# TÉCNICAS PARAMÉTRICAS DE UPMIXING EN AMBISONICS: EVALUACIÓN PERCEPTUAL

Grado en Ingeniería
en Tecnologías de Telecomunicación

Trabajo Fin de Grado

Autora: Alicia Pérez García

Tutor: Ricardo San Martín Murugarren

Pamplona, Junio de 2021

upna Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

## Resumen

Ambisonics es un método para la reproducción inmersiva de audio espacial con ventajas técnicas relacionadas con la interactividad y la realidad virtual. Del mismo modo, también se conocen deficiencias en la resolución espacial cuando se utilizan grabaciones de primer orden. Utilizar órdenes superiores corrige estas deficiencias a costa de una mayor complejidad técnica y esfuerzo económico en los micrófonos que se utilizan.

Cuando sólo se dispone de material en primer orden, pueden utilizarse estrategias de upmixing para aumentar la resolución espacial y el *sweet spot* o punto óptimo. Existen diferentes estrategias, y son estas las que se van a evaluar en este proyecto utilizando una esfera completa de 24 altavoces. Dado que las señales de los altavoces pueden convertirse en señales binaurales por medio de altavoces virtuales, las tres estrategias (DirAC, HARPEX y COMPASS) se comparan también reproduciendo las escenas a través de auriculares.

El objetivo de este proyecto es presentar y analizar mediante una prueba de audio llevada a cabo en la UPNA los diferentes algoritmos de upmixing utilizados para convertir señales de Ambisonics de primer orden a señales Ambisonics de tercer orden por medio de rutinas de software.



## Lista de palabras clave

Ambisonics, listening test, MUSHRA, upmixing, binaural, HARPEX, COMPASS, DirAC.

# ÍNDICE

# 1. Ambisonics

## 1.1 Características principales

[1] Ambisonics (del latín *ambire*: rodear, ir alrededor, y *sonus*: sonido) es un formato para codificar el campo sonoro teniendo en cuenta sus propiedades espaciales. Se basa en que el campo sonoro en un punto (conjunto de sonidos que nos rodean en un espacio delimitado) puede representarse como una superposición de ondas planas. Su primera versión (First Order Ambisonics o FOA) fue desarrollada en 1970 por P. Felgett [17], M.A. Gerzon [18] y P. Craven [19], y en la década de los 90 se extiende su desarrollo utilizando órdenes más altos que mejoran la precisión espacial, naciendo el formato HOA (Higher Order Ambisonics).

En los formatos de audio multicanal tradicional (p.e. estéreo, envolvente 5.1 y 7.1), cada canal contiene la señal correspondiente a un altavoz determinado. Sin embargo, en Ambisonics cada canal lleva información sobre determinadas propiedades físicas del campo sonoro, como la presión, la velocidad acústica o sus derivadas de orden superior.

Ambisonics permite trabajar con la espacialización del sonido en dos y tres dimensiones enlazando dos etapas independientes: una de codificación y otra de decodificación. La información espacial está codificada en la propia señal de audio, siendo la señal portadora de la información espacial y la señal sonora la misma, por lo que no se registra información en forma de metadatos más allá de la señal de audio per se.

De manera similar a las series de Fourier, Ambisonics discretiza espacialmente el campo sonoro utilizando los denominados armónicos esféricos, de forma que se puede reconstruir el campo sonoro original mediante la combinación de sus componentes individuales. La cantidad de armónicos empleados está relacionada con el orden del formato Ambisonics.

1. Orden cero: Ambisonics contiene información sobre el campo de presión en el origen. Equivale a la grabación de un micrófono omnidireccional en el origen. El canal para el campo de presión se llama convencionalmente W.
2. Primer orden: Ambisonics agrega información sobre la velocidad acústica en el origen. Equivale a la grabación de tres micrófonos figura de 8 a lo largo de cada uno de los ejes. Estos canales se denominan X, Y, Z. Siguiendo la ecuación de Euler, el vector de velocidad es proporcional al gradiente de la presión sonora a lo largo de cada uno de los ejes.
3. Segundo orden y órdenes superiores: Ambisonics agrega información acerca de las derivadas de orden superior del campo de presión.

Así pues, a mayor cantidad de armónicos, mayor es el orden y mayor es la resolución espacial obtenida. En 3D, el orden N de Ambisonics y los armónicos guardan la relación expresada en la *ecuación 1.*

$$Armónicos = (N+1)^2$$

$Armónicos = 2N + 1$ si la reproducción es en dos dimensiones.

*Ecuación 1*

Dentro de Ambisonics, el formato AmbiX es el estándar actual que determina el orden de los canales y su normalización. Debido a una fuerte correspondencia entre los armónicos esféricos y los patrones polares de directividad y al hecho de que esos patrones polares tienen direcciones claramente definidas, resulta útil ordenar y nombrar los componentes con referencia a los ejes en un sistema siguiendo la regla de la mano derecha o del sacacorchos.

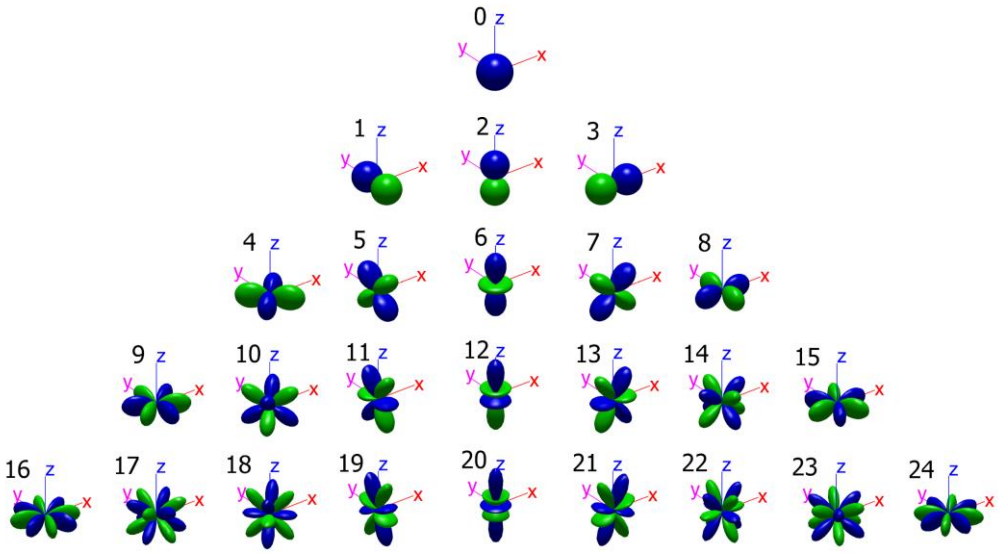



*Figura 1 – Armónicos esféricos para cada orden*

[4] Los armónicos en AmbiX se numeran siguiendo la convención numérica ACN (Ambisonics Channel Number), comenzando con el canal 0 (*figura 2*).

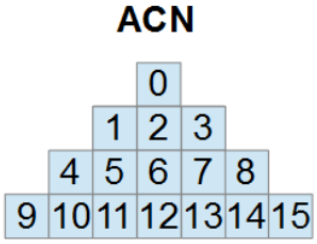



*Figura 2 – Notación ACN*

También existe la convención FuMa (Furse-Malham), que utiliza una notación con letras (*figura 3*) pero tiene un uso más restringido ya que sólo sirve para órdenes menores. Los órdenes 2 y 3 comienzan con su miembro simétrico z-rotacional y después saltan hacia afuera, derecha e izquierda, con los componentes horizontales al final. La diferencia con el ACN es que los canales ya están ordenados alfabéticamente (Para FuMa: WXYZ y para ACN: WYZX).

