calcular una función de auto correlación entre las señales de ambos canales yL(n) e yR(n). De la teoría de tratamiento digital de señales se sabe que la multiplicación de DFTs corresponde a la convolución circular de las secuencias. Para obtener una convolución lineal hay que asegurar que la convolución circular tiene el efecto de la convolución lineal.

Consideremos otra vez una secuencia s(n) de longitud L y una secuencia h(n) de longitud H. Supongamos que se desea combinar esas dos secuencias convolucionándolas, de manera que se obtiene otra secuencia y(n)

$$\mathbf{y(n)} = \mathbf{s(n)} * \mathbf{h(n)} = \sum_{\mathbf{k=-\infty}}^{\mathbf{+\infty}} \mathbf{s(k)} \cdot \mathbf{h(n-k)}$$

con $h(n) = 0$ para $n > H$. Es claro que el producto $s(k) \cdot h(n-k) = 0$ siempre que $n < 0$ y $N > L + H - 2$. Es decir,

$$\mathbf{y(n) = 0, \qquad n > L + H - 2}$$

Por tanto, $(L + H - 1)$ es la longitud máxima de y(n) resultante de la convolucion lineal de una secuencia de longitud L con otra secuencia de longitud H.

3.6.3. La convolución circular como una convolución lineal con solapamiento

Como se ha comentado, el que una convolución circular (correspondiente al producto de dos DFTs de N puntos) tenga el mismo valor que la convolución lineal de las secuencias de longitud finita, dependerá de la longitud de la DFT (denotada por la letra N) en relación con la longitud de las secuencias de longitud finita. Una interpretación muy útil de la relación entre la convolución lineal y la circular se puede dar en términos de *solapamiento temporal* [17, 18].

Primero, póngase de manifiesto que la DTFT $\mathbf{S(e^{j\omega})}$ de una secuencia s(n) , muestreada a las frecuencias

$$\boldsymbol{\omega_k = \frac{2\pi k}{N}},$$

corresponde a los coeficientes del desarrollo en serie de Fourier de la secuencia periódica:

$$\boldsymbol{\tilde{s}(n) = \sum_{r=-\infty}^{+\infty} s(n - rN)}$$

De esto se deduce que la secuencia de longitud finita

$$\boldsymbol{S(k) = \begin{cases} S\left(e^{j2\pi k/N}\right), & 0 \le k \le N-1 \\ 0, & resto \end{cases}}$$

corresponde a la DFT de un solo periodo de $\tilde{\mathbf{s}}(\mathbf{n})$. Es decir:

$$\boldsymbol{s_p(n) = \begin{cases} \tilde{s}(n), & 0 \le n \le N-1 \\ 0, & resto \end{cases}}$$

Obviamente, si s(n) tiene longitud menor o igual que N, no se producirá solapamiento temporal y $\mathbf{s_p(n) = s(n)}$. Sin embargo, si la longitud de s(n) es mayor que N, $\mathbf{s_p(n)}$ puede no ser igual a s(n) par algunos valores de n o para ninguno. El subíndice p indica que una secuencia es un periodo de la secuencia periódica resultante de realizar la IDFT de una DTFT muestreada.

La secuencia y(n) tiene como DTFT

$$\boldsymbol{Y(e^{j\omega}) = S(e^{j\omega}) \cdot H(e^{j\omega})}$$

Si definimos la DFT de y(n) como:

$$\boldsymbol{Y(k) = Y\left(e^{j2\pi k/N}\right), \quad 0 \le k \le N-1}$$

se desprende que

$$\boldsymbol{Y(k) = S\left(e^{j2\pi k/N}\right) \cdot H\left(e^{j2\pi k/N}\right), \ 0 \le k \le N-1}$$

Con ello,

$$\boldsymbol{Y(k) = S(k) \cdot H(k)}$$

Esto indica que la secuencia resultante de realizar la IDFT de Y(k) es

$$\boldsymbol{y_p(n) = \begin{cases} \sum_{r=-\infty}^{+\infty} y(n - rN), & 0 \le n \le N-1 \\ 0, & resto \end{cases}}$$

y utilizando la ecuación anterior se tiene

$$y(n) = s(n) \circledast_N h(n)$$

donde el operador $\circledast_N$ denota la convolucion circular de N puntos. Luego, se concluye que la convolución circular de dos secuencias de longitud finita es equivalente a la convolución lineal de dichas secuencias, seguida por el solapamiento temporal que indica la ecuación para $y_p(n)$.

Si N es mayor o igual que L o H, S(k) y H(k) representan exactamente a s(n) y h(n), pero $y_p(n) = y(n)$ para todo n solo si N es mayor o igual que la longitud de la secuencia y(n). Como se sabe, si s(n) tiene longitud L y h(n) tiene longitud H , la longitud máxima de y(n) será (L + H – 1).

Por consiguiente la convolución circular correspondiente a $S(k)\cdot H(k)$ es idéntica a la convolución lineal correspondiente a $S(e^{j\omega})\cdot H(e^{j\omega})$ si N (la longitud de las DFT) cumple que:

$$N \geq L + H - 1$$

Además, si $N = L = H$, todos los valores de la convolución circular pueden ser diferentes de los de la convolución lineal. Pero si $H < L$ algunos valores de la convolución circular coincidirán con los valores correspondientes de la convolución lineal.

En general, siempre que $H < L$ solo el termino $y(n + L)$ se solapara en el intervalo $0 \leq n \leq L - 1$. Cuando se suman esos términos, los últimos $(H - 1)$ puntos de $y(n + L)$, que se extienden desde $n = 0$ hasta $n = H - 2$, se sumaran con los $(H - 1)$ primeros puntos de $y(n)$ y los últimos $(H - 1)$ puntos de y(n), que se extienden desde $n = L$ hasta $n = L + H - 2$ serán descartados.
Finalmente, $y_p(n)$ se forma extrayendo la porción de $0 \leq n \leq L - 1$.

Como los últimos $(H - 1)$ puntos de $y(n + L)$ y los últimos $(H - 1)$ puntos de $y(n)$ son idénticos, el proceso de formar la convolución circular $y_p(n)$ se puede ver de forma alternativa como una convolución circular con solapamiento, tomando los $(H - 1)$ valores de $y(n)$ desde $n = L$ hasta $n = L + H - 2$, y sumándoselos a los primeros $(H - 1)$ puntos de y(n).

3.6.4. Desarrollo teórico

El algoritmo de solapamiento y suma soluciona el problema de la extensa convolución, dividiendo la operación en sucesivas convoluciones del filtro FIR h(n) con segmentos pequeños de la señal s(n), denotados por xi(n). Ya que esto resulta en toda la señal s(n) filtrada con el mismo filtro FIR h(n), es más adecuado para propósitos de filtrado adaptativo modificar este algoritmo para el problema de la auralización de audio. Cambiando de filtro h(n) en cada convolución, concretamente, usando J filtros FIR hi(n) para los J puntos de la trayectoria, se consiguen desarrollar los efectos psicoacústicos de espacialidad.

Se comienza dividiendo la señal de audio original s(n) en J segmentos o tramas de entrada xi(n) de longitud arbitraria L:

$$\mathbf{x_i(n)} = \begin{cases} \mathbf{s(n + i \cdot L)}, & \mathbf{n = 1, 2, \dots, L} \\ \mathbf{0}, & \mathbf{resto} \end{cases}$$

con $i = 1, 2, \dots, J$. Podemos reconstruir la señal original s(n) a partir de las tramas xi(n) mediante la relación:

$$\mathbf{s(n)} = \sum_{\mathbf{i=1}}^{\mathbf{J}} \mathbf{x_i(n - i \cdot L)}$$

La longitud de bloque L se elegía como potencia de 2 para el algoritmo FFT (aunque hay métodos mas efectivos). Con esto, la señal final y(n) se obtiene mediante la suma de todas las convoluciones:

$$\boxed{\mathbf{y(n)} = \sum_{\mathbf{i=1}}^{\mathbf{J}} \mathbf{x_i(n - i \cdot L) * h_i(n)} = \sum_{\mathbf{i=1}}^{\mathbf{J}} \mathbf{y_i(n - i \cdot L)}}$$

El lado derecho de la ecuación muestra que cada convolucion genera una trama de salida yi(n), que tienen longitud (en muestras) $L + H - 1$.
Para cualquier parámetro $N = L + H - 1$, esto es equivalente a la convolución circular de N puntos de xi(n) con h(n) en la región [1, N] (N más cercano a potencia de 2).

La gran ventaja de la convolución circular es que puede computarse muy eficientemente, por ejemplo con el teorema de convolución circular mediante la *transformada rápida de Fourier (Fast Fourier Transform FFT)* [17]:

$$\mathbf{y_i(n)} = \mathbf{IFFT\{FFT\{x_i(n)\} \cdot FFT\{h(n)\}\}}$$

donde IFFT{} y FFT{} denotan las transformadas rápidas inversa y directa de Fourier evaluadas sobre N puntos. Se supone que L > M.

Definamos una señal de entrada de audio mono s(n), de larga duración l:

$$\vec{\mathbf{s}} = \mathbf{s(n)} = [\mathbf{s_1}\ \mathbf{s_2}\ \dots\ \mathbf{s_L}\ \dots\ \mathbf{s_l}]$$

donde l viene dado en muestras. El primer índice empieza siempre en uno para que todo el desarrollo pueda evaluarse directamente en Matlab, el lenguaje preferido para este proyecto.
Deseamos trocear la señal de audio en J tramas, cada trama con L muestras, para procesarla y auralizar el audio que contiene. Al trocear se obtienen J tramas de entrada xi(n):

$$\overrightarrow{\mathbf{x_i}} = \mathbf{x_i(n)} = [\mathbf{x_1}^{\mathbf{i}}\ \mathbf{x_2}^{\mathbf{i}} \dots\ \mathbf{x_L}^{\mathbf{i}}], \qquad \mathbf{i = 1, 2, \dots, J}$$

Al expandirlas, estas relaciones quedan:

$$\begin{aligned}&\mathbf{x_1} = [\mathbf{x_1}^{\mathbf{1}}\ \mathbf{x_2}^{\mathbf{1}}\ \dots\ \mathbf{x_L}^{\mathbf{1}}]\\ &\mathbf{x_2} = [\mathbf{x_1}^{\mathbf{2}}\ \mathbf{x_2}^{\mathbf{2}}\ \dots\ \mathbf{x_L}^{\mathbf{2}}]\\ &\dots\\ &\mathbf{x_J} = [\mathbf{x_1}^{\mathbf{J}}\ \mathbf{x_2}^{\mathbf{J}}\ \dots\ \mathbf{x_L}^{\mathbf{J}}]\end{aligned}$$

Para calcular J (cuántas tramas se deben obtener de una señal de longitud l) empleamos:

$$\mathbf{J} = \mathbf{ceil}\left(\frac{\mathbf{l}}{\mathbf{L}}\right)$$

Si l / L no es entero, significa que las últimas muestras de señal entran de sobra en la última trama xJ(n), que tendrá que ser rellenada con algunos ceros. En principio, el número total de tramas J está asociado al número de puntos de la trayectoria ingresado en la aplicación, ambos iguales y de valor J.

Para los filtros FIR, la base de datos consta de J filtros hi(n) , i = 1, 2, ..., J, de longitud constante y valor M:

$$\overrightarrow{\mathbf{h_i}} = \mathbf{h_i(n)} = \left[\mathbf{h_1}^{\mathbf{i}}\ \mathbf{h_2}^{\mathbf{i}} \dots\ \mathbf{h_M}^{\mathbf{i}}\right], \qquad \mathbf{M = longitud\ de\ todos\ los\ filtros\ BRIR}$$

Por completitud, decir que estos filtros son respuestas binaurales de la sala (BRIR) calculadas como convolución de la HRTF (extraída para los ángulos que correspondan al momento i-ésimo) con la respuesta al impulso de la sala (RIR) que corresponda, en el momento i-ésimo:

$$\overrightarrow{\mathbf{h_i}} = \mathbf{h_i(n)} = \mathbf{BRIR_i(n)} = \mathbf{HRTF_i(\theta_i, \varphi_i)} * \mathbf{RIR_i}, \qquad \mathbf{n = 1, 2, \dots, M}$$

con i = 1, 2, ..., J. Con estos filtros se procesarán los sucesivos trozos de señal xi(n) para obtener las tramas de salida yi(n). En principio tenemos que convolucionar para ambos oídos, luego:

$$\mathbf{yL_i(n)} = \mathbf{x_i(n)} * \mathbf{hL_i(n)} = \sum_{\mathbf{k=-\infty}}^{+\infty} \mathbf{x_i(k)} \cdot \mathbf{hL_i(n-k)}$$
$$\mathbf{yR_i(n)} = \mathbf{x_i(n)} * \mathbf{hR_i(n)} = \sum_{\mathbf{k=-\infty}}^{+\infty} \mathbf{x_i(k)} \cdot \mathbf{hR_i(n-k)}$$

con

$$\overrightarrow{\mathbf{hL_i}} = \mathbf{hL_i(n)} = \mathbf{HRTF_{L_i}(\theta_i, \varphi_i)} * \mathbf{RIR_i}$$
$$\overrightarrow{\mathbf{hR_i}} = \mathbf{hR_i(n)} = \mathbf{HRTF_{R_i}(\theta_i, \varphi_i)} * \mathbf{RIR_i}$$

con i = 1,2, ..., J y n = 1, 2, ..., M. No obstante, este análisis toma por simplicidad solo uno de los dos conjuntos de filtros hi(n). Se supone que L > M.

Las tramas de salida yi(n) son el resultado de la convolución de las tramas de entrada xi(n) con los filtros BRIR hi(n):

$$\mathbf{y_i(n)} = \mathbf{x_i(n)} * \mathbf{h_i(n)} = \sum_{\mathbf{k=-\infty}}^{+\infty} \mathbf{h_i(k)} \cdot \mathbf{x_i(n-k)}, \qquad \mathbf{i = 1, 2, \dots, J}$$

donde n = 1, 2, ..., (N = L + M – 1).
Las componentes de cada trama de salida son:

$$y_i(n) = [y_1{}^i\ y_2{}^i \dots y_N{}^i]$$

Conviene expandir también estas relaciones como:

$$\mathbf{y_1} = \mathbf{x_1(n)} * \mathbf{h_1(n)} = [\mathbf{y_1}^1\ \mathbf{y_2}^1\ \dots\ \mathbf{y_N}^1]$$
$$\mathbf{y_2} = \mathbf{x_2(n)} * \mathbf{h_2(n)} = [\mathbf{y_1}^2\ \mathbf{y_2}^2\ \dots\ \mathbf{y_N}^2]$$
$$\dots$$
$$\mathbf{y_J} = \mathbf{x_J(n)} * \mathbf{h_J(n)} = [\mathbf{y_1}^J\ \mathbf{y_2}^J\ \dots\ \mathbf{y_N}^J]$$

El proceso más complejo es la reconstrucción de la señal de salida de manera que se solape y sume solo lo que sea necesario. Se analiza para el caso de que las tramas de entrada no estén solapadas y para el caso de que haya un solapamiento constante de D muestras entre tramas de entrada sucesivas.

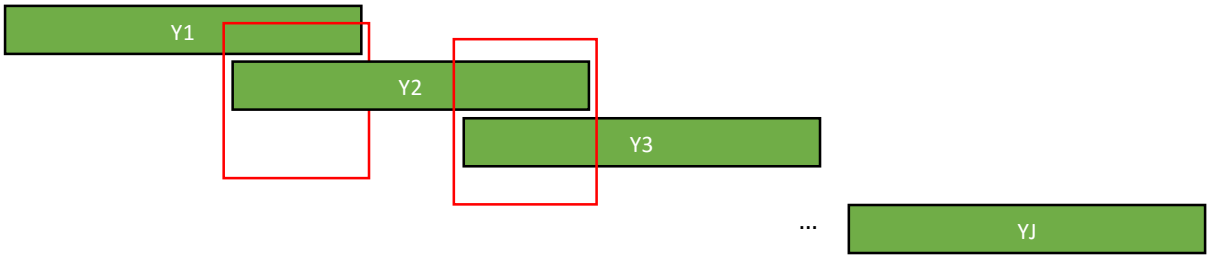
➔ *Solapamiento a la entrada de 0 muestras*

El esquema de bloques para las tramas de entrada es el que sigue, donde cada trama xi(n) tendrá las siguientes muestras de la señal original s(n)

$$\mathbf{x_1(n)} = [\mathbf{s_1}\ \mathbf{s_2}\ \dots\ \mathbf{s_L}]$$
$$\mathbf{x_2(n)} = [\mathbf{s_{M+1}}\ \mathbf{s_{M+2}}\ \dots\ \mathbf{s_{2L}}]$$
$$\dots$$
$$\mathbf{x_J(n)} = [\mathbf{s_{(J-1)\cdot L+1}}\ \dots\ \mathbf{s_I}\ \mathbf{0}\ \mathbf{0}\ \dots \mathbf{0}]$$

donde n = 1, 2, ..., L.

Las señales de salida van a estar solapadas debido a que las tramas van a tener una longitud N = L + M – 1 mayor que la longitud de trama de entrada L debido a las propiedades de la operación de convolución. El esquema de bloques para las tramas de salida es:



Todas las tramas de salida yi(n) tienen longitud N. Cuánto más porcentaje de solapamiento haya, más se solaparan las señales y más sumas habrá que realizar. El número de muestras de señal A que se solapan a la salida es, en el caso de que el solapamiento a la entrada sea de D = 0 muestras (solo se solapan M – 2 muestras a la salida):

$$A(D = 0) = N - (L + 1) = (L + M - 1) - (L + 1) = M - 2$$

Es decir la longitud del filtro BRIR hi(n) menos dos.