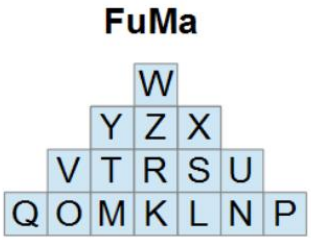



*Figura 3 – Notación FUMA*

Se dice que un flujo Ambisonics es de orden n cuando contiene toda la señal de órdenes 0 a n. Por ejemplo, una señal Ambisonics de orden 3 contiene 16 canales:

- 1 de orden 0
- 3 de orden 1
- 5 de orden 2
- 7 de orden 3

## 1.2 FOA (First Order Ambisonics)

El formato general de Ambisonics de primer orden o B-format, consiste en 4 señales W, X, Y y Z donde:

- W representa la salida de un micrófono de presión (se obtiene la presión de un punto).
- X, Y y Z son canales que representan la salida de micrófonos de gradiente de presión (se obtiene la diferencia de presión entre dos puntos), colocados en los ejes de un espacio tridimensional, con su cara positiva apuntando hacia el frente, izquierda y arriba, respectivamente.

La información registrada por estos canales representa una aproximación al campo sonoro envolvente a partir de sus armónicos esféricos de primer orden, como muestra la *figura 4.* Esta distribución de canales permite que el sonido pueda percibirse como una completa esfera de sonido [3].

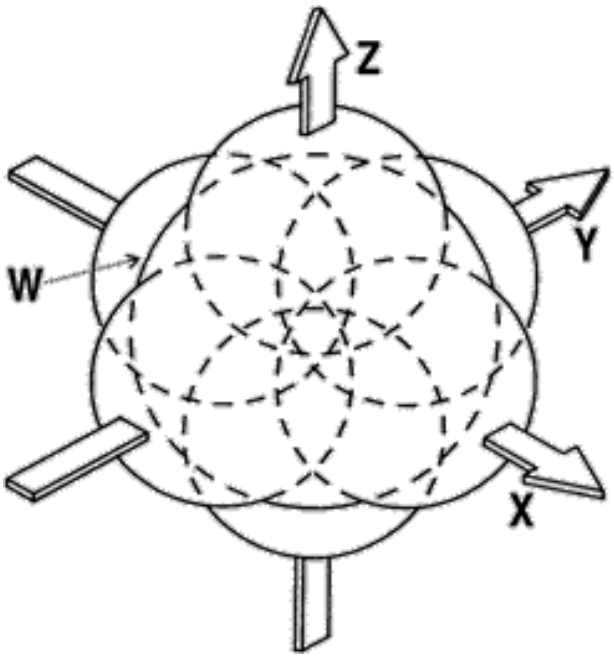



*Figura 4 - Disposición de señales W, X, Y, Z*

Las ecuaciones de estos patrones son las siguientes:

$$W = \frac{1}{\sqrt{2}}$$

$$X = cos\phi cos\delta$$

$$Y = sin\phi cos\delta$$

$$Z = sin\delta$$

Donde $\phi$ es el ángulo de azimut, que codifica la dirección de incidencia de la onda sonora en el plano horizontal (plano XY o plano paralelo al suelo) (*figura 5*), y $\delta$ es el ángulo de elevación, que es aquel formado entre el plano horizontal y la línea visual de un observador hasta el objeto (*figura 5*) y que codifica la dirección de incidencia de la onda sonora en el plano medio, que es el plano YZ que define la distinción entre derecha e izquierda.

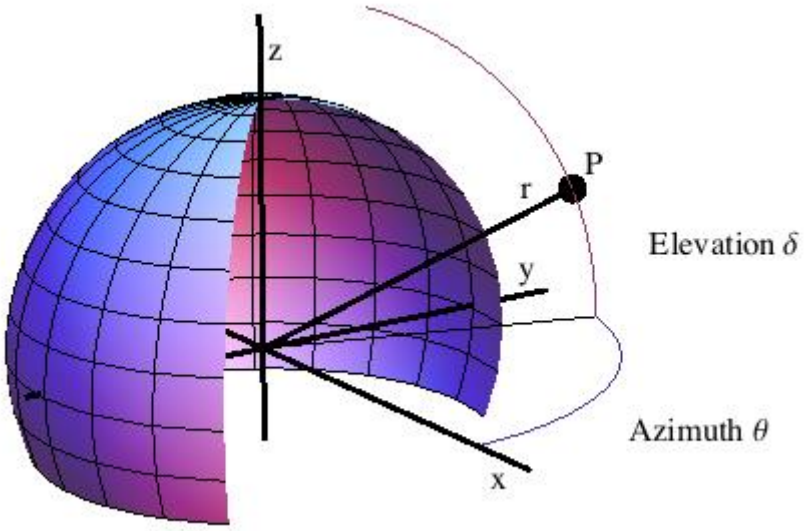



*Figura 5 – Ángulos azimut y elevación*

### 1.2.1 Codificación

[1] Como primer paso en la cadena Ambisonics, un campo sonoro debe codificarse, bien sintetizándolo a partir de una grabación mono (un canal), o grabándolo, utilizando un micrófono especial. En cada uno de los cuatro canales se registra entonces una señal monofónica que corresponde al armónico esférico de orden cero, más tres armónicos esféricos de primer orden, respectivamente.

Sea s(t) una señal mono procedente de una fuente sonora, W(t) el canal proporcional al campo de presión en el origen, y los canales X(t), Y(t) y Z(t) los proporcionales a la velocidad acústica en cada uno de los ejes, para un conjunto general de múltiples fuentes localizadas (múltiples ondas planas) con señales $S_i$ provenientes de la dirección ($\phi_i$, $\delta_i$), dichos canales Ambisonics se pueden calcular utilizando la *ecuación 2.* De esta manera es posible generar artificialmente una señal Ambisonics a partir de una grabación mono, posicionando esa grabación en cualquier dirección angular definida por su azimut y elevación.

$$W(t) = \sum_i s_i(t)/\sqrt{2}$$
$$X(t) = \sum_i s_i(t)\cos\phi_i\cos\delta_i,$$
$$Y(t) = \sum_i s_i(t)\sin\phi_i\cos\delta_i,$$
$$Z(t) = \sum_i s_i(t)\sin\delta_i.$$

*Ecuación 2*

En Ambisonics también es posible codificar otro tipo de ondas distintas a las ondas planas, sin embargo, la codificación es más complicada. En cualquier caso, trabajar con ondas planas suele ser suficiente porque cualquier fuente puede codificarse en términos de ondas planas múltiples, y una onda esférica, a una distancia del origen de varias longitudes de onda puede aproximarse con buena precisión en términos de ondas planas.

### 1.2.2 Grabación

[1,2] Después de que Cooper y Shiga [20] trabajaran en la expresión de estrategias de panoramización para configuraciones arbitrarias de altavoces envolventes en términos de una serie direccional de Fourier, Felgett [17], Gerzon [18] y Craven [19] desarrollaron la noción y tecnología de Ambisonics.

Con ello surgió el concepto de Ambisonics 2D de primer orden (FOA o First-Order Ambisonics), que consta de una señal correspondiente a un patrón de captación omnidireccional (llamado W) y dos señales correspondientes a los patrones de captación en figura de ocho alineados con los ejes cartesianos (X e Y).

Para grabar los canales W, X e Y en 2D FOA, se puede utilizar la técnica MS (*figura 6.1, figura 6.2*). Esta consiste en utilizar dos micrófonos de diferente polaridad colocados en un ángulo recto de 90º, donde el micrófono central (generalmente cardioide) registra una señal central (M), y el micrófono lateral bidireccional (figura de ocho) una señal lateral (S).