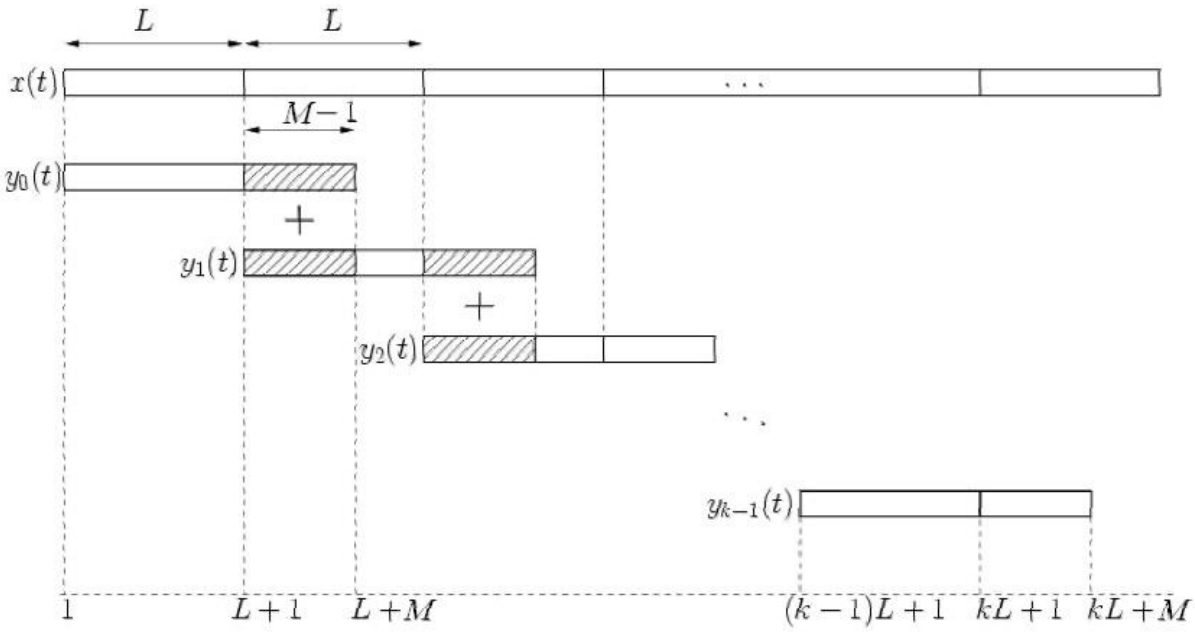
La señal de salida y(n) se compone de las tramas de salida solapadas y sumadas según el solapamiento deseado:

$$\vec{\mathbf{y}} = \mathbf{y(n)} = [\mathbf{y_1(1:L)},\quad (\mathbf{y_1(L+1:N)} + \mathbf{y_2(1:M-2)}),\quad \mathbf{y_2(M:L)},\quad \dots]$$

Por consiguiente, la longitud de solape en muestras es N – (L + 1) = M – 2. Esto es fácil de demostrar: ya que N = L + M – 1 entonces N – L – 1 = (L + M – 1) – L – 1 = M – 2. Se supone que L > M.

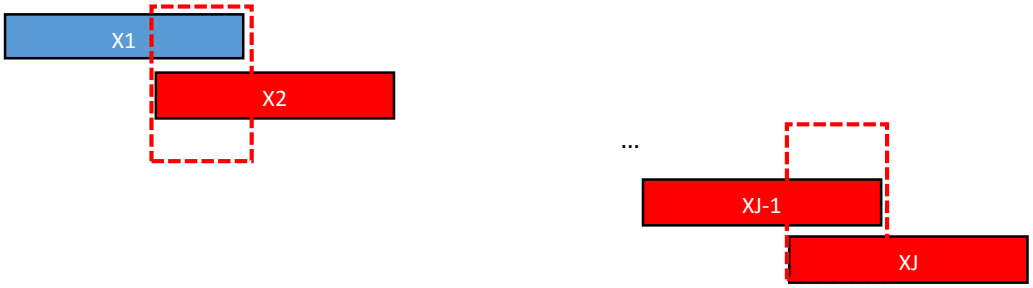
Los índices (ni) = (n1, n2, …, nj) serán calculados dependiendo del solapamiento D, la longitud del filtro BRIR M y la longitud de tramas de entrada L.

La figura siguiente ilustra todo el algoritmo OLA para el caso de tiempo continuo y de un solo filtro h(n). La señal continua de entrada x(t) es filtrada por bloques de L muestras mediante un solo filtro h(t) de M muestras. Se obtienen tramas de salida yk(t) de longitud N = L + M – 1, que se solapan entre si de la manera que se muestra.



➔ *Solapamiento a la entrada de D muestras (0 < D < L)*

Ahora el esquema de bloques para las tramas de entrada es el siguiente:



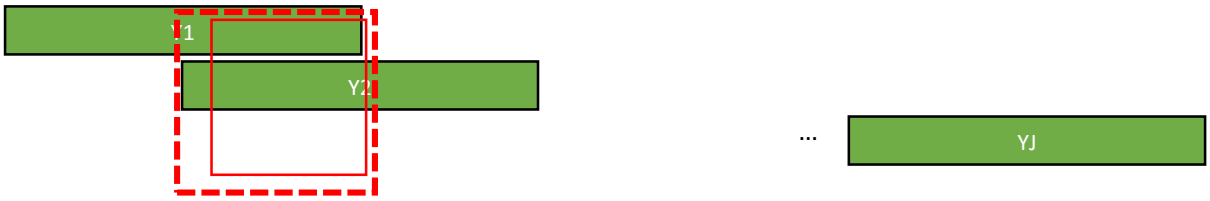
En este caso es necesario sumar el número de muestras de que se solapan entre tramas de entrada (D) al número de muestras solapadas en las tramas de salida. El problema es que ahora los sitios en los que hay solape han cambiado. El número de tramas J será mayor o igual en estas nuevas condiciones.

Volvamos a examinar el esquema de bloques. Ahora, cada trama xi(n) contendrá las siguientes muestras de la señal original s(n)

$$\mathbf{x_1(n)} = [\mathbf{s_1\ s_2} \dots \mathbf{s_L}]$$
$$\mathbf{x_2(n)} = [\mathbf{s_{L+1-D}\ s_{L-D+2}} \dots \mathbf{s_{2L+1-D}}]$$
$$\dots$$
$$\mathbf{x_J(n)} = [\mathbf{s_{(J-1)L-D}} \dots \mathbf{s_l\ 0\ 0} \dots \mathbf{0}]$$

donde n = 1, 2, …, L. Muy probablemente x2(n) tenga muestras de señal que ya están en x1(n), x3(n) tenga muestras que ya aparecen en x2(n), y así sucesivamente. Para la reconstrucción en la salida, ahora las señales se han solapado al inicio, El solape a la entrada, de D muestras más, hará que las tramas yi(n) tengan que ser posicionadas empezando en $n = i \cdot (L - D)$.

El esquema de bloques para las tramas de salida es ahora:



El rectángulo trazado en línea continua muestra el solapamiento normal debido a la convolución. El rectángulo en línea discontinua presenta la nueva situación, donde las tramas de salida yi(n) se tienen que desplazar una cantidad D hacia el inicio, creando aún mas solapamiento

El número de muestras A que se solapan a la salida es ahora:

$$\mathbf{A(D > 0) = (L + M - 1) - (L - D) = M + D - 1}$$

3.6.5. Coste computacional del método

El coste de la convolución se puede asociar al número de multiplicaciones complejas involucradas. El mayor coste es debido a la FFT, que para un algoritmo radix-2 aplicado a una señal de longitud N consume $c = (N/2) \cdot \log2(N)$ multiplicaciones complejas. Para el método de solapamiento y suma el número de multiplicaciones complejas el coste c(N) es:

$$\mathbf{c(N) = \left\lceil \frac{T}{N - M - 1} \right\rceil \cdot N \cdot (log2(N) + 1)}$$

donde T es el periodo de la trama de señal si(t) (si se puede calcular o estimar). El número c tiene en cuenta la FFT, la multiplicación por el filtro y la IFFT.
El mejor valor para N se puede estimar buscando el mínimo

$$c(N) = c(2^m)$$

con

$$\log2(M) < m < \log2(T).$$

Siendo N potencia de dos, las FFTs del método de solapamiento y suma se computan eficientemente. Una vez computado el mejor valor de N, se descubre que el particionado óptimo de s(n) se da para una longitud de trama L igual a:

$$\mathbf{L = N - M + 1.}$$

Con todo, se demuestra que el orden del método de solapamiento y suma es de

$$\mathbf{O(T \cdot log2(N)).}$$

No obstante, hay que saber que esta medida solo tiene en cuenta el coste de las multiplicaciones complejas, sin contar con las otras operaciones involucradas en el algoritmo.

3.6.6. Conclusión. Convolución variante en el tiempo

La situación es pues que existe una fuente sonora que se mueve y cuya posición y forma de onda son $\sigma(t) = \vec{p}(t)$ y $s(t)$, respectivamente. La señal que se observa en la posición $\vec{q}(t = t_s)$ puede derivarse de la ecuación de ondas para fuentes en movimiento como [19]:

$$\mathbf{y(t)} = \int \mathbf{s(t_s)}\, \mathbf{h}\big(\mathbf{t - t_s}, \vec{\mathbf{p}}(\mathbf{t_s})\big)\, \mathbf{dt_s}$$

Esta operación se suele *denominar convolución variante en el tiempo* para sistemas variantes en el tiempo.

En tiempo discreto esta ecuación se puede aproximar como:

$$\boxed{\mathbf{y(n)} = \sum_{\mathbf{n_s}=\mathbf{0}}^{\infty} \mathbf{s(n_s)}\, \mathbf{h}\big(\mathbf{n - n_s}, \vec{\mathbf{p}}(\mathbf{n_s})\big)}$$

La función $h\big(n\,, \vec{p}(n_s)\big)$ es la respuesta al impulso de la fuente en $\vec{p}(n_s)$ al punto de recepción. Esta ecuación implica que la señal de salida $y(n)$ puede obtenerse si la señal de la fuente $s(t)$ y las respuestas al impulso $h\big(n\,, \vec{p}(n_s)\big)$ en todas y cada una de las posiciones pueden obtenerse, incluso si la fuente se mueve. Notar que:

la frecuencia de muestreo debe ser mayor que dos veces la frecuencia máxima de la señal para tener en cuenta el efecto Doppler debido al movimiento de la fuente (i.e. mayor o igual a 44.1 kHz para audio con una banda de frecuencias 0.0-20.0 kHz).

La ecuación anterior se puede expresar en forma matricial como dos vectores y una matriz:

$$\vec{\mathbf{y}} = \mathbf{H} \cdot \vec{\mathbf{s}}$$
$$\vec{\mathbf{s}} = \big(\mathbf{s(1)}, \mathbf{s(2)}, \dots, \mathbf{s(Ls)}\big)$$
$$\vec{\mathbf{y}} = \big(\mathbf{y(1)}, \mathbf{y(2)}, \dots, \mathbf{y(Ls + Lh - 1)}\big)$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}(\mathbf{1}, \vec{\mathbf{p}}(\mathbf{1})) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{h}(\mathbf{2}, \vec{\mathbf{p}}(\mathbf{1})) & \mathbf{h}(\mathbf{1}, \vec{\mathbf{p}}(\mathbf{2})) & \ddots & \vdots \\ \vdots & \mathbf{h}\big(\mathbf{2}, \vec{\mathbf{p}}(\mathbf{2})\big) & \ddots & \mathbf{0} \\ \mathbf{h}(\mathbf{Lh}, \vec{\mathbf{p}}(\mathbf{1})) & \vdots & & \mathbf{h}\big(\mathbf{1}, \vec{\mathbf{p}}(\mathbf{Ls})\big) \\ \mathbf{0} & \mathbf{h}\big(\mathbf{Lh}, \vec{\mathbf{p}}(\mathbf{2})\big) & & \mathbf{h}\big(\mathbf{2}, \vec{\mathbf{p}}(\mathbf{Ls})\big) \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{h}\big(\mathbf{Lh}, \vec{\mathbf{p}}(\mathbf{Ls})\big) \end{bmatrix}$$

Aquí, $\vec{\mathbf{s}}$ denota el vector de la señal fuente, $\vec{\mathbf{y}}$ es el vector de señal observado y H es una matriz de convolución variante en el tiempo que es una extensión de la matriz de convolucion invariante en el tiempo [19]. Si se define el patrón de movimiento de la fuente como el vector posición $\vec{\mathbf{p}}(\vec{\mathbf{n}})$ que depende del vector de tiempo discreto $\vec{\mathbf{n}}$, H queda determinada por el patrón $\vec{\mathbf{p}}(\vec{\mathbf{n}})$ y el punto de observación $\vec{\mathbf{q}}$ **.** Aquí, se asume que la señal tiene duración Ls y todas las respuestas al impulso tienen la misma longitud Lh.

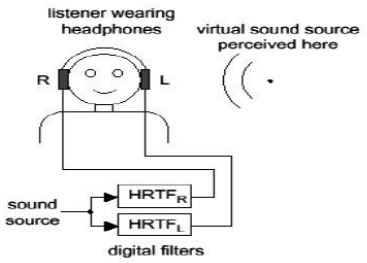
3.7. “Switching” BRIRs

3.7.1. Introducción. Concepto de imagen sonora en movimiento.

Para conseguir que se perciba que una fuente sonora se mueve virtualmente se puede ir cambiando de BRIR o de HRTF, que incluyen las características de transferencia espacial. Debido a la naturaleza de estas funciones de transferencia, ocurren discontinuidades de onda en el momento de cambiar, lo que degrada la calidad del audio. Las características de la discontinuidad dependerán del esquema usado para convolucionar. La discontinuidad de onda se refleja como un *click* o ruido impulsivo en el audio y ocurre cada vez que se procesa y convoluciona una trama del audio de entrada [20].
El método más simple usando altavoces es mover físicamente el altavoz. No obstante, esto es poco práctico. En general, se suelen usar tres métodos cuando se procesa el audio para auralizarlo: el cambio o *switching* simple, el método de solapamiento y suma (OLA) y el método de fade-in-fade-out.

Los tres métodos han sido implementados y probados sucesivamente conforme la complejidad y calidad del software de auralización iba mejorando. En este apartado se resumen las características principales de cada método.

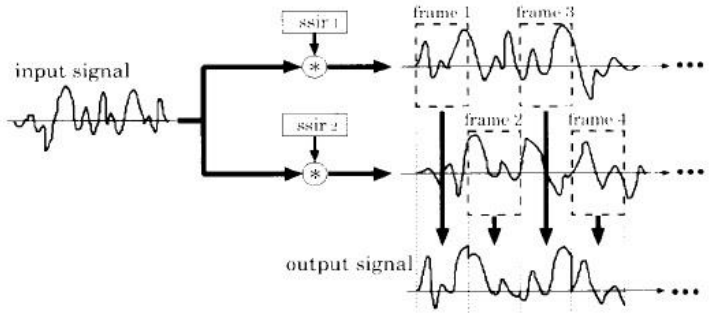
La figura siguiente muestra la relación entre las funciones de transferencia necesarias para realizar la imgen sonora en movimiento. Como se puede apreciar, el audio de entrada se convoluciona con dos funciones transferencia para el oído izquierdo y derecho (L y R) que cambiarán cuando se quiera cambiar la posición percibida de la fuente. En la imagen se ha simplificado el esquema de procesado para aclarar la idea fundamental subyacente.



Las discontinuidades de onda se deben a las diferencias en tiempo y amplitud (ITDs e ILDs) entre las funciones de transferencia BRIR o HRTF consecutivas y aumentan cuando estas diferencias aumentan.

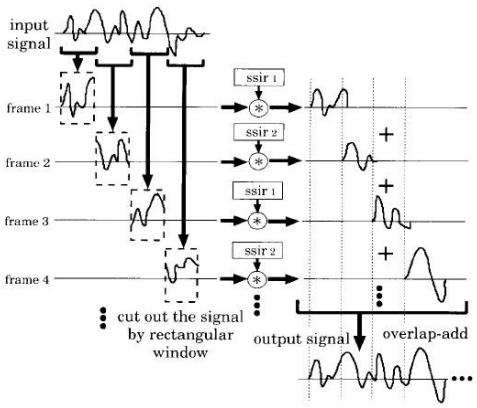
3.7.2. Switching simple

La manera más simple de ir cambiando de funciones de transferencia se muestra en la siguiente figura [20]. La señal de entrada es convolucionada por las BRIRs o funciones de transferencia espaciales SSIR1 y SSIR2 (Spatial Sound Impulse Response), que pueden ser coincidir con las BRIRs o también ser solamente igual a las HRTFs (o a otras funciones de transferencia si se quieren tener en cuenta más o menos factores acústicos y/o psicoacústicos). La sensación de movimiento de la fuente es creada concatenando directamente las tramas. Este método tiene la problemática de generar una discontinuidad de onda en cada punto de concatenación. El esquema completo se muestra a continuación.



3.7.3. Método de solapamiento y suma

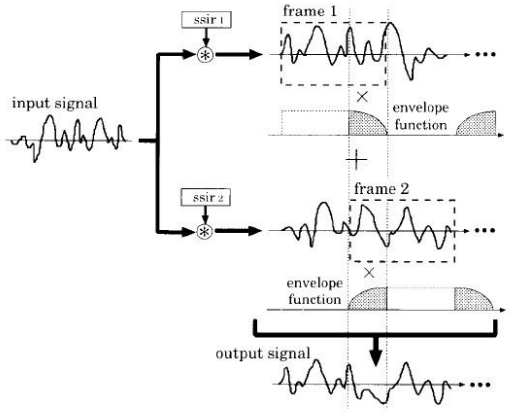
En el método se solapamiento y suma OLA, como se explico en el apartado 3.x.x., la señal de entrada es tomada por bloques o tramas. También se expresa esto diciendo que la señal es envantanada trama a trama. La trama actual es convolucionada con cada SSIR que se suponen diferentes. Las tramas de salida, resultado de la convolucion de la trama de entrada con ambas SSIRs, son solapadas y sumadas sucesivamente para crear el efecto de una imagen sonora en movimiento. Este método sigue creando una discontinuidad de onda en las áreas de solape.



Los clicks debidos a la discontinuidad pueden ser parcialmente suprimidos implementando un segundo filtrado, filtrando cada trama de entrada con una ventana (p.ej. de Hamming modificada). Las tramas de salida son ponderadas otra vez por la ventana y solapadas y sumadas normalmente.
No obstante, los ruidos seguirán presentes en el audio, aunque en menor medida.

3.7.4. Método fade-in-fade-out

En el método fade-in-fade-out, la señal de entrada es enventanada y convolucionada con las SSIRs, de manera que las tramas de salida se solapan. En las áreas de solape, las señales son ponderadas por las funciones f(n) y g(n), que satisfarán $f^2 + g^2 = 1$. Se aplica a continuación el algoritmo OLA, de manera idéntica a como se hace en el método de solapamiento y suma. El esquema de convolución se puede ver en la siguiente figura.



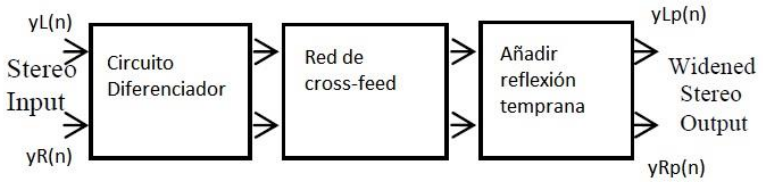
3.8. Post-procesado. Widening estéreo.

3.8.1. Introducción

Los auriculares estereofónicos han sido la herramienta elegida en este proyecto para "visualizar" el audio 3D sintetizado mediante el software desarrollado. En la reproducción de audio mediante auriculares, ya sea el audio del tipo que fuere, el oyente siempre percibirá las imágenes sonoras a lo largo del eje interaural (que cruza los dos oídos). En general, los auriculares consiguen una impresión espacial de sólo unos 60º. La exposición a este efecto no natural por periodos prolongados de tiempo puede ocasionar que el sonido parezca provenir de dentro de la cabeza (in-head localization), además de fatiga auditiva.
Las componentes de baja frecuencia de cualquier tipo de audio, sobre todo música (p.ej. el bajo eléctrico) siempre parecen provenir de dentro de la cabeza. Varios métodos han sido propuestos para externalizar el audio usando características binaurales (ITD, ILD e ICC) y todavía queda mucho por mejorar.

En este proyecto se ha optado por el esquema descrito en [21] que, en resumen, se basa en la relación señal central C a señal lateral S (CSR Centre-to-Side signal Ratio) y usa líneas de retardo apropiadas y parámetros de ganancia para explotar las características binaurales. Además, incorpora un circuito de reflexiones tempranas para generar más espacialidad en el audio estéreo.
La señal central C se define: $C(n) = \frac{1}{\sqrt{2}}(yL(n) + yR(n))$
y la señal lateral S se define: $S(n) = \frac{1}{\sqrt{2}}(yL(n) - yR(n))$
donde n es el número de muestras de la señal de audio, que se supone igual para ambos canales. Ambas señales C y S son perpendiculares entre si y forman un nuevo sistema de coordenadas.

El dotar de profundidad a un archivo de audio (llamado *widening* estéreo) se puede conseguir generando una señal diferencia L – R que se sumará a ambos canales L y R y usar un decorrelador para disminuir la correlación entre los canales L y R con la ayuda de las características binaurales.
Una red de cross-feed para simular el efecto de Crosstalk inherente en los altavoces ayuda a mejorar aún más la sensación de envolvimiento. El Crosstalk presente en la escucha mediante altavoces hace que las frecuencias bajas del audio sean difractadas y las altas atenuadas casi 20dB. El Crosstalk es simulado para auriculares mezclando una versión retardada y filtrada paso-bajo de un canal con el otro canal.



Sistema de Widening estéreo

El sistema implementado consta de tres partes: 1) un circuito diferenciador (con un decorrelador optativo) para sumar una señal diferencia L – R al audio, con el fin de hacer las componentes laterales más prominentes, 2) un circuito de cross-feed que simula el entorno de escucha mediante altavoces usando las ITDs e ILDs y 3) un circuito de reflexiones tempranas que extrae una reflexión temprana dominante y la añade al audio original con diferente retardo para los dos canales. Esto disminuye la ICC pues desfasa las señales L y R.

Hay cuatro metas fundamentales a la hora de implementar un sistema de widening estéreo [21]:

- *decorrelación de los canales L y R*
- *aumentar la SCR y calcular la medida de widening*
- *obtener la mínima coloración espectral debida a las redes de retardo realimentadas positiva o negativamente*
- *crear Crosstalk para una sensación mas realista*

Sin embargo, estos problemas no están muy relacionados entre sí, con lo que existen diferentes soluciones. El coeficiente de correlación puede minimizarse usando una técnica de decorrelación estática que pasa los canales L y R por dos filtros paso-todo diferentes (bien con una característica de fase aleatoria o bien complementaria). Una correlación dinámica se puede conseguir calculando el filtro paso-todo con respuesta en fase aleatoria, para cada trama [21]. Para decorrelar dinámicamente se usan filtros IIR en vez de FIR, pues los coeficientes se pueden actualizar fácilmente variando aleatoriamente las distancias de polos y ceros al circulo unidad.

En el estéreo convencional, la señal central parece provenir del centro de la cabeza del oyente. La señal lateral parece venir de los alrededores en vez de de la fuente virtual directamente.
La SCR es una medida del widening de una señal. Una señal amplia posee un alto SCR y viceversa. La manera de aumentar la SCR y con ello el widening es añadiendo la señal lateral S atenuada y retardada a las componentes L y R del audio.

Por otro lado, ya que las redes de retardo usadas generan coloración espectral, es mejor usar solamente una relfexion temprana para minimizar esta coloración.

Este post-procesado es aplicado al audio binaural sintetizado, pero funciona igual de bien con archivos de música o cualquier otro audio estéreo al que se le quiera dar profundidad.

Existen aplicaciones para dispositivos móviles que también post-procesan en cierta medida nuestros archivos de música. Para aplicaciones iOS, por ejemplo Beautyfier® [22] puede servir como ejemplo ilustrativo de una aplicación de post-procesado de audio.

3.9.2. Aspectos sobre técnicas de decorrelación

En esta sección se quiere presentar una técnica usada actualmente para conseguir un ancho arbitrario percibido de una fuente virtual en sistemas de audio 3D [23]. La discusión se centra en las técnicas que usan la decorrelación como un medio para reducir el *coeficiente de correlacion cruzada interaural (Interaural Cross-Correlation Coefficient IACC),* lo que tiene un efecto directo sobre la extensión percibida de la fuente virtual.

La *extensión espacial o ancho (width)* de una fuente sonora se define como el tamaño espacial percibido de la fuente. El ancho de una fuente sonora es un fenómeno natural: la costa de una playa, el viento que sopla entre los arboles de un bosque, una cascada, etc… son todos ejemplos de fuentes con una extensión espacial considerable. Es por esto que la extensión espacial de una fuente también es una característica perceptual importante.

Un esquema de decorrelación dependiente de la frecuencia puede emplearse para crear el efecto de una fuente sonora que se divide en frecuencias o bandas de frecuencia y cada frecuencia o frecuencias se posiciona virtualmente en el espacio en sitios distintos. Esto es especialmente útil para la producción de música a nivel profesional.
En los auditorios y salas de concierto, un valor del IACC bajo mejora la sensacion de amplitud de la sala y el ancho de la fuente [24].

Otras investigaciones se han basado en el ancho percibido de ruido reproducido mediante altavoces y auriculares. Se ha concluido que la cantidad de correlacion entre los dos canales tiene un impacto drástico en la percepción de extensión espacial [25]. Cuando se presentan señales de ruido no correladas mediante altavoces, el ruido parece llenar todo el espacio entre los altavoces. Mientras que señales correladas producen una fuente virtual estrecha justo entre los altavoces.

La decorrelación tiene cinco efectos positivos notables en la percepción de la imgen sonora:

· la coloración del timbre y el efecto de “combing” (filtro peine) asociado a las interferencias constructivas o destructivas de multiples señales retardadas son, perceptualmente, eliminados,

· los canales decorrelados producen campos sonoros (más) difusos,

· los canales decorrelados producen (más) externalización,

· la posición percibida del campo sonoro no cambia si hay cambios en la posición del oyente relativa a los altavoces y

· el efecto de precedencia o efecto Haas, que colapsa la imagen sonora al altavoz más cercano, disminuye, permitiendo presentar exactamente la misma señal usando múltiples altavoces.

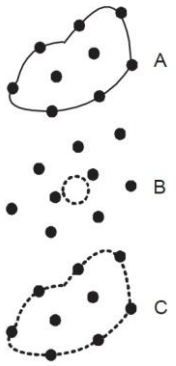
En este punto, es importante recordar la función de correlación cruzada o simplemente de correlación entre dos señales discretas yL(n) e yR(n) (se suponen de igual duración l):

$$\mathbf{R_{LR}(\tau)} = \frac{\mathbf{1}}{\mathbf{l}} \sum_{\mathbf{n=-\infty}}^{\mathbf{+\infty}} \mathbf{yL(n)} \cdot \mathbf{yR(n-\tau)}\,, \qquad \boldsymbol{\tau} = \mathbf{0}, \pm\mathbf{1}, \pm\mathbf{2}, \ldots$$

con $n = 0, 1, 2, \ldots, (l - 1)$. La medida de correlación (también llamada coeficiente de correlación) se expresa como un número entre -1 y $+1$ y es el valor de la función de correlación que mayor amplitud (en valor absoluto) tenga. Si las señales son muy similares, la medida de correlación será de ± 1 (dependiendo de si están en fase o en contrafase) y si las señales son completamente diferentes la medida de correlación será de 0. Ya que las señales en este caso son las señales para el oído izquierdo y derecho, la función se denomina *función de correlación cruzada interaural.*
El coeficiente IACC se define como el máximo del valor absoluto de la función de correlación cruzada interaural normalizada:

$$\mathbf{IACC(\tau)} = \frac{\sum_{\mathbf{n=-\infty}}^{\mathbf{+\infty}} \mathbf{yL(n)} \cdot \mathbf{yR(n-\tau)}}{\sqrt{\sum_{\mathbf{n=-\infty}}^{\mathbf{+\infty}} \mathbf{yL^2(n)} \cdot \sum_{\mathbf{n=-\infty}}^{\mathbf{+\infty}} \mathbf{yR^2(n)}}}$$

Un valor de IACC cercano a 1 producirá una fuente virtual "estrecha" o densa.
El coeficiente IACC es un parámetro ampliamente utilizado en Acústica para determinar la amplitud y envolvimiento percibido de sauditorios y salas [Auralization], ["Spatial Hearing", Blauert] ["The nature and technology of acoustic space", Tohyama et al]. Se ha demostrado que el sistema binaural humano es capaz de calcular coeficientes IACC por bandas de frecuencia y que es sensible a fluctuaciones temporales del IACC ["Spatial Hearing", Blauert].



Una técnica extendida para conseguir la expansión de la percepción de tamaño en las fuentes virtuales se basa en el hecho de que una fuente sonora ancha real puede descomponerse en múltiples fuentes puntuales posicionadas en sitios distintos (en la figura, A). No obstante, para que este efecto se de las señales emitidas por las fuentes puntuales no pueden estar correladas entre sí. Esto se debe a que, si la correlacion fuera alta, el sistema binaural humano percibiría todo como un solo evento auditivo en el centro de gravedad de las fuentes ["Spatial Hearing", Blauert] (en la figura, B). El centro de gravedad depende de las posiciones e intensidades de las fuentes puntuales. En el caso contrario, en el que las señales generadas por las fuentes puntuales están poco correladas, el sistema binaural humano percibe las fuentes como eventos auditivos distintos. Esto resulta en la percepción de ese conjunto de fuentes puntuales como una sola fuente virtual extensa en el espacio 3D (en la figura, C).

No obstante, en la práctica puede no ser posible distinguir cada fuente puntual por separado si las fuentes puntuales están posicionadas de forma muy compacta y densa, pues el sistema auditivo produce la impresión final de una fuente sonora única, espacialmente grande.

A continuación se presentar varias técnicas para obtener señales decorreladas de una señal monoaural.

Decorrelación en todo el espectro

La manera más fácil de introducir correlación es mediante convolución. Para producir un par de señales de salida con una medida de correlación deseada, la señal de entrada puede ser convolucionada con dos señales similares que estén correladas entre sí por la medida de correlación deseada.

Otra manera de obtener señales decorreladas es introducir un pequeño retardo entre ellas. Este método, aunque simple, solo permite producir un número limitado de señales decorreladas ya que el retardo máximo está restringido por la percepción de un eco: este retardo no debe ser superior a 40 ms.
Para altavoces esta técnica no funciona bien debido a efectos de "combing" causados por los retardos.

Otra manera común de decorrelación es filtrando la señal de entrada con filtros paso-todo con respuesta en fase aleatoria, ruidosa ["The decorrelation of audio signals and its impact on spatial imagery", Kendall, 1995]. Debido a la inestabilidad a las variaciones de fase del oído humano y por preservar la amplitud del espectro de la señal por usar filtros paso-todo, las señales obtenidas son perceptualmente iguales pero estadísticamente *ortogonales* [26].
La decorrelación de filtros paso-todo se puede implementar en arquitecturas FIR o IIR.

Este método permite obtener solo un número muy limitado de señales decorreladas, ya que, debido a la longitud finita de los filtros, habrá pronto un valor de correlación alto entre un par de señales. Por ello, las respuestas en fase de los filtros también tienen que ser ortogonales y ser obtenidas por un proceso de selección cuidadoso. Con esta técnica, en [23] son capaces de obtener solo cinco o seis señales decorreladas con una longitud de filtro de 100 polos y ceros.

Decorrelación dinámica

La decorrelación dinámica o variable en el tiempo se realiza usando filtros paso-todo variantes con el tiempo. La ventaja sobre la decorrelación estática es que se pueden obtener mas señales no correladas. Este tipo de decorrelación introducirá niveles de correlación también dinámicos dependiendo de la ortogonalidad de las respuestas en fase de los filtros; pero si estas variaciones son suficientemente rápidas y no pueden ser percibidas y seguidas por el oído, el valor medio de correlación percibido será bajo.

La decorrelación dinámica se consigue calculando una nueva respuesta en fase aleatoria para cada nueva trama. Las estructuras FIR o IIR en configuración "lattice" son apropiadas para esta tarea debido a su estabilidad (se suelen producir inestabilidades cuando se actualizan frecuentemente los coeficientes de los filtros).

En [27] también se dice que la decorrelación dinámica crea micro-variaciones que simulan las fluctuaciones temporales causadas por el aire en movimiento, lo que mejoraría el efecto 3D en caso de ser implementado en el software. No obstante la decorrelación dinámica puede tener también un efecto de distracción e incluso crear fatiga debido a cambios apreciables de las posiciones de las fuentes en una escena grabada con micrófonos. Esto puede deberse a las diferencias de fase entre fuentes puntuales que produzcan una ITD.

Decorrelación en sub-bandas

Esta técnica permite alterar la decorrelacion de manera diferente para cada banda de frecuencias. Mediante esta técnica se puede conseguir un conjunto de señales para las que, p.ej., las componentes de baja frecuencia no estén correladas mientras que las de las altas frecuencias se dejan correladas.

Usando el método de fuentes puntuales discutido, esta técnica produce efectos muy interesantes y aplicable donode la extensión espacial virtual de la fuente varía con la frecuencia.
Esto permite dividir la señal en distintas bandas de frecuencia, cada banda con una posición virtual distinta y con ancho distinto. Este efecto se puede llamar *efecto de descomposición espacial de Fourier.*

Para la decorrelacion en sub-bandas, como primer paso la señal de entrada es dividida en bandas de frecuencia mediante un banco de filtros usando filtros paso-bajo, paso-banda y paso-alto de órdenes altos. Cada sub-banda se decorrela usando cualquiera de las técnicas anteriormente expuestas (u otras). A continuación se usan cross-faders para controlar la cantidad de correlación en cada sub-banda. Esto se hace realimentando algo de la señal común en cada señal decorrelada (para total decorrelacion no se realimenta nada). Se prefiere una técnica de cross-fading en potencia constante para que no haya cambios en el nivel de señal si se cambia el factor de los cross-faders.
Finalmente, las sub-bandas se suman para formar el conjunto de señales parcialmente decorreladas.

Se puede combinar la decorrelacion dinámica con la decorrelacion en sub-bandas para obtener niveles de correlacion variantes en tiempo y frecuencia.

Decorrelación variable en tiempo

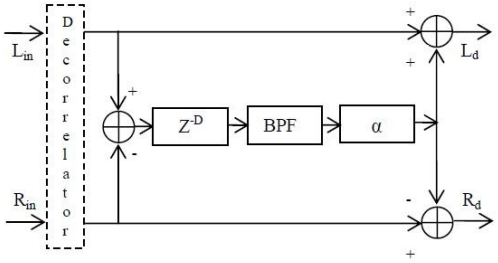
La correlacion variable con el tiempo se obtiene realimentando periódicamente la señal original con las señales decorreladas. Para decorrelacion hasta los 10 kHz con IACCs variando entre 0 y 1, esta técnica crea una fuente virtual con extensión espacial variando constantemente. Para señales a mas de 10 kHz, el efecto deja de funcionar debido a la imposibilidad del sistema auditivo de derivar un IACC a esas frecuencias.

Hay que notar que esta sección pretende ser solamente una introducción a la decorrelación; lo expuesto aquí no ha podido ser implementado en el proyecto pues hubiera aumentado la complejidad considerablemente. Además, el sistema de widening estéreo implementado funciona razonablemente bien sin el decorrelador.
Esto no quiere decir, bajo ningún concepto, que este sub-sistema no sea necesario: para usar el sistema de widening como elemento separado para procesar archivos de música, sería necesario incluir el sistema de decorrelación ya que aumentaría considerablemente las posibilidades del audio resultante. También para audio 3D, una decorrelación por tramas es muy beneficiosa. El decorrelador es siempre útil en material de audio donde los canales estéreo están altamente correlados.