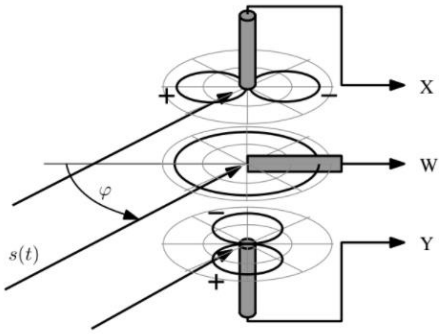



*Figura 6.1 – Grabación de FOA en 2D*

*Figura 6.2 - Configuración de grabación 2D FOA*

La grabación de FOA en 2D también se puede hacer mediante tres cardioides en ángulo de 120º (*figuras 7.1 y 7.2*).

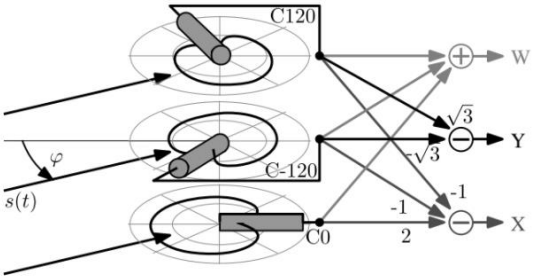



*Figura 7.1 – FOA 2D con 3 micrófonos cardioides*

*Figura 7.2 – Configuración de grabación*

El esquema se amplía a Ambisonics 3D de primer orden mediante un tercer micrófono en figura de ocho que apunta hacia arriba y hacia abajo tal y como muestra la *figura 8.1*, de manera que consiste en una señal W correspondiente a un patrón de captación omnidireccional y tres señales (X, Y y Z) correspondientes a patrones de captación en figura de ocho alineados con los ejes de coordenadas cartesianos.

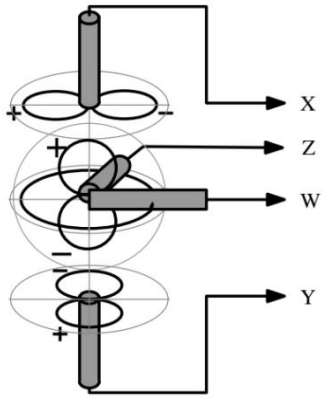



*Figura 8.1 – Grabación de FOA en 3D*

*Figura 8.2 – Configuración de grabación 3D FOA*

También es posible realizar la grabación de Ambisonics 3D mediante una disposición tetraédrica de micrófonos con patrón cardioide (*figura 9.1, figura 9.2*), colocados de manera que cubran un área en forma esférica. Este tipo de configuración conlleva la creación de dispositivos que contienen los cuatro micrófonos integrados creando un formato propio de grabación, usualmente denominado formato A. Existen micrófonos como el micrófono Soundfield utilizado en este proyecto, que coloca cuatro cápsulas cardioides o subcardioides en los vértices del tetraedro, tal y como se puede observar en la *figura 10.*

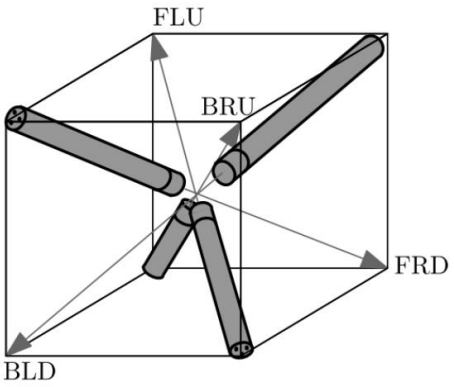



*Figura 9.1 – Configuración tetraédrica con cuatro cardioides*

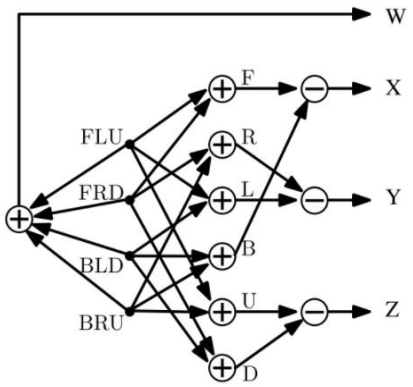



*Figura 9.2 Codificador de señales de micrófono*


*Figura 10 – Micrófonos Soundfield*

Para cada micrófono de campo sonoro se pueden distinguir tanto el A-format (grabación en bruto de cada una de las cuatro cápsulas de micrófono), como el B-format (componentes de los canales Ambisonics W, X, Y y Z).

Dado que para cualquier micrófono cardioide se sostiene que:

$$\text{Micrófono cardioide} = \frac{1}{2}\,(\text{Micro omnidireccional} + \text{micrófono figura de 8}),$$

se puede utilizar un sistema lineal de ecuaciones para convertir las grabaciones en formato A en señales Ambisonics en formato B. Además, es necesario agregar algunos filtros para corregir el hecho de que las cuatro cápsulas no son exactamente coincidentes, lo cual es más relevante conforme aumenta la frecuencia.

Para algunos Soundfields, el propio preamplificador de micrófono es quien realiza la conversión de formato A a B en el hardware (*figura 11*), de modo que el preamplificador ya tiene 4 salidas etiquetadas como W, X, Y, y Z. Sin embargo, para otros, el preamplificador de micrófono entrega la salida de las cápsulas en A-format y la conversión a B-format debe realizarse externamente (normalmente mediante ordenador).

*Figura 11 - Micrófono Soundfield TSL que realiza la conversión de A-format a B-format*

El formato Ambisonics de primer orden sigue siendo la base de las aplicaciones de realidad virtual y las transmisiones de audio de 360º en Internet de hoy en día. Además de la reproducción potencial por altavoces, permite la reproducción interactiva en auriculares con seguimiento de la cabeza (head-tracking) para evitar que la escena del sonido acústico resulte estática para el oyente. La grabación Ambisonics de primer orden FOA tiene la ventaja de que se puede realizar con una pequeña cantidad de micrófonos de alta calidad. Sin embargo, la distribución de grabaciones Ambisonics de primer orden a los altavoces de reproducción no suele ser convincente sin ir a órdenes superiores y mejoras direccionales.

### 1.2.3 Decodificación

[1] Para la presentación de Ambisonics mediante altavoces distribuidos alrededor del oyente, es necesario un proceso de decodificación. Dado un conjunto de N altavoces, la decodificación alimenta a cada altavoz $i$, ubicado en una dirección $\hat{u}_i(\phi_i, \delta_i)$, con una señal $s_i$ dada por la *ecuación 3.*

$$s_i(t) = w_i W(t) + x_i X(t) + y Y(t) + z_i Z(t)$$

*Ecuación 3*

donde el conjunto de 4N parámetros ($w_i$, $x_i$, $y_i$, $z_i$) definen la decodificación. Para inferir estos parámetros se pueden adoptar dos estrategias básicas: reconstrucción básica del campo sonoro o reconstrucción psicoacústica.

### a) Decodificación básica

[5] En la decodificación básica, los coeficientes se determinan bajo el supuesto de coherencia entre las señales que llegan de los altavoces. El requisito es reproducir en el origen la presión y el vector velocidad acústica con precisión.