3.9.3. Circuito diferenciador

El diagrama de bloques del primer subsistema para el sistema de widening se muestra en la figura siguiente. Sean yL(n) e yR(n) las señales binaurales o estéreo de entrada. Las señales pasan a través del elemento decorrelador (opcional), explicado anteriormente

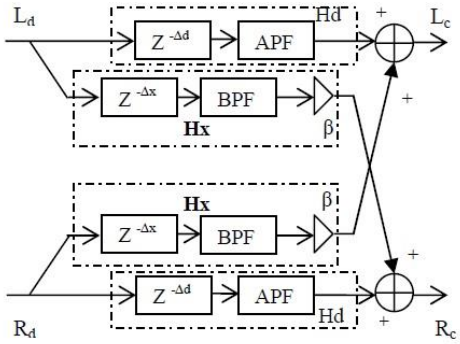
El "widening" o apertura o espaciamiento del sonido se consigue realzando la información direccional en cada canal. Se calcula la señal lateral S y se retarda en un rango de 10 a 25 ms (15 ms fue el valor elegido por compromiso ya que un valor superior a 25 ms podría incorporar eco y coloraciones). La señal lateral (señal resta) es filtrada a continuación por un filtro paso banda con frecuencias de corte 250 Hz y 12 kHz. A la salida, la señal se atenúa mediante un parámetro de profunidad de widening (α, $\alpha = 0.5$), y se suma al canal L. La misma señal se resta al canal R. Esto aumenta el SCR, mejorando la fidelidad espacial de la señal.



3.9.4. Circuito de cross-feed

Es un fenómeno conocido que cuando se está escuchando audio mediante altavoces, algo del canal izquierdo llega al oído derecho y viceversa (el ya discutido Crosstalk). Para crear un efecto similar en los auriculares, se puede usar una red de cross-feed para crear Crosstalk; es justo lo contrario a la cancelación de Crosstalk.
Hay que volver a notar que las frecuencias bajas se difractan al pasar detrás de la cabeza y los sonidos por encima de 1.5 kHz son bloqueados cuando la longitud de onda es pequeña en comparación con las dimensiones de la cabeza (efecto de apantallamiento del cráneo). Las ITDs e ILDs son características importantes por debajo de 1.5 kHz y el efecto de apantallamiento es el predominante por encima de 1.5 kHz.

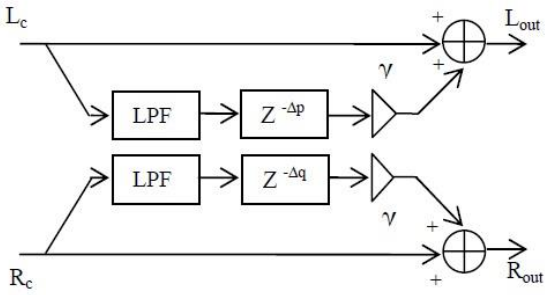
La figura siguiente muestra el subsistema para crear diafonía. El retardo se elige $\Delta d = 0.3$ ms, que es lo que tarda aproximadamente en llegar el sonido de un altavoz al oído ipsilateral en situaciones de escucha típicas (1m de distancia al altavoz). En muestras, $fs \cdot \Delta d = 44100 \cdot 0.0003 \cong 15$ muestras.
El retardo introducido en el camino de diafonía es $\Delta x = \Delta d + ITD$, que definiendo una $ITD = 0.2$ ms (ITC constante e igual a 9 muestras), queda establecido en unas 24 muestras. El camino directo se pasa por un filtro paso todo. La señal de diafonía se pasa pasa por un filtro de Crosstalk Hx que no es mas que un filtro paso-bajo con frecuencia de corte 2 kHz. A la salida del filtro se atenua la señal resultante con un parámetro de ganancia de cross-feed (β, $\beta = 0.3 \ldots 0.7$) y se suma al canal contrario para cada canal. Como se intuye, β controla la ILD entre el camino directo y el de diafonía.



3.9.5. Circuito de reflexiones tempranas

El efecto de localizacion externa (exterior a la cabeza) se puede simular añadiendo la cantidad adecuada de sonido indirecto al sonido directo. El subistema de reflexiones tempranas simula la adicion de una única reflexión temprana dominante usando una red de retardo realimentada positivamente. Las reflexiones tempranas proveen de mucha de la ifnoramcion espacial, directividades reconocibles asi como los distintos tiempos de llegada en un entorno o sala. El filtro paso bajo se añade ya que las paredes de un entorno tienden a absorber más frecuencias altas que bajas. Aquí, cada canal se añade a una versión retardada y atenuada mediante otro parámetro ($\gamma, \gamma = 0.5$) de ese mismo canal, para reducir la coherencia inter-canal ICC pero sin degradar la calidad de la señal.. Para minimizar la ICC se usan dos retardos ligeramente diferentes para cada canal L y R. Las restricciones en los retardos son: $\Delta p - \Delta q = 3$ ms y que ambos estén comprendidos entre 5 ms y 10 ms.

La reducción de la ICC mejora la "legibilidad" espacial y la inmersión en el audio.



3.9.6. Resultados

Se realizaron tests subjetivos y objetivos con diferentes sujetos ajenos y no ajenos al ámbito del procesado de audio. En los tests subjetivos, el oyente tenía que decidir la mejora percibida en la espacialidad del audio y también definir la sensación de externalización percibida. Varios archivos de

audio binaural fueron presentados a los oyentes usando auriculares y se les pidió realizar una clasificiación de la mejora según la misma escala que la usada en [21]:

5 Widening y sensación de externalización
4 Widening
3 mejor que el original
2 no se aprecia diferencia
1 mala calidad

Los resultados obtenidos por el autor de este Proyecto Fin de Carrera fueron que la mayoría de sujetos clasificaban el audio entre el 3 y 5.

En la publicación en la que se basa este apartado, [21], también se obtuvieron estos resultados, con lo que el sistema desarrollado es satisfactorio para los términos de este proyecto.

BIBLIOGRAFÍA

[1] "Fundamentals of binaural technology", Moller, H. Applied Acoustics, 36, pp 172-218, 1992.

[2] "A model for Room Acoustics", McGovern, S. 2003. http://www.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator/content/rir.m

[3] "Fast Convolution", McGovern, S. 2003. http://www.mathworks.com/matlabcentral/fileexchange/5110-fast-convolution/content/fconv.m

[4] "Efficient Convolution without Input-Output Delay", Gardner, W. AES, 1995.

[5] "Characterising studio monitor loudspeakers for auralization" Supper, B. Paper 7994. AES 128th Convention, 2009.

[6] "HRTF interpolation in azimuth, elevation and distance", Gamper, Aalto University Finland, 2010.

[7] "Estructuras dinámicas de Datos en lenguaje C", Camarero, I., 2005 http://progrmandoenc.webcindario.com

[8] "Data Structures and Algorithm Analysis in C", Weiss, M. A. Florida University, 1990.

[9] "Adaptive 3D Sound Systems", chap. 2, Garas, J. Kluwer Academic Publishers, 2000.

[10] "RF and Microwave Transmitter Design", Grebennikov, A. Wiley, 2011.

[11] "Individual headphone compensation for binaural synthesis", Brinkmann, F. TU Berlin, 2011. https://www2.ak.tu-berlin.de/~akgroup/ak_pub/abschlussarbeiten/2011/Brinkmann_MagA.pdf

[12] "Handbook of Signal Processing in Acoustics", chap. 33, Vorlander, M., Havelock, D. Springer, 2008.

[13] "Characterising studio monitor loudspeakers for auralization", Supper. AES, 2010.

[14] " A Study of Low-Frequency Near- and Far-field loudspeaker behaviour", Vanderkooy *et al*. AES, 2009.

[15] "Can one perform quasi-anechoic loudspeaker measurements in normal rooms?", Vanderkooy *et al*. AES, 2008.

[16] "The Art of Sound Reproduction", Watkinson. Focal, 1998.

[17] "Tratamiento de señales en tiempo discreto", Oppenheim, A. V., Schafer, R. W. Pearson, 2000.

[18] "Señales y Sistemas", Oppenheim, A. V., Wilsky, A. S. Prentice Hall, 1996.

[19] "Inverse Filtering of Rooms Acoustics", Miyoshi, M., Kaneda, Y. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 36, IEEE, 1988.

[20] "A study on switching of the transfer functions focusing on sound quality", Kudo, A., Hokari, H. and Shimada, S. Acoustical Society of Japan, Acoust. Sci. & Tech. 26. 3. ASJ, 2005.

[21] “Stereo widening system using binaural cues for headphones,” Basha, S. M. A, Gupta, A. and Sharma, A. Proc. of International Conference on Signal Processing and Communication Systems, Australia, pp. 378-382, 2007.

[22] “Beautifyer™”, Auro Technologies NV. 2012.
Disponible en la App Store de Apple®:
https://itunes.apple.com/us/app/beautifyer/id551335372?mt=8

[23] “Decorrelation techniques for the rendering of apparent source width in 3D audio displays”, Potard y Burnett, DAFX, 2004.

[24] “Concert and opera halls: how they sound”, Beranek. AES, 1996.

[25] “The decorrelation of audio signals and its impact on spatial imagery”, Kendall. JSTOR, 1995.

[26] “Principios de probabilidad, variables aleatorias y señales aleatorias”, Peebles, P. Z. Jr. McGraw-Hill, 2001.

4 RESULTADOS, CONCLUSIONES Y TRABAJO FUTURO

4.1. Resultados

El proyecto desarrollado en Matlab se compone de los siguientes archivos:
auralization.m -
fconv.m -
get_angles.m -
get_angles2.m -
get_HRTF.m -
get_HRTF_d.m -
HRTF_interpol_in_3D.m -
load_audioinput.m -
OLA.m -
plotting.m -
post_proc.m (hace uso del Signal Toolbox) -
read_cal_file.m -
rir.m -
room_parameters.m -
room_parameters0.m -
tonegen.m -
traj_computing.m -
X_Talker.m -

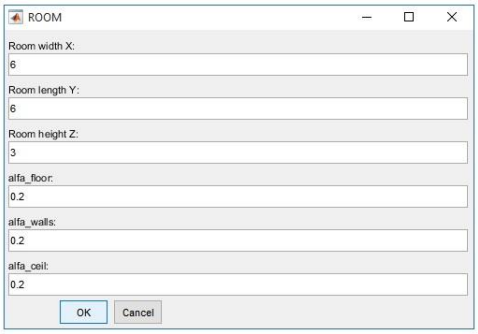
y las siguientes bases de datos:
· KEMAR Dummy Head HRTF Database Media Lab MIT
· Distance-dependent HRTF Database PKU Peking University
· Audio inputs

Además, se incluye una carpeta para realizar modelos paramétricos de HRTFs y convertir audio binaural a sonido 5.1 Surround, con los siguientes archivos:
central.m -
delay.m -
dominant_51Gains.m -
DOWNMIX_3a2.m -
esfera.m -
genera_HRTF.m -
head_shadowing.m -
pinna_reflect.m -
shelving.m -
shoulder_torso.m –
simpleHRIR.m -
suma_vect.m -
to_51Surround.m -

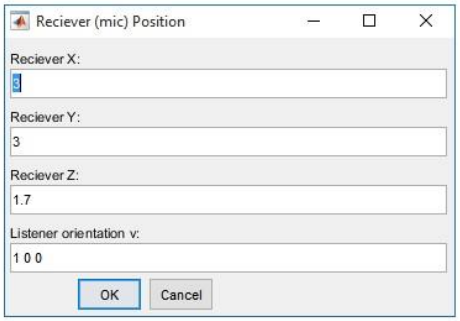
4.2. Pruebas y conclusiones

Una vez se consigue implementar el software y que funcione correctamente, se procede a probarlo. Para ello, se incializa el software de auralización. El software pide al usuario una serie de datos que necesita para procesar el audio.

El programa comienza abriendo una ventana para definir las dimensiones (x, y, z) de la sala, así como los coeficientes de absorción de suelo, paredes y techo que luego se usarán ponderados para calcular parámetros acústicos así como para calcular las RIRs.

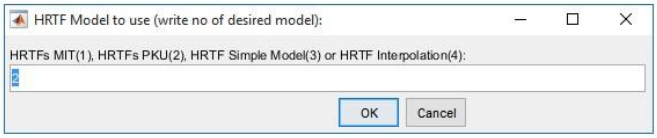


Al pulsar OK se cierra la ventana y se abre otra para determinar las coordenadas (x, y, z) del oyente.

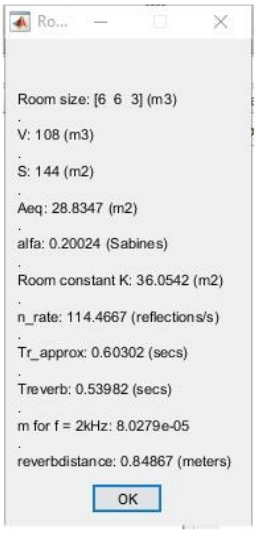


Al pulsar OK se abre una tercera ventana para determinar qué tipo de HRTFs se desean usar. Se han implementado cuatro opciones:

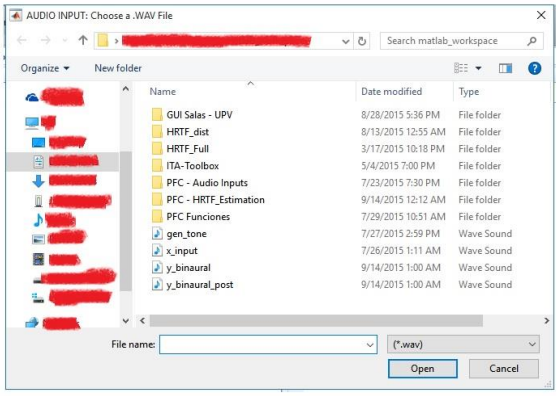
- usar la base de datos de HRTFs del MIT Media Lab
- usar la base de datos de HRTFs de la PKU dependientes de la distancia
- usar un modelo paramétrico simple de HRTFs
- usar la base de datos de HRTFs de la PKU usando interpolación (mejores resultados)



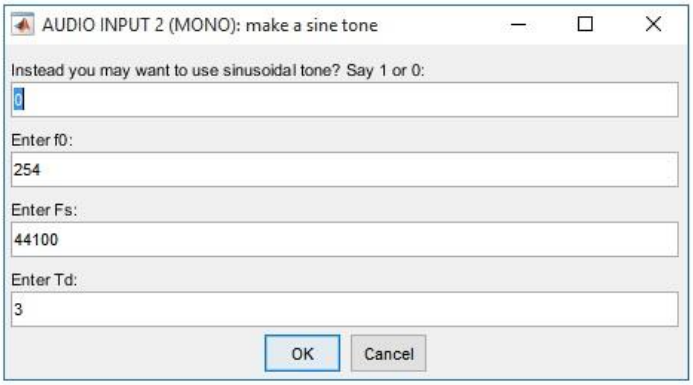
Estas tres ventanas provienen de la funcion room_parameters0.m. Esta función, una vez recopilados los datos necesarios, llama a la funcion room_parameters.m para mostrar un resumen de los parámetros acústicos de la sala que se han estimado en funcion de los datos introducidos. Esta ventana permanecerá abierta.



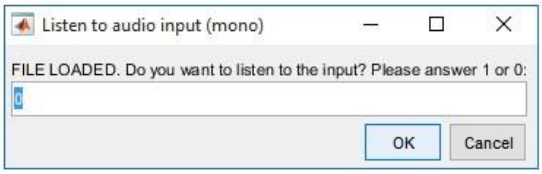
A continuación se abre una ventana, llamada por la función load_audioinput.m, para adquirir el audio a auralizar.



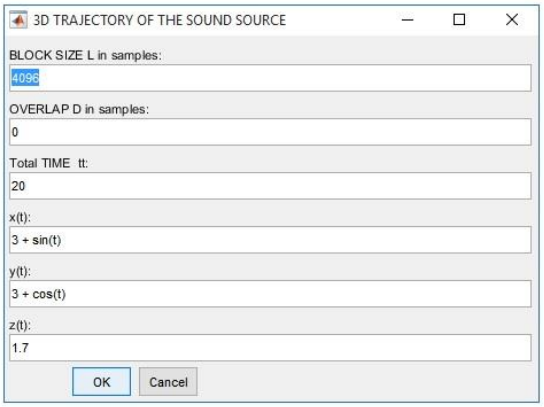
El software dará a continuación de usar un tono sinusoidal definido por el usuario, en vez de un archivo de audio. Una nueva ventana se abre para tal fin:



Luego se abre otra ventana que permite al usuario escuchar lo que ha cargado en el software, antes de que éste lo procese.



La última ventana que se abre es generada por la función traj_computing.m. Esta funcion recoge los parámetros necesarios para definir la trayectoria y realizar la convolución. En concreto, los parámetros que el programa necesita son el tamaño de bloque para la convolucion por tramas (L), el solapamiento en muestras (D) en caso de que en la convolución se quieran solapar las tramas de entrada, el tiempo de la trayectoria (tt) que, en caso de ser mayor que el audio se ajustará automáticamente al audio y en caso de ser menor reducirá la duración del audio.

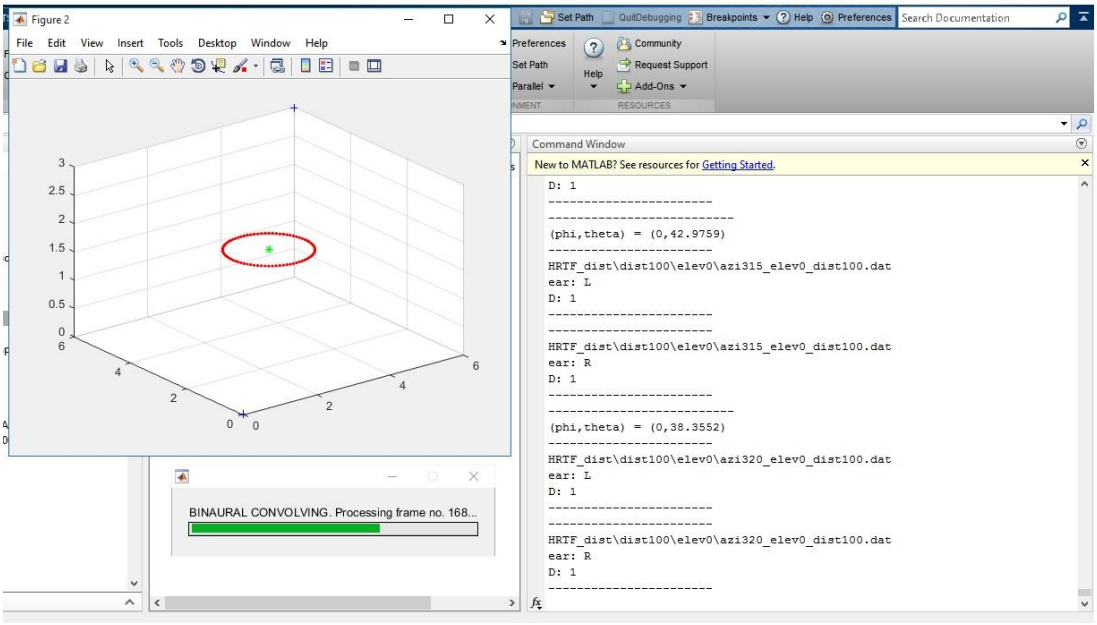


En las ultimas tres casillas se piden las componentes (x, y, z) de la trayectoria que describirá la fuente en el espacio σ(t):

$$\boldsymbol{\sigma(t) = (x(t), y(t), z(t))}$$

Las componentes de la trayectoria son vectorizadas para poderse usar para los cálculos de posición con el fin de obtener HRTFs y RIRs dependientes de posición (θ, φ) y distancia (r) .

Al cerrar esta ventana, el programa ya tiene todos los datos que necesita y comienza la convolución por tramas. El software muestra lo siguiente mientras convoluciona.



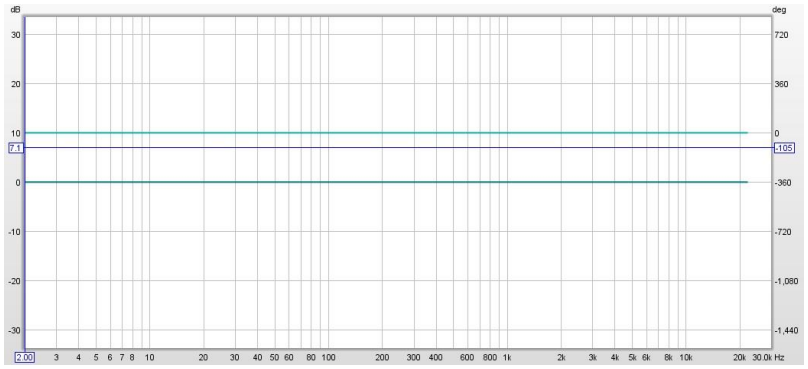
En una ventana se dibuja la trayectoria que se ha definido, en rojo. Los ejes están ajustados a las dimensiones de la sala. El punto verde representa la posición del oyente y dos aspas marcan las dos esquinas opuestas (0, 0, 0) y (rm1, rm2, rm3). La barra de progreso muestra qué numero de trama se está procesando actualmente. En la "Command Window" de Matlab se muestran detalles de la HRTF correspondiente a la trama actual; ángulos (θ, φ), distancia en el caso de la base de datos de HRTFs de la PKU (D), oreja (L o R) y ruta del archivo. Dependiendo de la elección de la base de datos, el proceso de convolución llevará más o menos tiempo.

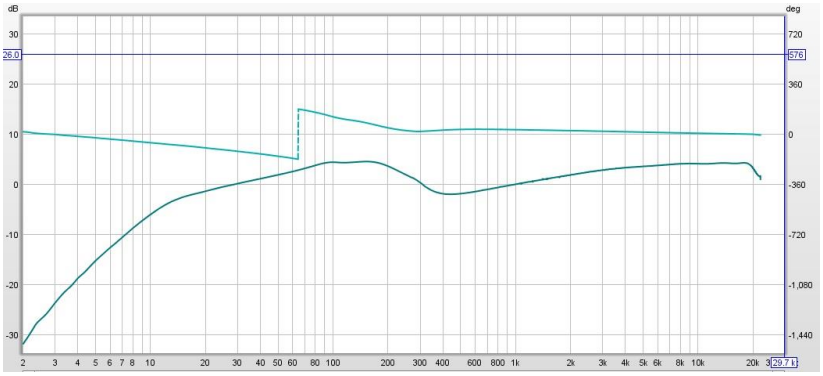
Cuando el programa finaliza el proceso de convolución, guarda el audio generado en un archivo. La ruta del archivo se muestra en una ventana. Luego, post-procesa el audio generado mediante el sistema de widening estéreo. El audio post-procesado es guardado en otro archivo de audio.
Finalmente, se muestran gráficas de algunas señales interesantes.

Se da la posibilidad de reproducir el audio generado mediante Matlab. Si no se desea, se puede cerrar Matlab y abrir el audio con cualquier otro programa desde los archivos guardados.

Para la toma de datos y reproducción del audio, el sistema usado ha sido el siguiente.

Un PC (Dell Studio 1558, Win10 x64) que tiene instalado Matlab (R2014b) se conecta a la interfaz de audio Focusrite (Saffire Pro 40) mediante un puerto Firewire IEEE 1394. La conexión a la interfaz de audio anula a la tarjeta de audio del PC, de manera que los conversores AD/DA usados son los de la interfaz, de calidad muy superior a los del PC.
Basta comparar las respuestas en frecuencia de ambas tarjetas de audio para hacerse una idea. La figura superior muestra la respuesta en frecuencia de la interfaz de audio Saffire Pro 40 de Focusrite. La figura inferior muestra la respuesta de la tarjeta de audio del PC Dell Studio 1558.





Usar la tarjeta de audio del PC, que por los resultados obviamente está dañada o defectuosa, hubiera implicado filtrar el audio con la respuesta en frecuencia inversa de la tarjeta o aceptar las coloraciones que esto hubiera introducido. Además de que, obviamente, el ordenador tampoco admite entradas TRS ¼'' o XLR, lo que descarta esta tarjeta integrada por impidir la conexión de altavoces profesionales. Una interfaz de audio más pequeña puede bastar [], pero si se quieren probar sistemas Surround 5.1 o 7.1 harán falta más entradas que las disponibles en las tarjetas de audio de gama baja.

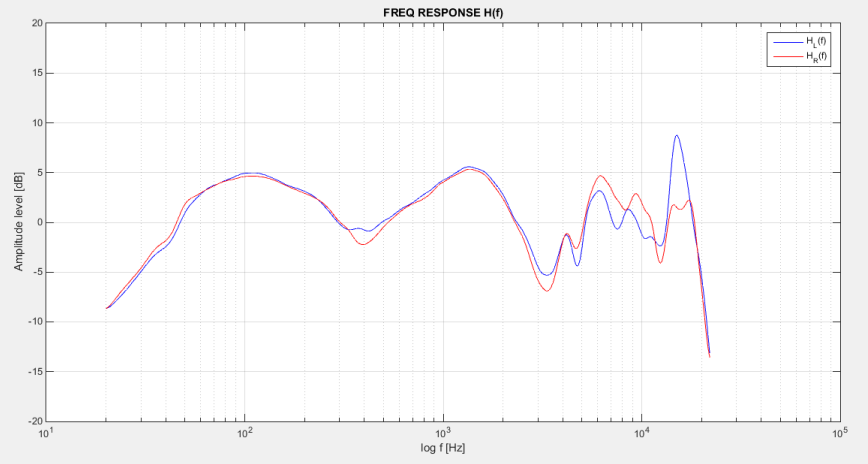
Además, la interfaz de audio permite la conexión de dispositivos de audio de carácter profesional mediante *conexiones balanceadas* TRS ¼'' (llamado a veces Jack) ó XLR, que reducen interferencias y ruidos y permiten longitudes de cable mucho mayores. La siguiente figura muestra el tipo de conectores que usa la interfaz.

Las líneas balanceadas se componen de dos conductores, uno porta la señal en fase y el otro en contrafase (desfasda 180º). La diferencia entre una línea balanceada y una no balanceada es básicamente que la no balanceada solo lleva la señal original. La mejora en las líneas balanceadas se basa en lo siguiente: si una interferencia logra atravesar la malla que recubre el cable, induce un transitorio en los dos conductores en el mismo sentido. En el receptor, para desbalancear la línea hay que invertir la señal que porta la contrafase y sumarla a la señal original (restar ambas señales) logrando así duplicar la amplitud de la señal resultante. Al invertir la contrafase, el transitorio queda invertido también, que al ser sumado a la señal con la fase anula la interferencia. Esto no ocurre en líneas no balanceadas.

A este tipo de interferencia se la conoce como señal de modo común y se caracteriza por la *razón de rechazo al modo común (common mode rejection ratio CMRR)*, que debería tener un valor de al menos 80 dB. Es por este motivo que se ha preferido usar líneas balanceadas, para asegurar una calidad de audio a prueba de interferencias.

Los altavoces usados (HARMAN-JBL LSR 305) son tipo monitores de estudio de dos vías, de 5 pulgadas y 1 pulgadas cada una, con una imagen sonora adecuada para reproducir audio 3D para más de un oyente así como una respuesta en frecuencia casi plana. Se conectan mediante cables balanceados con conectores XLR a la interfaz de audio.

Los auriculares usados son unos Pioneer (HDJ1500), con un cable no balanceado y con una respuesta en frecuencia no plana, como se aprecia en la siguiente figura. Pero, para no introducir coloraciones en el audio y tener una respuesta en frecuencia casi plana, se personalizó un archivo de calibración para estos auriculares gracias a la empresa Sonarworks®, que incluye la respuesta en frecuencia inversa de cada canal de los auriculares. Al convolucionar el audio binaural con esta respuesta inversa se pueden cancelar efectos de coloración de los auriculares.



En la figura se muestran las inversas de los filtros de ecualización proporcionados por Sonarworks®. Hay que notar que esto no es exactamente la respuesta en frecuencia original de los auriculares para cada canal L y R, debido a que esta gráfica muestra el resultado de multiplicar por -1 los filtros para ecualizar la respuesta en frecuencia. La inversión de la respuesta en frecuencia es algo no trivial y es por eso que se contactó con una empresa profesional para hacerlo, pues la medición sin un maniquí para audio binaural es muy compleja y, en caso de poder obtener la respuesta, los procesos de inversión requieren invertir teniendo en cuenta factores psicoacústicos. Todo esto sobrepasaba los fines de este proyecto.

Una vez descritos los componentes del sistema de reproducción, se pasa a la caracterización del esquema. La configuración completa se muestra en la siguiente figura. En ella se indica que, mediante la interfaz, se pueden grabar señales (micrófono, instrumento, entrada de audio, …) en caso de querer auralizarlas también. La interfaz se usó con un micrófono omnidireccional XREF20 de Sonarworks® calibrado para realizar pruebas de mediciones de RIRs y respuestas en frecuencia de los altavoces. Los micrófonos de medición profesionales necesitan de alimentación Phantom de 48 V.

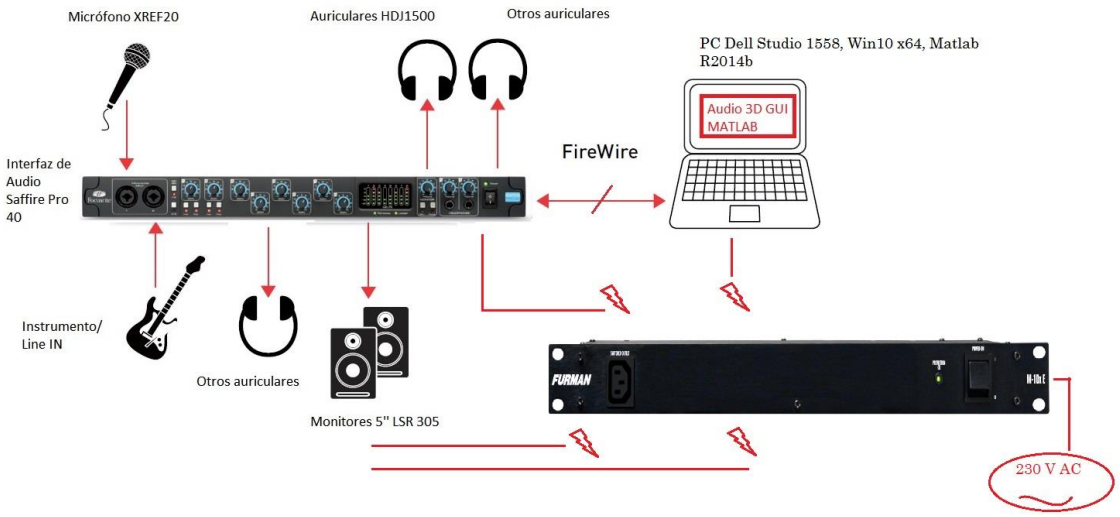
Todo el sistema usa un acondicionador de potencia Furman® M-10x E con 11 salidas a 230V-10A , que proporcionan una energía de alimentación limpia y constante, además de que el sistema protege a los componentes conectados a él de rudios debidos a acomplamiento de RF e interferencias, además de salvaguardar los dispositivos conectados a él en el caso de sobretensiones.

Las flechas en rojo indican las entradas y salidas, mientras que los símbolos rojos en forma de rayo indican las entradas de alimentación.

El puerto digital de entrada/salida de datos es un puerto Firewire IEEE 1394b a 400Mb/s. Viene indicado por una flecha doble con una barra cruzada en medio.

La ventaja de una interfaz de audio de estas características, en concreto que disponga de 10 salidas, es que permite las conexiones necesarias para las pruebas de un sistema 5.1 o 7.1 Surround en caso de disponer de él. Para este proyecto fin de carrera las pruebas más relevantes no conciernen a sistemas 5.1 o 7.1 y no se han descrito en esta memoria.

Sistema para pruebas del software de audio 3D desarrollado



Una vez que el sistema está completamente conectado y operativo, se lanza Matlab® y el software de audio 3D desarrollado mediante el programa "auralization.m". La aplicación recoge una serie de datos, procesa convenientemente el audio de entrada preferido y ofrece a su salida dos pares señales de alimentación para auriculares (sin y con post-procesamiento para widening estéreo) y dos para altavoces (las mismas que para auriculares pero post-procesadas mediante un cancelador de Crosstalk).

El oyente puede sacar sus propias conclusiones en base a la reproducción del audio 3D sintetizado, las gráficas ofrecidas para algunas señales características, así como mediante los parámetros acústicos y psicoacústicos que el programa ofrece.

4.3. Limitaciones tecnológicas de los sistemas de audio 3D

4.3.1. Limitaciones relacionadas con el procesado de señales fisiológicas

La misión de los sistemas de audio 3d es sintetizar, manipular y renderizar campos sonoros en tiempo real y que no existen en el entorno físico real, por ello sumergiendo al usuario en un entorno de realidad virtual sonora.

Muchas investigaciones han proveido de avances importantes en las áreas del procesado de video e imágenes. Sin embargo, el procesado de audio y en particular el del audio 3D o virutal no han avanzado al mismo ritmo. Las limitaciones que se dan en los campos del procesado de señal, las consideraciones acústicas del entorno, las características del aparato auditivo humano así como el movimiento del oyente son aspectos a considerar al realizar un sistema de audio 3D.

La reproducción precisa de sonido en el espacio puede incrementar notablemente la sensación de inmersión en un sistema de realidad virtual para aplicaciones en las que es importante conseguir una buena localización del sonido relativo a imágenes visuales. Los seres humanos son los uncos que son capaces de localizar e identificar sonidos en un entorno 3D con precisión relativamente buena. Por ejemplo, la resolución de las diferencias del tiempo de llegada en los humanos es de aprox. 7μs.
Como se ha comentado, la percepción sonora esta basada en características que incluyen las ILDs e ITDs asi como los efectos derivados de funciones de transferencia dependientes de la dirección causados por una reflexión en el pabellón auricular, la cabeza y el torso (las HRTFs propiamente dichas).
Ademas de estas características direccionales, los humanos usamos otras características como: timbre, respuesta en frecuencia y rango dinamico. Hay otras carcteristicas subjetivas que varian de oyente en oyente pero que son igual de importantes para conseguir la “suspensión de incredulidad” deseada en un sistema de realidad virtual. Estas características incluyen atributos como el ancho de fuente aparente, el envolvimiento del oyente, la claridad y calidez del sonido, …

Se diferencias dos clases de limitaciones: la primera comprende las limitaciones fundamentales debidas a los límites de las leyes físicas y la segunda consiste en las restricciones que impone la tecnología elegida y usada en cada aplicación.

Las limitaciones que impone la primera clase han de ser estudiadas cuidadosamente y su conocimiento es esencial para determinar la realizabilidad de una tecnología en particular al compararla con los limites físicos. Muchas de estas limitaciones fundamentales no dependen directamente de la elección del sistema de audio; mas bien están relacionanadas con el proceso de propagación del sonido en salas de geometría irregular. Las propiedades físicas del entorno acústico se codifican en el campo sonoro. Tienen que ser decodificadas por un sistema de audio para poder simular de manera realista los sonidos originales. La influencia del entorno acústico local en el que se encuentra el oyente se refleja en la percepción de atributos o características espaciales como direccion y distancia, como también el tipo de sala (si es una sala viva o más bien apagada y aburrida) y el tamaño de las fuentes. La situación es aun mas compleja; el hecho de que el proceso de decodificación tiene que incluir las transformaciones asociadas con las que realiza el sistema auditivo humano, complica el proceso de síntesis de audio 3D. Estas transformaciones en concreto son, por ejemplo, la conversión de las características del sonido 3D en diferencias de amplitud y tiempo (ILDs e ITDs) y efectos en la respuesta en frecuencia del sistema global debidos al pabellón aurivular, el cráneo y el torso y los hombros que se codifican conjuntamente en la HRTF.

La segunda clase de limitaciones, las que vienen de la elección de la tecnología usada en cada implementación particular, son de igual importancia para entender las aplicaciones potenciales de un sistema dado y son impuestas por la tecnología elegida para el sistema que se desea realizar.
Por ejemplo, el proceso de codificar características relacionadas con la acústica de salas en los campos sonoros puede ser modelado mediante métodos numéricos BEM o FEM.