Suponiendo que todos los altavoces están a la misma distancia del origen, y asumiendo que todos los altavoces se suman coherentemente, la presión en el origen es (descartando el factor de decaimiento, que se supone que es igual para todos los altavoces):

$$p(t) = \sum_{i=1}^{N} s_i\,(t)$$

y la velocidad acústica normalizada es:

$$\vec{v}(t) = \sum_{i=1}^{N} s_i\,(t) \cdot \hat{u}_i$$

El coeficiente de direccionalidad de la velocidad acústica, $r_v$, puede definirse como

$$r_v = \frac{||\vec{v}||}{|p|}$$

Para una onda plana (una fuente completamente localizada) $r_v = 1$, mientras que para una fuente completamente deslocalizada (por ejemplo, campo difuso) $r_v = 0$. En general, $r_v \in [0, \infty)$ ($r_v \to \infty$ por ejemplo, para una onda estacionaria en un nodo de presión).

Esta estrategia de decodificación reproduce la impresión del sonido original a bajas frecuencias (por debajo de 500 Hz aproximadamente), y a distancias del centro no mayores que una fracción de la longitud de onda considerada más corta.

Para decodificar un sistema Ambisonics con esta estrategia se resuelve un sistema de ecuaciones lineales con 4 ecuaciones escalares y N incógnitas, que son las señales $s_i(t)$ alimentadas a cada uno de los altavoces, teniendo en cuenta que en el origen el valor de la presión y la velocidad sea el que corresponde a W(t) y (X(t), Y (t), Z (t)) respectivamente.

Siempre que los altavoces estén bajo configuraciones razonables, esta ecuación tiene una solución única con N = 4. Con N > 4, el sistema presenta múltiples soluciones diferentes (eligiendo finalmente la solución con menos energía global).

Generalmente para diseños regulares con N > 4, $s_i(t)$ presenta la forma de la *ecuación 3* con:

$$w_i = \frac{\sqrt{2}}{N}$$

$$x_i = 3cos\phi_i cos\delta_i / N$$

$$y_i = 3sin\phi_i cos\delta_i / N$$

$$x_i = 3sin\delta_i / N$$

Y para ecuaciones de decodificación en 2D

$$w_i = \frac{\sqrt{2}}{N}$$

$$x_i = 2cos\phi_i / N$$

$$y_i = 2sin\phi_i / N$$

Teniendo en cuenta que W(t) es la señal registrada por un micrófono omnidireccional, y que X(t), Y(t), Z(t) son las señales grabadas por tres micrófonos figura de ocho, la combinación de estos cuatro micrófonos es también un micrófono, en este caso un micrófono supercardioide apuntando en la dirección del altavoz (*figura 12*).

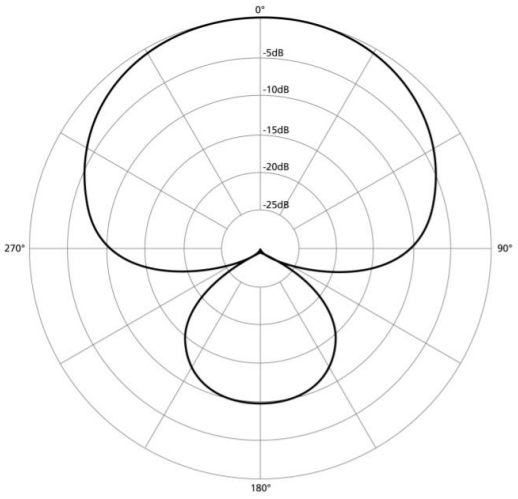



*Figura 12 - Ejemplo de un patrón polar de un micrófono supercardioide*

La decodificación básica solo funciona bien para bajas frecuencias (hasta 500 Hz aproximadamente) y muy cerca del centro de reproducción.

## b) Decodificación psicoacústica

[5] En las decodificaciones psicoacústicas, se asume una suma incoherente de las señales. El requisito necesario es que la decodificación reproduzca la energía original y la intensidad acústica en el origen.

Dentro de la hipótesis de la suma incoherente, y asumiendo que cada una de las ondas entrantes es una onda plana, la energía de la señal normalizada en el origen es:

$$w(t) = \sum_{i=1}^{N} |s_i(t)|^2$$

y el vector de energía es:

$$\vec{E}(t) = \sum_{i=1}^{N} |s_i(t)|^2 \hat{u}_i$$

De donde el coeficiente de direccionalidad de la energía se define como

$$R_E = \frac{||\vec{E}||}{w}$$

Para una onda plana (una fuente completamente localizada) $R_E$= 1, mientras que para una fuente completamente deslocalizada (por ejemplo, un campo difuso) $R_E$ = 0.

Para ondas planas reproducidas, es físicamente imposible cumplir la condición $R_E$= 1 sumando incoherentemente la señal de varios altavoces.

También existe una decodificación llamada max-$R_E$ en la que se intenta maximizar este valor (de ahí el nombre). En este caso las ecuaciones son mucho más complicadas de resolver porque no son lineales. Sin embargo, cuando los altavoces están en un diseño regular (sólido platónico o similar), se puede encontrar una solución similar a la decodificación básica. Esta decodificación funciona bien para frecuencias medias a altas (a partir de 500 Hz) o cuando la audiencia está lejos del punto óptimo.

Existe otra estrategia de decodificación similar a la decodificación max-$R_E$, pero con la restricción adicional de que los diferentes altavoces no emitan en oposición de fase simultáneamente. Proporciona una localización más sólida para los oyentes que están lejos del punto central (sweet spot), que es donde se coloca el micrófono durante la grabación, o donde se encuentra el denominado punto óptimo del oyente durante la reproducción. En la denominada decodificación in-phase, las propiedades de direccionalidad de en fase son algo peores que las de max-$R_E$ (Para 3D,$R_E$= 0,5 y para 2D $R_E$= 0,677), aunque funciona mejor con audiencias extendidas, donde los altavoces que emiten en oposición de fase pueden resultar una distracción.

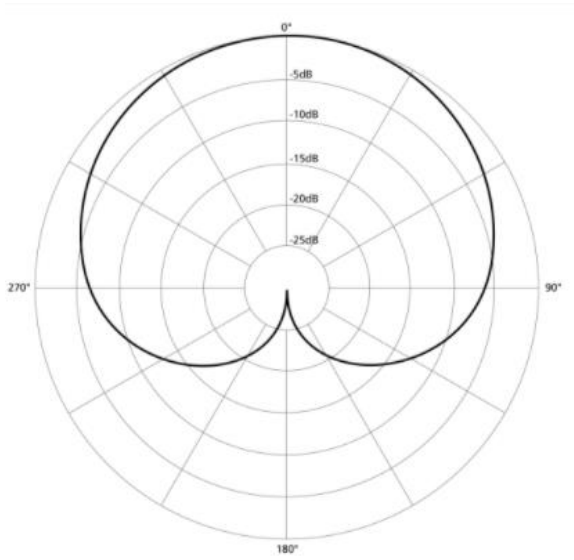



*Figura 13 – Patrón polar de un micrófono cardioide*

A veces se usa una estrategia de decodificación mixta, donde para bajas frecuencias se usa la decodificación básica, y para altas frecuencias se emplea una decodificación en fase o max-$R_E$, realizando el cruce a través de un par de filtros.

## 1.3 HOA (Higher Order Ambisonics)

Uno de los principales inconvenientes de Ambisonics es su pobre resolución espacial. Ambisonics de primer orden (FOA) codifica cualquier campo sonoro en términos de grabación de un micrófono omnidireccional (orden cero) y 3 micrófonos figura de ocho (primer orden). HOA extiende la expansión en términos de micrófonos de orden superior para mejorar la direccionalidad del formato.

HOA descompone el campo de sonido en términos de la grabación de un conjunto de micrófonos llamados armónicos esféricos, cuya notación es $Y_{lm}(\phi, \delta)$, donde $l$ es el orden Ambisonics, m = $-l,..,\ l$ indica el coeficiente particular, y ϕ y δ son los ángulos de azimut y elevación respectivamente.