Los sistemas más actuales renderizar campos sonoros usando la Auralización, una técnica que combina modelos a escala, filtrado digital y hardware de propósito especial para convolución en tiempo real con el

fin de predecir y renderizar el campo sonoro deseado. Conforme avanza la potencia de procesamiento de los sistemas de tratamiento de señales digitales (DSP) la capacidad de los sistemas de inmersión de audio y video se incrementa proporcionalmente.

La tecnología de audio 3D se basa en implementaciones digitales de las HRTFs. En principio se pueden conseguir excelentes resultados; no obstante y como se ha mencionado anteriormente en otras secciones, esto requiere conocer las HRTFs de cada individuo, algo casi inabarcable. Las investigaciones actuales, muy centradas en este área, intentan conseguir una buena localización para una amplia gama de oyentes usando HRTFs sintéticas o no individualizadas. Se suelen identificar cuatro barreras en la construcción de los sistemas que usan HRTFs de este tipo [1]: a) errores psicoacústicos como confusión frontal/atrás (típicas de los sistemas de audio 3D con auriculares), b) demasiado volumen de datos para representar todas las HRTFs con precisión, c) errores en la respuesta en frecuencia y fase originados por desadaptaciones en las HRTFs no individualizadas y las reales y d) las limitaciones técnologicas de los sistemas de computación para un usuario medio. No obstante la cuarta barrera ha sido parcialmente solventada actualmente mediante la migración de las aplicaciones a hosts provistos con GPUs, dejando la CPU libre para otras tareas. Hay que tener en cuenta que la síntesis de los entornos acústicos virtuales requiere el calculo de una RIR y la convolucion de esta con la HRTF del oyente en tiempo real. Las RIR típicas tienen una duración de 2 segundos, que muestreadas a 48 kHz requieren de un procesador capaz de operar a mas de 13 Gflops/canal. En cualquier caso, la meta es reducir el numero de cálculos requeridos para estas convoluciones.

4.3.2. Renderizado de audio espacial con auriculares

Aunque actualmente existen sistemas de audio 3D para auriculares en el mercado como p.ej. [hoja características DSPeaker], se identifican en general cuatro desventajas para estos sistemas:

a) Una información individualizada de cada oyente mediante sus HRTFs no es posible y las HRTFs genéricas imposibilitan adaptar cada percepción del sonido, individual y diferente para cada persona.
b) Hay errores grandes en la localización en el espacio virtual 3D de las fuentes sonoras originales, especialmente para la direccion mas importante: la frontal.
c) Los auriculares son incomodos
d) Es realmente complejo externalizar sonidos para contrarrestar la sensación de localización dentro de la cabeza (in-head localization).
e) Hay situaciones donde los auriculares son la única opción (p. ej. en cabinas de piloto o entornos multi-usuario)

Como se ha visto, el uso de altavoces consigue reducir varias de estas barreras pero solamente si se usa una cancelación de Crosstalk con una complejidad acorde al grado de realismo deseado.

4.3.3. Altavoces vs. Auriculares

A continuación se muestra una tabla que intenta resumir lo más brevemente posible las diferentes influencias para la elección del dispositivo de reproducción del audio 3D. Las características se marcan en tres colores: rojo, amarillo y verde. Las marcadas en rojo muestran desventajas de esa tecnología. Las marcadas en amarillo pueden ser ventajas en ciertas circunstancias y desventajas en otras. Las marcadas en verde son ventajas de la tecnología de reproducción.

	ALTAVOCES	AURICULARES
Ergonomía	Usuario "liberado"	Incomodos tras un tiempo
	No aíslan al usuario	Aislan al usuario
	Requieren instalación	Sin instalación
	Calidad dependiente de la posición del oyente	Calidad dependiente de la posición del oyente
Entorno acústico	Sala preparada acusticamente	Sala sin requisitos
Algoritmia	EQ altavoces	EQ auriculares
	Correccion de Crosstalk	Canales independientes
	Filtros muy largos para posiciones alejadas	
Calidad del audio obtenida y sensación de percepción de espacialidad	Experiencias espaciales muy realistas	Mejor forma de presentar el sonido en 3D
	Muy sensible a la posición del oyente	Independiente de la posición del oyente
	Las posiciones virtuales lejanas a las líneas de los altavoces son difíciles de conseguir	Muy versátil
	Externalizacion y sensación de distancia sencillas de conseguir	Suelen generar sensaciones demasiado cercanas (efecto de localización in-head)

4.3.4. Limitaciones relacionados con la Acústica de salas

Aunque los métodos de la Acústica Geométrica proveen buenos resultados de simulación, no se debería olvidar que estos métodos de descomposición de ondas sonoras tienen implícitas limitaciones importantes. Con esto se quiere poner de manifiesto que los métodos de la Acústuca Geométrica solo pueden aproximar la propagación del sonido del mundo real.

Salas grandes. La limitación mas importante de la Acustica Geometrica es la restricción a salas muy grandes. El "Método de Imágenes" como también el "Ray Tracing" usan reflexiones geométricas de las ondas sonoras que tengan longitudes de onda pequeñas en comparación con la geometría de la sala. Efectos acústicos de fase como modos de la sala, que influyen significatiamente el rango de bajas frecuencias, no se tienen en cuenta en los métodos geométricos.

Señales sonoras de banda ancha. Se requieren señales con un espectro rico para la Auralización, ya que eliminan efectos de fase que ocurrirían con señales monocromáticas o señales tipo tonos puros.

Modelos CAD de la sala. Es importante para obtener buenos resultados en la simulación que se tengan modelos adecuados y precisos de la sala. En este contexto, un modelo adecuado no tiene porque significar un alto grado de detalle pero sí una reproducción precisa de las características de reflexión de objetos y bordes de la sala usando coeficientes de absorción y scattering (dispersión) como los descriptores de una superficie.

Coeficientes de absorción pequeños. El "Metodo de Imágenes" es exacto solamente para coeficientes de reflexión $R = 1$ y $R = -1$ aunque sigo siendo una buena aproximación para coeficientes de absorción pequeños. No obstante, el campo de onda plana que se asume en el momento de la reflexión es adecuado solo si se considera una onda esférica con un espacio grande entre la pared y la fuente y el receptor y con angulo de incidencia constante (esto ultimo solo es valido para angulos hasta 60º entre la reflexión incidente y la normal a la superficie).

4.4. Audio 3D usando una GPU con CUDA®

El procesamiento de señales acústicas ha sufrido cambios importantes en los últimos años. La incorporación de información espacial en un entorno virtual audiovisual o en videojuegos dota a la aplicación de mayor sensación de presencia. El audio 3D consiste en reproducir señales acústicas con características espaciales incorporadas. Esta información espacial permite al oyente identificar las posiciones virtuales de las fuentes correspondientes a diferentes sonidos. El audio 3D con auriculares se obtiene filtrando las diferentes señales de audio de las fuentes con los filtros HRTF.
La computación eficiente puede jugar un papel imporante cuando el numero de fuentes a manejar aumente. Esto aumentaría el numero de operaciones de filtrado, requiriendo de gran capacidad de computo, especialmente en el caso de que las fuentes se muevan en el espacio describiendo trayectorias en 3D. Las *Unidades de Procesamiento Gráfico (Graphics Processings Units GPU)* consisten en co-procesadores programabables con alto grado de paralelizacion y proveen de una capacidad de computo impresionante comparada con las *Unidades Centrales de Procesamiento (Central Processing Units CPUs)* típicas de los ordenadores recientes, exceptuando ciertos procesadores como las versiones más potentes de Intel® Core i7™ o Intel® Xeon™. No obstante, obtener el mejor rendimiento en las GPUs supone paralelizar adecuadamente las instrucciones para el cómputo.

La aparición del *Compute Unified Device Architecture (CUDA)* por parte de la compañía fabricante de tarjetas gráficas NVIDIA® ha permitido que actualmente se pueden usar las GPUs para aplicaciones mas alla del procesamiento de imágenes y video. Programar eficientemente una GPU requiere el conocimiento de la arquitectura de la GPU asi como saber cómo distribuye una GPU sus tareas por sus distintas unidades de procesamiento.

El objetivo de esta sección es presentar una aplicación para audio 3D con auriculares cuyo procesamiento se lleva a cabo exclusivamente en una GPU. La aplicación supone una interaccion con el oyente que seleccionaría cambiaria y movería las fuentes de sonido en tiempo real. La característica mas notable de un sistema asi es su escalabilidad. La aplicación debe funcionar en cualquier dispositivo con una GPU de NVIDIA.
La siguiente parte describe el uso de las HRTFs para esta aplicación. Luego, se estudian algunas características arquitectónicas y de programación de las GPUs que permiten usar CUDA. Finalmente, se describe la aplicación de audio 3D y la manera de hacer que interactue con el oyente.

4.4.1. Uso de las HRTFs

Sea $x(n)$ el buffer o vector de entrada compuesto por L muestras de la señal de audio de entrada $s(n)$, siendo n el índice de la muestra en el buffer. En una aplicación de audio en tiempo real, cuando el buffer se llena se envía a la unidad de procesamiento mientras que otro buffer $x(n)$ es rellenado con muestras de nuevo, y así sucesivamente. Los buffers de salida $yL(n)$ e $yR(n)$ representan los sonidos ya procesados que van a los canales L y R, respectivamente. Se obtienen convolucionando la señal $x(n)$ con las dos HRTFs $HRTF\_L(\theta, \varphi, n)$ y $HRTF\_R(\theta, \varphi, n)$. De este modo, para obtener los sonidos en 3D para múltiples fuentes, será necesario primero sumar todas las salidas $yi(n)$ para cada sonido y para cada canal antes de reproducir el audio.
Como ejemplo, si se tienen dos fuentes $\{x1(n), x2(n)\}$, $n = 0, 1, \ldots, L - 1$ y queremos la fuente $x1(n)$ en la posición:

$$(\theta 1, \varphi 1) = (0º, 90º)$$

y la fuente $x2(n)$ en la posición:

$$(\theta 2, \varphi 2) = (225º, 0º),$$

las señales de salida serán el resultado de la convolución con las dos HRTFs, de longitud también L para simplificar:

$$yL(n) = HRTF\_L(\theta 1, \varphi 1, n) * x1(n) + HRTF\_L(\theta 2, \varphi 2, n) * x2(n)\ ,$$

$$yR(n) = HRTF\_R(\theta 1, \varphi 1, n) * x1(n) + HRTF\_R(\theta 2, \varphi 2, n) * x2(n)\ ,$$

donde (*) denota la operación de convolución. Las dos salidas yL(n) e yR(n) tendrán longitud $N = L + L - 1 = 2L - 1$.
El problema reside en que las fuentes se mueven, bien porque el oyente la ha cambiado manualmente usando una interfaz o bien porque una aplicación, como un videojuego, requiere un reposicionado de las fuentes de sonido. Por ejemplo, si la fuente x1(n) se desea mover desde la posición anterior a la posición:

$$(\theta 1', \varphi 1') = (0º, 45º)$$

y la fuente x2(n) desde su posición anterior a la posición:

$$(\theta 2', \varphi 2') = (135º, 0º)$$

una manera común de simular este movimiento es cambiar de filtros HRTF_L y HRTF_R, ya que estos dependen de la dirección de procedencia del sonido.
El problema es que este cambio produce un "clipping" o sonido en forma de click perfectamente audible debido a la gran diferencia entre HRTFs. En el apartado de "Switching BRIRs" se analizó una publicación que sugería tres métodos para cambiar de HRTFs y se concluyó que mediante el método de fade-in-fade-out no se percibían ya apenas estos molestos sonidos. Usando las funciones rampa f(n) y g(n) las nuevas señales serían, mediante el método de fade-in-fade-out:

$$yL(n) = (HRTF\_L(\theta 1, \varphi 1, n) * x1(n))\cdot g(n) + (HRTF\_L(\theta 1', \varphi 1', n) * x1(n))\cdot f(n) + (HRTF\_L(\theta 2, \varphi 2, n) * x2(n))\cdot g(n) + (HRTF\_L(\theta 2', \varphi 2', n) * x2(n))\cdot f(n)$$

$$yR(n) = (HRTF\_R(\theta 1, \varphi 1, n) * x1(n))\cdot g(n) + (HRTF\_R(\theta 1', \varphi 1', n) * x1(n))\cdot f(n) + (HRTF\_R(\theta 2, \varphi 2, n) * x2(n))\cdot g(n) + (HRTF\_R(\theta 2', \varphi 2', n) * x2(n))\cdot f(n)$$

Ambas funciones f(n) y g(n) pueden tener formas diferentes, como senos, cosenos, raíces cuadradas, rampas, ventanas de Hamming, … La longitud de las funciones será la del resultado de la convolución, N. Una función rampa funcionará bien en la mayoría de los casos. Las dos funciones ponderan los resultados de las convoluciones par evitar el sonido oído como un "click" cada vez que se cambia de par de HRTFs.

4.4.2. Las Unidades de Procesamiento Gráfico (GPUs) y CUDA®

Compute Unified Device Architecture (CUDA) es una plataforma de computación paralela e *interfaz de programación de aplicaciones (Application Programming Interface API)* que aprovecha toda la capacidad de procesamiento de las GPUs programables. Las GPUs pueden tener varios *multiprocesadores stream (SM)*, donde cada SM consiste bien en 8 cores si la versión de CUDA es 1.x o en 32 cores si la versión es 2.x. Las GPUs pueden tener una gran capacidad de memoria off-chip (memoria global) y una rápida y pequeña memoria on-chip (memoria compartida). En el modelo CUDA, el programador se encarga de definir la función del kernel. El código que será ejecutado en la GPU está escrito en el kernel.

Una configuración en forma de malla, que defina el número de hilos a utilizar y cómo están distribuidos y agrupados, ha de estar implícita en el código principal de la aplicación. Se puede definir una malla como una malla de bloques, cada una con una malla de hilos. Así, un hilo se puede identificar mediante unas coordenadas dentro de un bloque (ThreadIdx.x, ThreadIdx.y, ThreadIdx.z), el cual estará definido dentro de una malla también mediante unas coordenadas (BlockIdx.x, BlockIdx.y, BlockIdx.z). El tamaño de bloque también tendrá que definirse (BlockDim.x, BlockDim.y, BlockDim.z), de la misma manera que el de la malla global (gridDim.x, gridDim.y, gridDim.z).

El kernel distribuye todos los bloques de hilos por los SMs cuando la GPU comienza a procesar. Cada SM crea, gestiona, planifica y ejecuta hilos en grupos de 32 hilos llamados *warps.* Estos warps son gestionados por un planificador de warps para la ejecución.

La manera en la que un bloque se divide en warps siempre es la misma. Cada warp contiene hilos consecutivos, incrementando el identificador del hilo "Idx", con el primer warp conteniendo Idx = 0. Cada warp ejecuta una instrucción común a todos sus hilos por unidad de tiempo; por esto la eficiencia máxima es conseguida cuando todos los 32 hilos de un warp coinciden en su camino de ejecución (no hay *serialización de warps*). Es importante tener múltiples warps en un bloque debido a que permite eliminar la latencia. Esto es, si todos los hilos de un warp tienen que ejecutar un acceso a memoria, este acceso puede tomar varios ciclos de reloj. Para eliminar la latencia, el planificador de warps elige uno que esté en estado de listo y va cambiando de warp para maximizar la utilización de los SMs.

Un aspecto importante a tener en cuenta a la hora de acceder a la memoria global es que se debería acceder de una forma *coalescente.* La coalescencia significa en computación que los hilos se escriben en un rango pequeño de direcciones de memoria, con un cierto patrón. Para ilustrarlo con un ejemplo, considere un puntero "array" que apunta a memoria global e "Idx" un puntero para el hilo; si el hilo "Idx" escribe en la dirección "array[Idx]" y el hilo "Idx+1" escribe en la dirección "array[Idx+1]", se ha conseguido coalescencia.

El uso de GPUs ha estado siempre asociado a aplicaciones de vídeo o imágenes. Desde la aparición de CUDA, varios grupos de investigación han usado las GPUs como un medio para acelerar los cálculos en ámbitos diferentes de la ciencia y la tecnología. También en el ámbito de la Acústica y el Procesamiento de Audio encontramos numerosos estudios que usan GPUs para desarrollar sus aplicaciones []

4.4.3. Ejemplo de software de audio 3D con auriculares usando una GPU

Como se ha comentado, el fin de muchas aplicaciones de audio 3D es convolucionar cada buffer de datos xi(n) de longitud L, que contiene muestras de cada señal fuente si(n), con los filtros HRTF correspondientes a la dirección de procedencia del sonido HRTF_L(i, n) y HRTF_R(i, n), por simplicidad también de longitud L. Luego, los dos conjuntos de resultados yi(n) se solapan y suman por separado para formas los dos vectores de salida yL(n) e yR(n). Ya que en una aplicación en tiempo real se trabaja con tramas y aquí se discute cómo implementar dicha aplicación en una GPU, no es relevante cuántas tramas xi(n) haya por fuente; lo importante es cómo se trabaja con cada trama, ya que van llegando en tiempo real.
Al arrancar la aplicación, todos los filtros en el dominio del tiempo (filtros HRIR) se transfieren de la CPU a la GPU; a continuación se llevan a cabo FFTs para obtener los filtros HRTF. Los pasos generales de una aplicación sencilla de audio 3D que use una GPU e interaccione con el usuario en tiempo real serían, a cada instante:

1) Llenado de los buffers de entrada con las L muestras de entrada de las diferentes fuentes. Matemáticamente, obtener los K vectores xi(n) de las K fuentes, donde i = 0, 1, …, K – 1, n = 0, 1, …, L – 1.

2) Transferir los K vectores xi(n) a la GPU

3) Transferir un vector de posición "POS", de longitud K. No obstante, en la GPU el vector POS será de longitud 2K; la segunda mitad se usará para guardar las nuevas posiciones cuando las fuentes de sonido cambien. Cuando esto ocurra, se tendrán que computar el doble de datos (ver Interaccion con el usuario).