El orden máximo en el que se realiza la expansión constituye el orden HOA. Cada orden $l$ tiene $(2l+1)$ canales. En total, Ambisonics de orden $l$ tiene $(l+1)^2$ canales. En 2D, el orden $l$ de Ambisonics tiene $(2l+1)$ canales. Sin embargo, esto tiene el coste de un mayor ancho de banda, ya que el número de canales aumenta cuadráticamente con el orden.

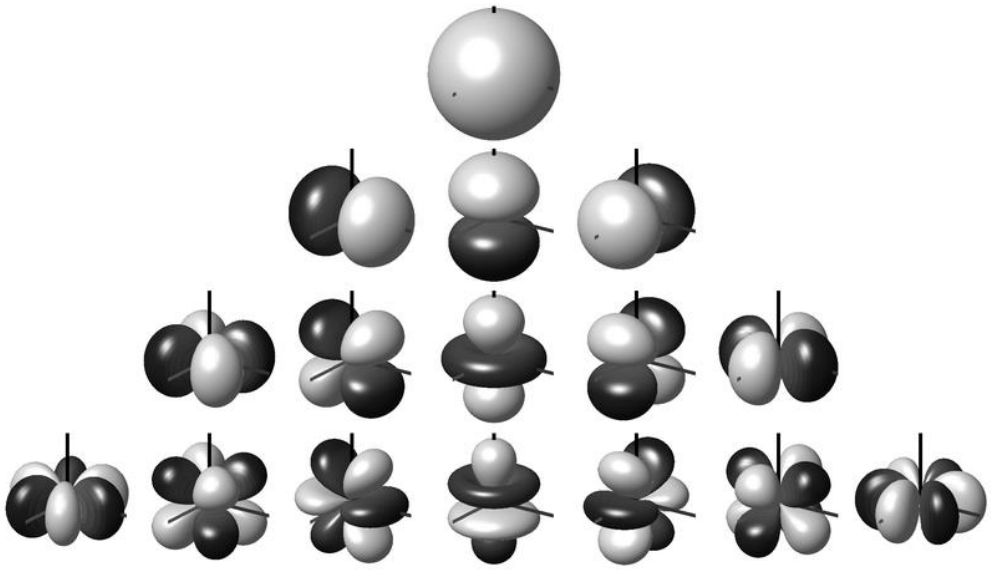


*Figura 14 – Patrones polares 3D de armónicos esféricos hasta tercer orden*

Cualquier distribución de fuentes se puede expresar en términos de campos de sonido de la siguiente manera:

$$S(t;\varphi,\delta) = \sum_{l=0}^{\infty} B_{lm}(t)\, Y_{lm}(t)\, (\phi,\delta)$$

donde $B_{lm}$(t) corresponde a los canales HOA y contiene W, X, Y y Z, y los armónicos esféricos $Y_{lm}(\phi$, δ) son la base de la expansión.

### 1.3.1 Codificación

[2] Dada una onda plana de una señal s(t), la codificación es $B_{lm}(t) = s(t)\, Y_{lm}(\phi_s, \delta_s)$.

Según la convención normalizada, los coeficientes de la expansión son precisamente los armónicos. Cualquier distribución de fuentes puede crearse con múltiples ondas planas.

El diagrama de bloques de la codificación HOA se muestra en la *figura 15*. El primer paso del procesamiento consiste en descomponer las muestras de presión p(t) de la matriz de micrófonos en sus coeficientes armónicos esféricos ψN (t). Con esto se obtiene la cantidad de las muestras que contienen patrones armónicos omnidireccionales, figura de ocho y otros patrones armónicos esféricos al que la disposición del micrófono permite la descomposición. Las muestras de presión sonora p(t) son armónicos esféricos descompuestos por la matriz $Y_N^T$ †, y las señales de coeficiente ψN(t) resultantes se convierten en señales ambisónicas $\chi_N$(t) mediante los filtros de agudizamiento ρn(ω).

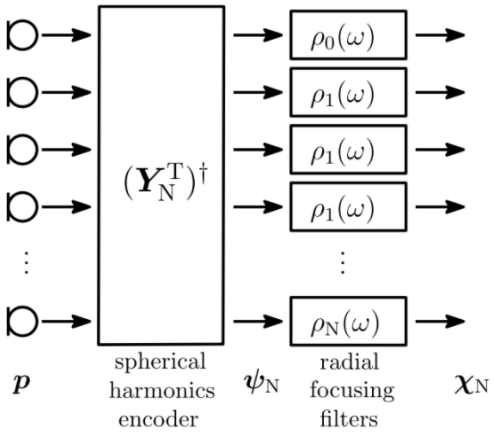



*Figura 15 – Codificación de micrófono HOA*

### 1.3.2 Grabación

Hay micrófonos que pueden grabar directamente en Ambisonics de orden superior, como el micrófono Eigenmike (figura 16), que consta de una esfera rígida de 32 cápsulas, cuya salida se puede convertir a HOA.


*Figura 16 – Micrófono em32 Eigenmike de 32 elementos [6]*

### 1.3.3 Decodificación

La decodificación de HOA procede de manera similar a los Ambisonics de primer orden. Hay dos posibilidades de decodificación: reconstruir los armónicos esféricos en el origen o intentar entregar una decodificación psicoacústica que maximice las propiedades de direccionalidad de la energía. Decodificar a diseños regulares es relativamente sencillo, pero decodificar a diseños irregulares es un que sigue en investigación: uno debe aplicar métodos algebraicos (decodificación pseudoinversa, básica) o métodos de búsqueda no lineales (decodificaciones psicoacústicas), o métodos de aproximación o combinaciones de estos.

De manera similar a los Ambisonics de primer orden, el resultado de la decodificación en diseños regulares es que cada altavoz emite la señal de un micrófono virtual apuntando en la dirección del altavoz.

Para la decodificación básica y Max-$R_E$, este micrófono virtual es supercardioide; para la decodificación en fase, el micrófono virtual es cardioide, como se puede observar en la *figura 17,* donde cada micrófono reproduce la señal de un micrófono virtual de orden superior que apunta en la dirección de ese altavoz.

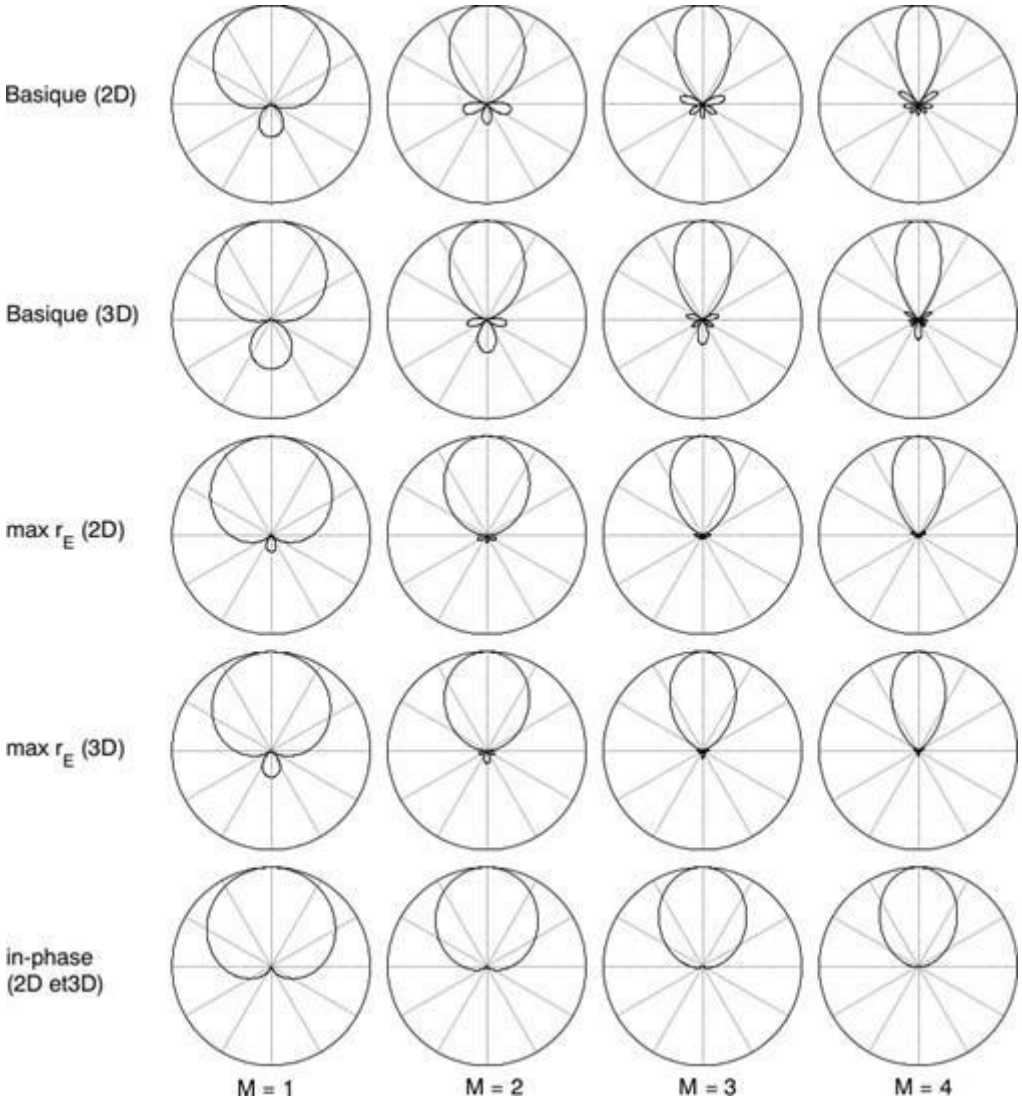



*Figura 17 - Micrófonos virtuales equivalentes utilizados para la reproducción de HOA*

Las propiedades de direccionalidad de HOA son mejores que las de primer orden Ambisonics, ya que esa es la razón principal para usar HOA. La *tabla 1* muestra los mejores coeficientes de direccionalidad de HOA en 3D dependiendo de la decodificación.

| Orden | Max-$R_E$ | En fase |
|---|---|---|
| 1 | 0,577 | 0,500 |
| 2 | 0,775 | 0,567 |
| 3 | 0,861 | 0,750 |
| 4 | 0,906 | 0,800 |

*Tabla 1 – Valores para el coeficiente de direccionalidad $R_E$ en HOA según el orden*

# 2. Test subjetivos

Los test subjetivos de audio permiten entender cómo perciben, procesan y responden las personas al audio. En este proyecto se van a analizar dos tipos de test o pruebas para calificar la percepción del sonido en las personas: los *grading test* y los test psicofísicos.

## 2.1. Grading test

**Grading test:** Son pruebas que sirven para evaluar las cualidades de sonido percibidas de los sistemas de audio utilizando escalas de calificación. Se puede establecer una relación entre la calidad del sonido percibido y los parámetros físicos asociados.

Antes de diseñar un grading test hay que tener en cuenta los siguientes factores:

1. Variables experimentales e hipótesis.
2. Tipo de diseño: intersujeto o intrasujeto.
3. Tipo de escala a utilizar para hacer la calificación: semántica o continua.
4. Tipo de método: múltiple estímulo o de estímulo único.
5. Tipo y número de sujetos.

### 2.1.1 Variables experimentales

En el diseño de los grading test es necesario definir las variables dependientes (DV) e independientes (IV) del proceso, entendiendo la variable dependiente como aquello que se mide y la independiente como aquello que se manipula. Es necesario valorar qué efecto produce la variable independiente sobre la variable dependiente y si dicho efecto es significativo o no. Un efecto estadísticamente significativo se entiende como la probabilidad de que la relación entre dos o más variables en un análisis no sea pura coincidencia, sino que sea causada por otro factor. Si lo es, además hay que evaluar en qué condiciones hay diferencias significativas.

La significancia estadística de la variable independiente (IV) sobre la variable dependiente (DV) se determina mediante pruebas de hipótesis. En estadística, una hipótesis es una suposición sobre la relación entre un conjunto de datos o una proposición acerca de una característica de la población de estudio. El resultado de una prueba de hipótesis permite determinar si es interesante evaluar dicha suposición o no.

Las pruebas de hipótesis se basan en dos hipótesis:

- Hipótesis inicial u original a evaluar.
- Hipótesis nula (H0): contrario de la hipótesis inicial. Supone que dos o más parámetros no tienen relación entre sí.

Si la probabilidad de la hipótesis nula de ser cierta es menor que un cierto porcentaje entonces H0 es rechazada. Por lo tanto, inversamente, también puede concluir que la probabilidad de que la hipótesis original sea cierta es muy alta.

El p-valor es la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta y determina si aceptar H0 o no. Cuando un p-valor es menor que o igual al nivel de significancia, se rechaza la hipótesis nula. Por convenio suele establecerse que si el p-valor es inferior al 5% (0,05) es lo suficientemente improbable que se deba al azar como para rechazar con una seguridad razonable la H0 y afirmar que la diferencia es real. Si es mayor del 5%, no se tendrá la confianza necesaria como para poder negar que la diferencia observada sea debida al azar.

### 2.1.2 Tipo de diseño

Para determinar el tipo de método estadístico más apropiado a utilizar, es necesario distinguir entre el tipo de diseño y de datos. Dependiendo del tipo de análisis (paramétrico o no paramétrico) conviene aplicar un tipo de diseño u otro.

En los análisis paramétricos se hacen suposiciones sobre los parámetros de la distribución de población de la que se extrae la muestra. Suelen suponer que los datos se distribuyen de manera normal (la variable dependiente). Por otro lado, en los análisis no paramétricos los datos están libres de distribución y las variables se consideran no normales.

Hay 3 tipos diferentes de diseño: inter-sujeto, intra-sujeto y mixto.

#### 2.1.2.1 Diseño inter-sujeto

Se comparan dos o más grupos separados. Los resultados se comparan entre ambos grupos y el objetivo es ver si un tratamiento es mejor que el otro. Para cada tema se recopila una puntuación y la puntuación de cada sujeto se promedia con los demás sujetos de su grupo de tratamiento. Finalmente se comparan las puntuaciones medias de cada uno de los grupos para ver si un tratamiento es más eficaz que el otro. Se necesita un número grande de sujetos.

**a) Análisis paramétrico**

Si se quiere analizar la media de 2 grupos la prueba más adecuada es el T-test (prueba t). La prueba se basa en el estadístico t, que asume que la variable tiene una distribución normal, la media es conocida y la varianza de la población se calcula a partir de la muestra.

Por otro lado, para determinar si 3 o más poblaciones son estadísticamente independientes entre sí, el método más apropiado es el análisis de varianza o ANOVA. Consiste en determinar, dados un conjunto de supuestos, si existen diferencias estadísticamente significativas entre las medias de 3 o más grupos mediante el uso de varianzas. La varianza inter-grupo (varianza entre las medias del grupo) se compara con

la varianza intra-grupo (varianza dentro de los grupos) para determinar si los grupos difieren más entre sí que dentro de sí.