4) La convolución se lleva a cabo, por ejemplo, en los siguientes pasos:

a. realizar una FFT de cada vector xi(n),
b. realizar multiplicaciones elemento a elemento de cada una de las L muestras de xi(n) con cada una de las L muestras de los filtros correspondientes HRTF_L y HRTF_R; ahora todos los vectores están en el dominio frecuencial,
c. realizar una suma elemento a elemento de los dos conjuntos de todos los bloques de salida yi(n) de longitud N,
d. realizar dos IFFTs, una para cada vector de salida, para transformarlos al dominio temporal. Los resultados son yL(n) e yR(n),
e. en caso de que las direcciones de procedencia de las fuentes cambiaran en ese instante,

los vectores de salida tendrían que ponderarse mediante las funciones f(n) y g(n), de longitud N.

5) Transferir los vectores yL(n) e yR(n) de la GPU a la CPU. Finalmente, las señales están listas para reproducirse usando auriculares.

En aplicaciones de audio en tiempo real, el paso 1) se ejecuta todo el rato durante todo el tiempo que dura la aplicación, ya que cada muestra de audio de cada fuente llega al buffer con periodo estándar de 1/fs segundos. La aplicación va a tener que usar dos buffers A y B de entrada. Cuando A se llena, se transfiere a la GPU y comienza el procesamiento. Mientras, las muestras de audio llenan el buffer B. Para conseguir el mejor rendimiento en la aplicación, el procesamiento del buffer A (pasos 2 a 4) debe acabar antes que se llene el buffer B (paso 1). Despues de esto se transfiere B a la GPU para empezar a procesarlo mientras que A se vuelve a llenar. La aplicación acabara normalmente cuando no haya mas muestras de audio que procesar.

El tiempo que se tarda en llenar el buffer es L/fs y es independiente del numero de fuentes. No obstante, el tiempo que se tarda en procesar sí depende tanto del numero de fuentes K como del uso de los recursos de la GPU para llevar a cabo el procesamiento. Por ello, es importante desarrollar una implementación eficiente en la GPU que permita conseguir máximo numero de fuentes K y a la vez satisfacer:

$$\text{tiempo de procesamiento} < \text{tiempo de llenado del buffer} = L/fs$$

4.4.4. Interacción con el usuario

Con el fin de interactuar con el usuario, las aplicaciones para CUDA suelen usar un objeto conocido como *stream.* Este objeto es una secuencia de instrucciones que se ejecutan en orden. Streams diferentes pueden no obstante ejecutar sus instrucciones en distinto orden uno con respecto al otro, o puede ser que lo hagan concurrentemente. Este objeto se suele usar básicamente para paralelizar los cálculos con las transferencias CPU <-> GPU. Para ello, la aplicación corre el stream 0 mientras que el stream 1 se mantiene a la escucha, que aguarda por los de posición de fuentes. Cuando se detecta un cambio, el stream 1 transfiere las nuevas posiciones de las fuentes a la GPU. Ya que esta transferencia se lleva a cabo en un stream distinto, se permite que esta transferencia se paralelice con el procesamiento del kernel. La aplicación detecta la llegada de nuevas posiciones y ejecuta las operaciones de convolucion y suma con el siguiente buffer de datos. Con este fin, las modificaciones que ocurren en los pasos 4.b y 4.c son tan simples como simular, solo para ese buffer de muestras, que el valor de K es $2 \cdot K$ y el numero de salidas es 4 (2 para el canal L y 2 para el canal R). Después del paso 4.d, las cuatro salidas se reducen a dos, ponderando las salidas con las funciones f(n) y g(n). Finalmente, las dos señales resultantes yL(n) e yR(n) se envían de vuelta a la CPU.

4.4.5. Implementación

De los pasos que lleva a cabo la aplicación de audio 3D propuesta, los pasos 4.a y 4.d son llevados a cabo automáticamente por la librería de FFTs de NVIDIA “CUFFT”.

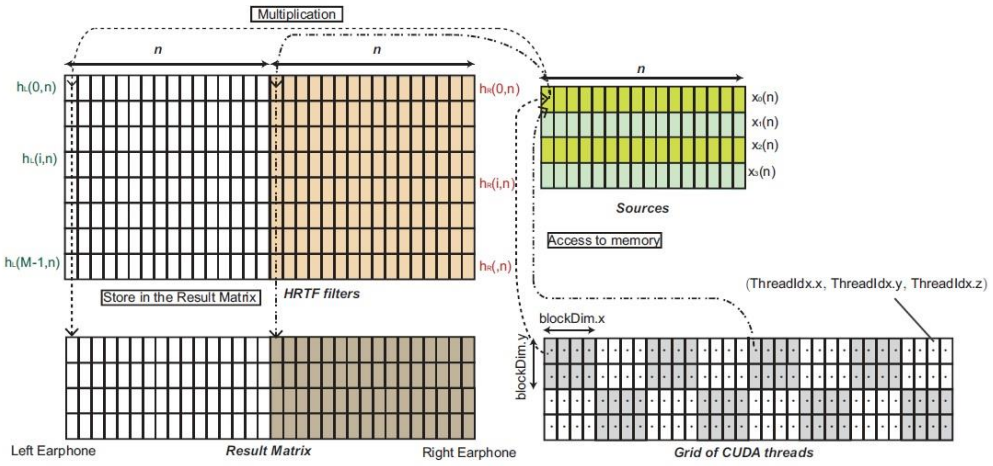
Hay algunas técnicas de programación para conseguir el mejor rendimiento de una aplicación que use CUDA:
· coalescencia máxima para acceder a memoria global y serialización warp minima,
· eliminar la latencia que ocurre por dependencias entre registros,
· mantener aproximadamente una ocupación minima de 25%,
· que el número de hilos por bloque sea si es posible múltiplo del tamaño del warp.

Es importante implementar kernels eficientes para multiplicaciones elemento a elemento (paso 4.b) y sumas elemento a elemento (paso 4.c). Dos ejemplos se muestran a continuación.

Multiplicación elemento a elemento

Un ejemplo de kernel sería uno que usara un hilo de la GPU para: tomar una muestra de una de las xi(n), multiplicar esa muestra con su correspondiente muestra de HRTF_L o HRTF_R y guardarla en una matriz "RESULTADO". Digamos que la longitud de las HRTFs se ha rellenado con ceros de manera que tienen longitud N. La matriz "RESULTADO" tendrá dimensiones 2·K x N y este núcleo usaría por tanto 2·K·N hilos.

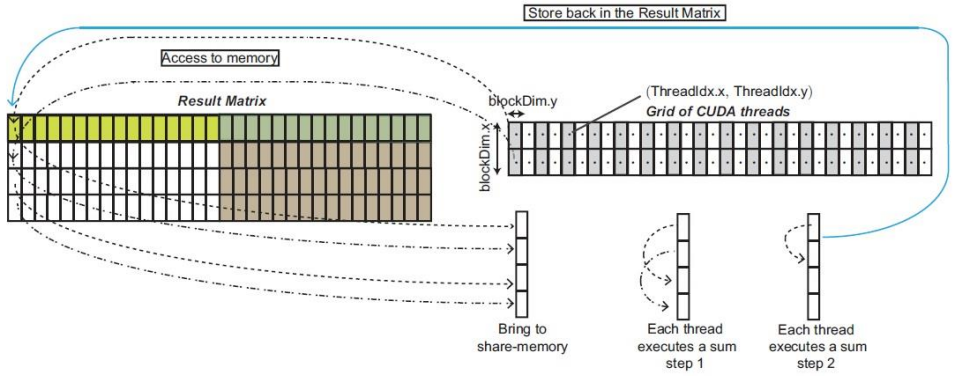
Una vez que la tarea de cada hilo esta clara y ha sido definida, se lanza el kernel. Para ello es necesario organizar los hilos, como se ha comentado, en bloques y también configurar una malla de bloques CUDA. La figura muestra las operaciones que lleva a cabo un hilo dentro de su bloque.



La matriz inferior derecha es una vista general de la malla de hilos. Cada punto en la matriz representa a un hilo y las dimensiones de un bloque van señaladas por blockDim.x y blockDim.y.
Si hay K fuentes, la matriz superior izquierda, de dimensiones K x 2N, guarda los 2K filtros de auralización HRTF_L y HRTF_R, ya transformados al dominio frecuencial. La matriz superior derecha guarda los K bloques de audio de entrada xi(n).
Las flechas muestran el flujo de datos. Se toma un hilo (un punto en la malla). Se accede a memoria para obtener una muestra de una señal xi(n). Esa muestra, transformada mediante una FFT, se multiplica 2 veces, una vez por cada una de las dos HRTFs que correspondan con esa xi(n). Las dos muestras resultantes se guarda en la matriz "RESULTADO" de tamaño K x 2N.

Suma elemento a elemento

El siguiente paso es sumar todas las componentes yi(n) de los dos conjuntos de salida; esto es, llevar a cabo una *reducción* de las filas de la matriz "RESULTADO" ["Algebra y geometría", Hernandez]. NVIDIA recomienda un algoritmo de reducción en [], que aprovecha la memoria compartida de la GPU. La idea fundamental es que los hilos dentro del mismo bloque llevan a memoria compartida 2·blockDim.x componentes. Posteriormente en la memoria compartida cada hilo ejecuta una suma de dos componentes. Este paso se repite hasta que se hayan sumado todas las componentes. Por ella, para cada instante de tiempo que se ejecuta una suma, se usan la mitad de hilos de la iteración previa.
Para adaptar este algoritmo de reducción a la estructura de datos de CUDA, tenemos que aplicarlo por columnas. Por este motivo, se definirán bloques unidimensionales, uno por columna. La siguiente figura muestra la adaptación de este algoritmo a la estructura de datos.



Este esquema funciona siempre que el numero de fuentes K sea menor que 1024, debido a que el tamaño máximo de bloque es 512 hilos.

En caso de querer consultar bibliografía adicional acerca de otros sistemas de procesamiento paralelo de audio 3D mediante GPGPUs, se recomienda consultar por ejemplo [3-8].

4.5. Otras aplicaciones profesionales que implementan audio 3D en Matlab®, Octave®, C++ ® y otros lenguajes de programación

· ITA Toolbox. Por la Universidad RWTH Aachen, Alemania. Para Matlab/Octave.
http://www.ita-toolbox.org/

· SLAB. Por el Ames Research Center, NASA. Para lenguaje C++.
http://human-factors.arc.nasa.gov/slab/

· SoundScape Renderer (SSR). Por el QU Lab de la universidad TU Berlin, Alemania.
Para GNU/Linux y MAC OS X.
http://spatialaudio.net/ssr/
http://www.tu-berlin.de/?id=ssr

· Auditory Image Model Toolbox. Por el CNBH. Para Matlab.
http://www.pdn.cam.ac.uk/groups/cnbh/aimmanual/

· Varias tecnologias de audio desarrolladas. Por el instituto IRCAM, Francia.
http://www.ircam.fr/eac.html?L=1

· Sustainable Software for Audio and Music Research
http://soundsoftware.ac.uk/tools

· Tecnologias de sonido Surround. Por Dolby Laboratories, Inc.
http://www.dolby.com/us/en/index.html

· Tecnologias de audio 3D para empresas (industria cinematográfica, sistemas AV, videojuegos, automóvil, broadcasting y aplicaciones móviles) y particulares. Por Auro Technologies.
http://www.auro-3d.com/

· PsyLab
http://www.hoertechnik-audiologie.de/psylab/

· Auditory Toolbox. Por Malcolm Slaney
https://engineering.purdue.edu/~malcolm/interval/1998-010/

· Binaural Toolbox. Por Michael Akeroyd. Para Matlab.
http://www.ihr.mrc.ac.uk/index.php/products/display/software

· Auditory modelling Toolbox. Por Jens Blauert et al. Para Matlab/Octave.
http://amtoolbox.sourceforge.net/

· Sound Field Synthesis-Toolbox (SFS). Para Matlab/Octave.
https://github.com/sfstoolbox/sfs

· The Large Time-Frequency Analysis Toolbox (LTFAT). Para Matlab/Octave
http://ltfat.sourceforge.net/

· Spatially Oriented Format for Acoustics (SOFA). Para Matlab/Octave
http://sourceforge.net/projects/sofacoustics/

· Tecnologias de audio. Por Isono.
http://www.iosono-sound.com/home/

· Ray Meddis
https://github.com/rmeddis/MAP

BIBLIOGRAFÍA

[1] “Fundamental and technological limitations of immersive audio systems”, Kyriakakis, C. Proceedings of the IEEE, Vol. 86, No. 5, 1998.

[2] “Headphone-based spatial sound with a GPU accelerator”, Belloch et al.

[3] “Real-time massive convolution for audio applications on GPU”, Beloch et al, 2011.

[4] “Ray acoustics using computer graphics technology”, Rober et al, DAFx, 2007.

[5] “Virtual Room Acoustics: A comparison of techniques for computing 3D-FDTD schemes using CUDA”, Webb, AES, 2011.

[6] “Computing Room Acoustics with CUDA – 3D-FDTD schemes with boundary losses and viscosity”, Webb, IEEE, 2011.

[7] “Spatial sound for video games and virtual environments utilizing real-time GPU-based Convolution”, Cowan, 2008.

[8] “GPU-based one-dimensional convolution for real-time spatial sound generation”, Cowan, 2009.

ANEXO

A Presupuesto y Planificación para el Proyecto Fin de Carrera

A.1. Introducción

En este capítulo se quiere dar una estimación del tiempo invertido en el desarrollo del proyecto, así como un posible presupuesto de inversión para el mismo. Para este fin, el proyecto se puede dividir en tres partes :

· Documentación sobre audio 3D, instalación y familiarización con el software usado para el desarrollo de la aplicación
· Desarrollo e implementación completo, la programación y puesta en funcionamiento del sistema completo para hacer pruebas y ver los resultados conseguidos comparándolos con otros softwares
· Elaboración de la memoria completa del PFC

El tiempo estimado para cada una de estas fases ha sido el siguiente:

· en cuanto a la documentación, el tiempo empleado ha sido todo el posible. No obstante para sentar bases y empezar a programar se ha necesitado un tiempo aproximado de cuatro meses, de octubre de 2014 a enero de 2015.

· El desarrollo de la aplicación junto con la formación sobre la ingeniería de audio 3D ha sido la parte complicada y ha llevado diez meses, de diciembre de 2014 a agosto de 2015.

· El desarrollo de la memoria se ha llevado en paralelo con la parte final del proyecto y ha llevado cinco meses, de julio de 2015 a diciembre de 2015.

El tiempo total de desarrollo del PFC ha sido de trece meses, de octubre de 2014 a noviembre de 2015. La dedicación ha sido a tiempo parcial, en jornadas de cuatro horas diarias, inclusive fines de semana. Al haber tenido que dedicar tiempo a estudiar y preparar los exámenes finales de carrera, el número de horas dedicadas ha sido de cuatro al día, todos los días.

A.2. Presupuesto aproximado

Para una estimación del coste teórica, nos basamos en los métodos usados en empresas y multinacionales. La estimación se hará según estos valores teniendo en cuenta dos aspectos fundamentales:

- el Proyecto Fin de Carrera ha sido desarrollado por *una sola persona* cuyo nivel profesional corresponde a un ingeniero licenciado (M. Ing) en Telecomunicación

- El desarrollo del protecto ha comprendido un tiempo de *trece meses*, de octubre de 2014 a noviembre de 2015.

El desglose del presupuesto comprende dos bloques: el coste de materiales humanos en el desarrollo del mismo y el coste de recursos humanos (RRHH).

A.2.1. Presupuesto para materiales

En cuanto al coste de los materiales, se ha dividido entre el hardware y el software adquirido para el PFC así como bibliografía en papel imprescindible para empezar a estudiar la ingeniería de sonido. El desglose se muestra a continuación

1 Equipamiento HW y SW

Precio unidad	Cant.	Artículo
-(pagado)	*1x*	*Ordenador portátil Laptop PC Studio™ 1558 de Dell®, Microsoft® Windows™ 10 x64, Intel® Core i5 430M @ 2.27 GHz, ATI® Mobility Radeon™ HD 4500 Series 512MB, Disco duro SSD 2.5'' 250 GB, Memoria DRAM 8 GB DDR3 @ 1333 MHz.*
-(pagado)	*1x*	*Licencia de un año del software Matlab R2014b™ de Matlab®.*
-50€	*2x*	*Memoria DRAM 4GB DDR3 @ 1333 MHz de Dell®.*
-120€	*1x*	*Memoria SSD 2.5'' 250 GB de Crucial®.*
-50€	*1x*	*Memoria HDD 2.5'' 1 TB de Toshiba®*
-70€	*1x*	*Impresora OfficeJet™ 5740 de HP®.*
-469€	*1x*	*Interfaz de Audio Saffire™ Pro 40 de Focusrite®.*
-10€	*1x*	*Cable Firewire IEEE 1394b 9/6 pines de Roline®.*
-150€	*2x*	*Monitor de estudio 5'' LSR 305 de HARMAN-JBL®.*
-26€	*2x*	*Cable XLR m/Jack m balanceado con conectores Neutrik®.*
-38€	*2x*	*Cable extensión Jack m/h balanceado con conectores Neutrik®.*
-95€	*1x*	*Micrófono omnidireccional XREF20 calibrado de Sonarworks®.*
-20€	*1x*	*Cable XLR m/h balanceado con conectores Neutrik®.*
-30€	*1x*	*Auriculares Earpods de Apple®.*
-15€	*1x*	*Auriculares Earphones MX270 Dynamic Audio de Sennheiser®.*
-175€	*1x*	*Auriculares circumaurales HDJ 1500 de Pioneer®.*
-50€	*1x*	*Calibración auriculares Pioneer® HDJ 1500 por Sonarworks®.*

Subtotal: -1519€

2 Bibliografía

Precio unidad	Cant.	Artículo
-50€	*1x*	*"Tratamiento de señales en tiempo discreto", Oppenheim, A. y Schafer, Pearson Prentice-Hall, 1989.*
-143€	*1x*	*"Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality", Vorländer, M. Springer, 2008.*
-125€	*1x*	*"Spatial Audio Processing", Breebaart, J. and Faller, C. Wiley, 2012.*
-40€	*1x*	*"Sound System Engineering", Davis, C. Focal Press, 1989.*
-32€	*1x*	*"Master Handbook of Acoustics", 2nd ed., Everest, A. McGraw-Hill, 1989.*

Subtotal: -390€

3 Consumibles utilizados

Precio unidad	Cant.	Artículo
-400€	*1x*	*Electricidad total consumida por el equipamiento*
-25€	*12x*	*Conexión a internet por cable (doce meses x 25€/mes)*
-150€	*1x*	*Material de oficina (tinta, papel, CDs, USBs, bolígrafos, …)*

Subtotal: -850€

A.2.2. Presupuesto para RRHH

Para calcular el coste de recursos humanos (RRHH) tenemos en cuenta los datos anteriores. Para poner el proyecto en el contexto de una empresa, se establece el sueldo del ingeniero a cargo del proyecto.