**b) Análisis no paramétrico**

Se aplica la prueba U de Mann-Whitney si existen dos variables independientes para contrastar si dos muestras proceden de poblaciones igualmente distribuidas. En cierto modo es el equivalente no paramétrico del t-test para la comparación de medias de dos distribuciones.

Por otro lado, el método de Kruskal-Wallis se utiliza para probar si un grupo de datos proviene de la misma población. Intuitivamente es idéntico al ANOVA reemplazando los datos por categorías. Es una extensión de la prueba de Mann-Whitney para 3 o más grupos.

#### 2.1.2.2 Diseño intra-sujeto

En este tipo de diseño, todos los test subjetivos se repiten en cada sujeto para después comparar puntuaciones. Se contempla un solo grupo de sujetos al que se le mide una o más variables de forma repetida. Esto proporciona tantos conjuntos de información como condiciones para cada participante. Es el más común en audio y es más eficiente que el diseño inter-sujeto cuando hay muchas condiciones para evaluar.

**a) Análisis paramétrico**

El método de ANOVA con medidas repetidas (MR) sirve para estudiar el efecto de uno o más factores cuando al menos uno de ellos es un factor intra-sujeto. Se entiende por medidas repetidas cuando el mismo sujeto participa de todas las condiciones de un experimento o siempre que se tiene de ellos múltiples valores en el tiempo. La diferencia entre el ANOVA y el ANOVA MR radica en que el primero trabaja con muestras independientes y el segundo con muestras relacionadas.

Para dos condiciones, el t-test para muestras relacionadas valora si existen diferencias estadísticamente significativas entre la media de ambas condiciones siempre y cuando la variable dependiente sea cuantitativa y se distribuya de manera normal.

**b) Análisis no paramétrico**

Para análisis que no se ajustan a ninguna distribución, se utiliza el test de Friedman. Es la alternativa no paramétrica a la prueba ANOVA cuando los datos son dependientes.

Sin embargo, para dos grupos relacionados se utiliza la prueba de Wilcoxon. Este test trabaja con rangos o posiciones que ocupan los datos una vez ordenados. Compara el rango de dos muestras pareadas y establece si existen diferencias entre ellas.

La *tabla 2* contiene un resumen de ambos tipos de diseño y sus diferentes pruebas.

| | | Paramétrico | No paramétrico |
|---|---|---|---|
| **Inter-sujeto** | Efecto principal | ANOVA | Kruskal-Wallis test |
| | Comparación por pares | T-test | Mann-Whitney test |
| **Intra-sujeto** | Efecto principal | ANOVA repetido | Friedman test |
| | Comparación por pares | T-test | Wilcoxon test |

*Tabla 2 - - Comparación diseños test inter-sujeto e intra-sujeto*

#### 2.1.2.3 Diseño mixto

Es un diseño combinado, siendo inter-sujeto para una cierta variable independiente (IV) e intra-sujeto para una cierta variable dependiente (DV).

### 2.1.3 Tipo de escala

Para calificar los resultados de los test se utilizan escalas semánticas y escalas continuas.

#### 2.1.3.1. Escalas semánticas

Las escalas semánticas son aquellas utilizadas para evaluar las magnitudes absolutas de los resultados de calificación (excelente, justo, pobre...). Este tipo de escala ofrece la posibilidad de identificar actitudes de los sujetos y su grado de satisfacción. No obstante, pueden resultar ambiguas para atributos de poco nivel. Funcionan mejor en test no paramétricosy permiten comparar resultados entre diferentes tipos de estudios o pruebas.

La mayoría de los métodos ITU-R y ITU-T utilizan este tipo de escala.

Se basan en las siguientes recomendaciones:

##### 2.1.3.1.1 ITU-T P.911

**a) Índices por categorías absolutas (ACR).**

[8] Es un juicio de categorías en el que las secuencias de prueba se presentan una tras otra y se califican de forma independiente en una escala de categorías: 5 (excelente), 4 (bueno), 3 (justo), 2 (pobre), 1 (malo), y es responsabilidad de cada sujeto interpretar estos niveles, de manera que se utiliza una referencia implícita. El método especifica que, después de cada presentación, se invite a los participantes a que evalúen la calidad de la secuencia presentada. El resultado se presenta mediante la MOS (Mean Opinion Score, media entre los resultados).

También se utilizan para los test MUSHRA.

**b) Índices por categorías de degradación (DCR).**

Implican la presentación de las secuencias de prueba por pares: el primer estímulo presentado en cada par es siempre la referencia fuente, mientras que el segundo estímulo es la misma fuente presentada a través de uno de los sistemas sometidos a prueba. Los niveles de degradación son 5 (imperceptible), 4 (perceptible pero no molesto), 3 (ligeramente molesto), 2 (molesto) y 1 (muy molesto).

#### 2.1.3.1.2 ITU-R BS.1116-3

**Estímulo triple con referencia oculta.**

[9] Es un método recomendado para estímulos con pequeñas diferencias. En el proceso se proporcionan dos notas en cada experimento y se hace posible que cada participante, de forma individual, compare directamente ambas notas y pueda realizarse un examen de estas comparaciones para todos los experimentos de dicho individuo.

En cada experimento puede calcularse la diferencia algebraica entre las dos apreciaciones restando siempre en el mismo sentido.

La razón por la que las escalas semánticas funcionan mejor en test no paramétricos es porque la distancia entre cada nivel adyacente no es consistente. No obstante, si existe un número considerable de muestras o sujetos (según el teorema central del límite, > 30), puede asumirse una distribución normal y por lo tanto este tipo de escala también puede aplicarse en test paramétricos.

### 2.1.3.2 Escalas continuas

En una escala continua, los encuestados califican los objetos colocando una marca en la posición adecuada en una línea que une un extremo de la variable con el otro, sin niveles intermedios.

Su uso resulta más apropiado si el objetivo es examinar las diferencias relativas entre las condiciones en lugar de la magnitud absoluta de cada calificación de estímulo. Resultan más útiles para test paramétricos ya que se trata de una escala de intervalo puro.

El problema principal de utilizar escalas continuas está en que los sujetos pueden utilizar de manera diferente el rango de escala incluso si el orden de calificación es el mismo. Para evitarlo, la recomendación ITU-R BS.1116 [9] propone el uso de datos normalizados mediante la *ecuación 4.*

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot S_s + x_s$$

*Ecuación 4*

Donde:

$Z_i$: resultado normalizado

$x_i$: nota del participante i

$x_{si}$: nota media para el participante i en la sesión s

$x_s$: nota media de todos los participantes en la sesión s

$s_s$: desviación típica para todos los participantes en la sesión s

$s_{si}$: desviación típica para el participante i en la sesión s.

### 2.1.4. Tipo de método

En ingeniería de audio, se utilizan dos tipos de métodos: los test de comparación de múltiple estímulo y los test de comparación de estímulo único.

#### 2.1.4.1 Test de comparación de estímulo múltiple

Los test de comparación de estímulo múltiple o multiestímulo son los más populares ya que permiten al sujeto cambiar entre estímulos libremente y escucharlos repetidamente. Se consideran más eficientes que los de estímulo único siempre y cuando el número de estímulos a comparar no sea muy elevado, ya que puede ocasionar imprecisión en los datos obtenidos.

Resultan más eficientes para medir diferencias relativas entre sistemas distintos en términos de atributos específicos de bajo nivel.

Un ejemplo de este método es el MUSHRA [10] (MUlti Stimulus test with Hidden Reference and Anchor) basado en la recomendación UIT-R BS.1534.