Un sueldo bruto para un ingeniero licenciado hardware y software (M. Ing. Telecomunicación), de aprox. 3000€/mes y tomando en consideración que han sido doce meses con *jornada a tiempo parcial de 4 horas diarias, inclusive fines de semana*, queda reducido a la mitad, 1500€/mes.

Esto da un total de 4h/día x 7 = 28 horas/semana, dedicadas a investigación de temas de Acústica, ingeniería de sonido y Psicoacústica, desarrollo del software propio de audio 3D y pruebas de software de otros autores.

El sueldo total percibido por todo el proyecto es, aproximadamente:

Precio unidad	Cant.	Artículo
-1500€	*13x*	*Sueldo ingeniero hardware y software (M. Ing. Telecomunicación), media jornada: doce meses x 1500€/mes*

Subtotal: -19500€

No se ha necesitado de ningún otro trabajador que afectara de forma directa al desarrollo del proyecto aparte del estudiante y del director de proyecto.

A.3. Tabla con el presupuesto del proyecto completo

En la tabla de la página siguiente se detalla el desglose total del Proyecto Fin de Carrera.

CANT.	ARTÍCULO/DESCRIPCIÓN	PRECIO 1x	TOTAL
	COSTE MATERIALES		
	1 Equipamiento HW y SW		
1x	Ordenador portátil Laptop PC Studio™ 1558 de Dell®	(pagado)	0€
1x	Licencia de R2014b de Matlab®	(pagado)	0€
2x	Memoria DRAM 4 GB DD3 @ 1333MHz	50€	100€
1x	Memoria SSD de 250 GB de Crucial®	120€	120€
1x	Memoria HDD de 1 TB de Toshiba®	50€	50€
1x	Impresora OfficeJet™ 5740 de HP®	70€	70€
1x	Interfaz de audio Saffire™ Pro 40 de Focusrite®	469€	469€
1x	Cable Firewire IEEE 1394b de Roline®	10€	10€
2x	Monitor de estudio de 5'' LSR 305 de HARMAN-JBL®	150€	300€
2x	Cable XLR m/Jack m balanceado con conectores Neutrik®	26€	52€
2x	Cable extensión Jack m/h balanceado con conectores Neutrik®	38€	76€
1x	Micrófono omnidireccional XREF20 para mediciones acústicas de Sonarworks®	95€	95€
1x	Cable XLR m/h balanceado con conectores Neutrik®	20€	20€
1x	Auriculares Earpods de Apple®	30€	30€
1x	Auriculares Earphones MX270 de Sennheiser®	15€	15€
1x	Auriculares circumaurales HDJ 1500 de Pioneer®	175€	175€
1x	Calibración de auriculares HDJ 1500 por Sonarworks®	50€	50€
	2 Bibliografía		
1x	"Tratamiento de señales en tiempo discreto", Oppenheim, A.	50€	50€
1x	"Auralization", Vorländer, M.	143€	143€
1x	"Spatial Audio Processing", Breebaart, J. y Faller, C.	125€	125€
1x	"Sound System Engineering", Davis, C.	40€	40€
1x	"Master Handbook of Acoustics", Everest, A.	32€	32€
	3 Consumibles utilizados		
1x	Electricidad total consumida	400€	400€
12x	Conexión a internet por cable	25€/mes	300€
1x	Material de oficina (varios)	150€	150€
	SUBTOTAL:		**2 882€**
	COSTE RRHH		
13x	Sueldo ingeniero hardware y software (M. Ing. Telecomunicación), media jornada	1500€/mes	19500€
	COSTE TOTAL DEL PFC:		***22 382€***

B HOJAS DE CARACTERÍSTICAS

Las siguientes páginas contienen las hojas de características mas representativas de los componentes utilizados para la reproducción del audio 3D sintetizado mediante el software desarrollado.

Manual 1. Altavoces monitores de estudio activos. HARMAN-JBL® LSR 305 5''.
Manual 2. Auriculares. Pioneer® HDJ 1500-S, 32 Ohm.
Manual 3 y 4. Interfaz de audio. Focusrite® Saffire™ Pro 40.
Manual 5 y 6. Acondicionador de potencia. Furman® M-10x E.
Manual 7. Dispositivo para audio 3D. HeaDSPeaker®.

Section 8: Specifications

Specifications

	LSR305	LSR308	LSR310S
Frequency Range:	43 Hz – 24 kHz	37 Hz - 24 kHz	27 Hz
Crossover:	1725 Hz 4th order acoustic Linkwitz-Riley	1800 Hz 4th order acoustic Linkwitz-Riley	- - -
Maximum Peak SPL:	108 dB SPL *	112 dB SPL *	113 dB **
Maximum Peak Input Level: -10 dBV / +4 dBu	+6 dBV / +20.3 dBu	+6 dBV / +20.3 dBu	+6 dBV / +20.3 dBu
Input Connectors:	1 x XLR, 1 x TRS Balanced	1 x XLR, 1 x TRS Balanced	2 x XLR, 2 x TRS Balanced
Input Sensitivity: (-10 dBV input)	92 dB / 1m	92 dB / 1m	92 dB / 1m
HF Driver Size:	25 mm (1")	25 mm (1")	- - -
LF Driver Size:	127 mm (5")	203 mm (8")	250 mm (10")
HF Driver Power Amp:	41 W Class D	56W Class D	- - -
LF Driver Power Amp:	41 W Class D	56W Class D	200W Class D
HF Trim Control:	+2 dB, 0, -2 dB @ 4.4 kHz	+2 dB, 0, -2 dB @ 4.4 kHz	- - -
LF Trim Control:	+2 dB, 0, -2 dB @ 115 Hz	+2 dB, 0, -2 dB @ 115 Hz	- - -
AC Input Voltage:	100-240 VAC +/- 10% 50/60 Hz	100-240 VAC +/- 10% 50/60 Hz	100-240 VAC +/- 10% 50/60 Hz
Enclosure Construction:	15 mm (5/8 in) MDF	15 mm (5/8 in) MDF	18 mm (3/4 in) MDF
Enclosure Finish:	Matte Black PVC	Matte Black PVC	Matte Black PVC
Baffle Construction:	Injection-molded structural ABS	Injection-molded structural ABS	- - -
Baffle Finish:	Metallic Black Acrylic Paint	Metallic Black Acrylic Paint	- - -
Dimensions (H x W x D):	298 mm x 185 mm x 231 mm (11.75 in x 7.28 in x 9.88 in)	419 mm x 254 mm x 308 mm (16.5 in x 10.0 in x 12.1 in)	448 mm x 381 mm x 398 mm (17.65 in x 15.0 in x 15.65 in)
Weight:	4.6 kg (10.12 lbs)	8.6 kg (18.9 lbs)	15.6 kg (34.3 lbs)
Display Carton (H x W x D):	354 mm x 244 mm x 299 mm (13.93 in x 9.6 in x 11.77 in)	473 mm x 312 mm x 358 mm (18.6 in x 12.2 in x 14.0 in)	505 mm x 466 mm x 476 mm (19.9 in x 18.3 in x 18.7 in)
Shipping Carton (H x W x D):	373 mm x 260 mm x 315 mm (14.69 in x 10.22 in x 12.4 in)	491mm x 326 mm x 371 mm (19.3 in x 12.8 in x 14.6 in)	520 mm x 478 mm x 488 mm (20.5 in x 18.8 in x 19.2 in)
Shipping Weight:	6 kg (13.2 lbs)	10 kg (22 lbs)	19.1 kg (42 lbs)

* Full Bandwidth Pink Noise Measured C-Weighted
** Measured in Half Space

19

Manual 1 Altavoces monitores de estudio activos. HARMAN-JBL® LSR 305 5".

Pioneer DJ

HDJ-1500

Professional DJ Headphones

ENGLISH

Please read through these operating instructions so you will know how to operate your model properly.

⚠ WARNING

- Adjust headphones sound to the proper volume. Loud sound may damage your ears.
- For traffic safety, never use the headphones while riding a bicycle, motorbike, or while driving a car. It is dangerous to increase the sound volume too much because you cannot hear external sounds. Also, pay attention to traffic around you.

⚠ CAUTION

- Never inspect the inside OR remodel this machine. If the customer remodels this machine, Pioneer DJ will no longer guarantee its performance.

Precautions for use

- Caution should be observed during use, since adjustable parts may pinch the user's hair, depending on the way in which headphones are used.
- Do not subject the headphones to strong forces or impacts, since damage could occur to the exterior appearance or product performance.
- When headphones are dirty, wipe with a dry, soft cloth. Take care not to blow into the speaker unit.
- Sound quality might deteriorate or sound might be interrupted if the plug is dirty. Keep the plug clean by wiping it with a soft, dry cloth occasionally.
- Ear pads may deteriorate over long periods of use or storage. In this event consult your dealer.
- If any itching or skin disorder occurs during use, cease use immediately.
- This product is not edible. Place out of reach of small children to prevent accidental ingestion of small parts.

■SPECIFICATIONS

Type	Fully enclosed dynamic headphones
Impedance	32 Ω
Sensitivity	106 dB
Frequency response	5 Hz to 30 000 Hz
Maximum input power	3 500 mW
Driver units	ø30 mm dome type
Cord	1.2 m long one-side connection coiled type (extended length 3.0 m)
Plug	ø3.5 mm stereo mini plug (gold plated, threaded type)
Ear pad	Polyurethane leather finish
Weight	265 g (without cord)
Accessories	ø6.3 mm stereo plug adapter (gold plated, threaded type), Carrying pouch

NOTE

Specifications and design are subject to possible modification without notice, due to improvements.

Manual 2 Auriculares. Pioneer® HDJ 1500-S, 32 Ohm.

Saffire PRO 40 Specifications

MIC
- Frequency Response: 20Hz - 20kHz +/- 0.1 dB.
- THD+N: 0.001% (measured at 1kHz with a 20Hz/22kHz band pass filter).
- Noise: EIN > 125dB (128dB analogue to digital): measured at –60dB of gain with 150 Ohm termination (20Hz/22kHz band pass filter).

LINE
- Frequency Response: 20Hz - 20kHz +/- 0.1dB.
- THD+N: <0.001% (measured with 0dBFS equivalent input and 22Hz/22kHz band pass filter).
- Noise: -90dBu (22Hz/22kHz band pass filter).

INSTRUMENT
- Frequency Response: 20Hz - 20kHz +/- 0.1dB.
- THD+N: 0.004% (measured with 0dBu input and 20Hz/22kHz band pass filter).
- Noise: -87dBu (20Hz/22kHz band pass filter).

DIGITAL PERFORMANCE
- Clock Sources:
  - Internal clock.
  - Sync to word clock on S/PDIF (coaxial input).
  - Sync to word clock on ADAT input.
  - Sync to word clock on optical S/PDIF input (when enabled).
- A/D Dynamic Range 110dB 'A-weighted' (all inputs).
- D/A Dynamic Range 110dB 'A-weighted' (all outputs).
- JetPLLTM PLL technology providing superb jitter reduction, for class leading converter performance.
- Clock Jitter < 250 pico seconds.
- Sample rates: 44.1 to 96kHz.
- 20 input channels to computer: Analogue (8), S/PDIF (2), ADAT (8) and Mix Loop-back (2).
- 20 output channels from computer: Analogue (10), S/PDIF (2) and ADAT (8).
- Fully assignable 18 input by 16 output mixer.

WEIGHT and DIMS
- 3kg - 35cm x 4.5cm x 26.5cm.

ANALOGUE INPUTS
- Mic / Line inputs on XLR Combo with auto-switching between XLR and TRS.
- Mic / Line / Instrument 1 & 2: 2 x XLR Combo on front panel.
- Mic / Line 3-8: 6 x XLR Combo.
- Instrument: As above, switched to Instrument (inputs 1 & 2 only).
- Mic Gain: +10dB to + 55dB.
- Line 1-8 Gain: -10dB to +36dB.
- Instrument Gain: +10dB to +55dB.
- Input Pad on inputs 1-2, -9dB.
- Phantom power switched in 4 channel groups on Mic. 1-4 and 5-8.
- Mic and instrument maximum input level +7dBu (+16dBu with pad on inputs 1 and 2).
- Line maximum input level +22dBu.

ANALOGUE OUTPUTS
- Line level 10 x 1/4 inch TRS Jack.
- Outputs Monitor 1 and 2 have anti-thump protection circuitry.
- Nominal output level 0dBFS = 16dBu, balanced.
- Frequency Response: 20Hz – 20kHz +/- 0.2dB.
- THD+N <0.0010% (-100dB) (measured with 0dBFS input 22Hz/22kHz band pass filter, un-weighted).
- Software switched attenuation (–20dB) on outputs 1 and 2 (for sensitive active monitors).
- Hardware and Software Controlled Digital Volume control for all outputs (assignable through control panel).
- Hardware and Software Controlled Digital Dim and Mute controls for all outputs (assignable through control panel).
- All outputs are useable as monitoring outputs

26

Manual 3 Interfaz de audio. Focusrite® Saffire™ Pro 40 (1)

DIGITAL I/O
• S/PDIF input and output (RCA phono) on rear panel, (24-bit, 44.1 - 96kHz) Output transformer isolated.
• ADAT In / Out 8 channels (44.1 / 48kHz), 4 channels S-MUX (88.2 / 96kHz).
• ADAT Input / Output can be re-configured to be optical S/PDIF input / output via control panel.

MIDI I/O
• 1 in / 1 out on rear panel.

FIREWIRE S400
• 2 ports.

POWER
• Internal Universal Input PSU. 90-250Vac

HEADPHONE MONITORING
• 2 x 1/4 inch TRS Jack on front panel (mirrors outputs 7-8 and 9-10).
• High power headphone drivers.

FRONT PANEL INDICATORS
• Metering of analogue inputs (channels 1-8), 5 segment (-42, -18, -6, -3 and 0dBFS).
• 'Lock' Indicator.
• 'FW Active Indicator.
• MUTE switch and LED.
• DIM switch and LED.
• 48V switches and LEDs.
• Inst switches and LEDs.
• Pad switches and LEDs.
• Power switch and LED.

Troubleshooting

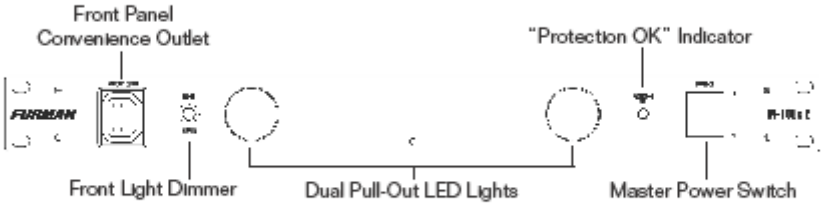
For all troubleshooting queries, please visit the Focusrite Answerbase where there are articles covering numerous troubleshooting examples. www.focusrite.com/answerbase.

E & O.E.

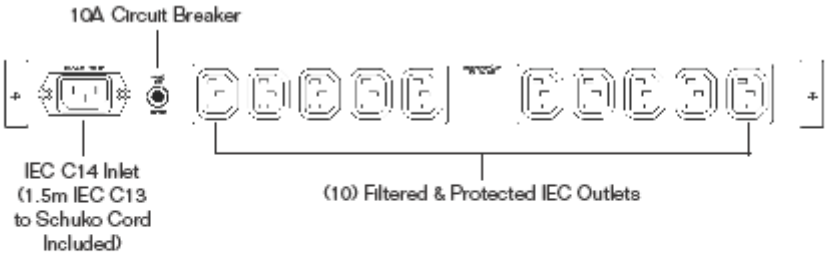
Manual 4 Interfaz de audio. Focusrite® Saffire™ Pro 40 (2).

M-10x E FRONT PANEL FEATURES

M-10Lx E FRONT PANEL FEATURES



MERIT E SERIES REAR PANEL FEATURES



6

Manual 5 Acondicionador de potencia. Furman® M-10x E (1)

SPECIFICATIONS

Maximum Output Current:
10 amps

Line Cords:
1.5 meter, Removable, IEC C-13 Female to Schuko Male

Lamps:
LED

Operating Voltage:
230VAC 50HZ

Spike Protection Modes:
Line to Neutral

Energy Dissipation:
305 Joules

Peak Impulse Current:
12,000 amps

Maximum Clamping Voltage:
650 Volts

Noise attenuation (Transverse Mode):
Greater than 20DB, 1.5MHz to 200MHz

Dimensions:
1.75" H x 19" W x 7.5" D

Weight:
2.3 Kg.

Construction:
Steel Chassis, Black Painted

Manual 6 Acondicionador de potencia. Furma® M-10x E (2)

HeaDSPeaker *Home* User's Manual

Revision History

Rev.	Date	Author	Affected chapters	Description
1.0	4.10.2010	LM, POj	All	Original version Finnish and English

Manual 7 Dispositivo para audio 3D. HeaDSPeaker®.

C PLIEGO DE CONDICIONES

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un **Sistema de audio 3D con auriculares**. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma,

por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad "Presupuesto de Ejecución de Contrata" y anteriormente llamado "Presupuesto de Ejecución Material" que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.