En los test MUSHRA se le presenta al oyente un cierto número de muestras de audio, una referencia, una versión oculta de la referencia y una o más anclas, las cuales son muestras de audio claramente buenas o malas. MUSHRA sirve para la evaluación de la calidad de audio intermedia comparando las características de audio de varias condiciones de prueba con deficiencias intermedias y proporciona resultados fiables y concisos.

En el método de prueba MUSHRA se utiliza una señal de referencia de gran calidad y se prevé que los sistemas sometidos a ensayo introduzcan degradaciones significativas. MUSHRA está previsto para evaluar los sistemas de audio de calidad intermedia. Si se utiliza MUSHRA con el contenido adecuado, en condiciones ideales la calificación del oyente oscila entre 20 y 80 puntos MUSHRA. Si la calificación de la mayoría de las condiciones de prueba oscila entre 80 y 100, puede que los resultados de la prueba no sean válidos.

El MUSHRA es un método de prueba doblemente ciega multiestímulo con referencia y patrones ocultos, a diferencia del de la Recomendación UIT-R BS.1116 que utiliza un método de prueba doblemente ciega de triple estímulo con referencia oculta. Se considera que el enfoque MUSHRA es más adecuado para evaluar degradaciones de nivel medio y grande [9].

En una prueba que implique pequeñas degradaciones, la dificultad para el participante consiste en detectar todo efecto perturbador que pueda estar presente en la señal. En esta situación, es

necesario incluir en la prueba una señal de referencia oculta, a fin de que el evaluador pueda evaluar la capacidad del participante para detectar satisfactoriamente estos efectos perturbadores. Por el contrario, en una prueba con degradaciones de nivel medio y grande, el evaluador no tiene dificultad para detectar los efectos parásitos, y por tanto no es necesaria a estos fines una referencia oculta. Además, la dificultad surge cuando el participante debe dar una nota de la incomodidad relativa de los diversos efectos parásitos. En este caso, el sujeto debe valorar su preferencia por un tipo de efecto perturbador respecto a otro.

Dependiendo del test se utiliza o no una referencia.

### 2.1.4.2 Test de comparación de estímulo único.

Los test de comparación de estímulo único utilizan un estímulo por prueba y el sujeto es quien califica la calidad del audio, generalmente en escala semántica. Esta escala está expresada en términos de una serie de niveles escalonados ordenados a lo largo de una secuencia característica de un atributo (figura 18).

Este test deposita la confianza en el juicio del sujeto ya que lo utiliza como referencia interna (porque no hay una referencia con la que compararlo) y el estímulo anterior puede servir de referencia para valorar el siguiente.

Se trata de un test más adecuado para medir la calidad general (ej. Quality of Experience) de manera absoluta.

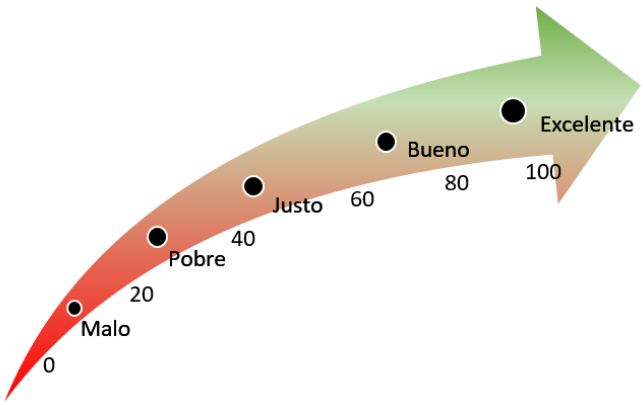



Figura 18 - *niveles escalonados en el test de comparación de estímulo único*

Esta prueba requiere menos tiempo para analizar un número de muestras. En general, un número mayor de muestras o de sujetos hace que sea más fácil alcanzar una distribución normal. El teorema del límite asume que una distribución normal es aquella con un número de muestras $N > 30$.

## 2.2. Test psicofísicos

El principal objetivo de los test psicofísicos es estimar el umbral perceptual o la sensibilidad en la detección de una señal y distinguir entre señales con diferentes intensidades.

En el proceso de escucha de estímulos auditivos principalmente intervienen dos procesos internos; el proceso sensorial y el proceso de decisión. El proceso sensorial transforma la señal de entrada en una sensación y el proceso de decisión toma una determinación de la respuesta al estímulo resultante del proceso sensorial. Los test psicofísicos miden la respuesta o salida resultante de ambos procesos internos.

### 2.2.1 Modelos de detección

Existen dos modelos de detección de la señal.

- Teoría del umbral: asume que el proceso sensorial tiene un umbral perceptual. El umbral es un estímulo mínimo de intensidad por encima del cual el proceso sensorial lleva a una correcta decisión o a una señal correctamente detectada. Se mide mediante umbral absoluto o JND (just noticeable difference).
- Teoría de detección de señal (SDT): supone que el umbral absoluto no existe en el proceso de decisión porque siempre existe un sesgo en el proceso de decisión debido a la existencia del ruido. Mide la discriminación del factor de sensibilidad (d') y el criterio de decisión del factor de sesgo (C) en lugar del umbral. Ambos son calculados desde las tasas de aciertos (señal presente) y falsas alarmas (señal ausente). Lo que diferencia las SDT de las teorías de umbral tradicionales es que el sujeto toma una decisión sobre si la señal está presente o no. Si la señal está presente, la persona puede decidir si está presente o ausente. Estos resultados se denominan aciertos y errores. Si la señal está ausente, la persona aún puede decidir si la señal está presente o ausente. Estos se denominan falsas alarmas o rechazos correctos (CR), respectivamente como se indica en la *tabla 3*.

| | | Presente | Ausente |
|---|---|---|---|
| **Decisión** | Presente | Acierto | Falsa Alarma |
| | Ausente | Error | CR |

*Tabla 3*

### 2.2.2 Métodos

Los métodos utilizados en los test psicofísicos se pueden clasificar en clásicos y adaptativos.

- Métodos clásicos: Fueron desarrollados por Fechner en 1860 y se pueden resumir básicamente en tres métodos:
    - Método de estímulos constantes (MCS): presenta repetidamente varios estímulos (entre 5 y 9) con diferentes intensidades de manera aleatoria, de

manera que el más débil esté claramente por debajo de cierto umbral y el más fuerte muy por encima. Obtiene una función psicométrica basada en el porcentaje correcto de respuestas obtenidas para cada intensidad en los estímulos, y el umbral se calcula a partir esa respuesta. Proporciona una información fiable y libre de sesgos, aunque requiere de mucho tiempo.

- Método de límites (MOL): mide el umbral ejecutando varias pruebas que presentan los estímulos de manera ascendente y descendente en términos de intensidad. No requiere una función psicométrica (relación entre una variable psicológica y una física) para estimar el umbral, pero las respuestas por parte del sujeto permiten determinar el valor del umbral absoluto y el umbral diferencial, entre otros. Requiere menos tiempo que el método de estímulos constantes, pero puede provocar errores de habituación.

- Método de ajuste (MOA): es el propio sujeto quien controla la intensidad de los estímulos para determinar el umbral, hasta que sea capaz de percibirlo (umbral absoluto) o hasta que sea igual a otro estímulo de comparación (umbral diferencial). Es el método más rápido y natural ya que es el propio sujeto quien manipula la intensidad, pero puede producir errores psicológicos (habituación).

- Métodos adaptativos: Son métodos que sirven para encontrar o estimar el umbral sensorial. Son más eficientes que los métodos clásicos ya que en estos generalmente se desconoce el umbral psicométrico y generalmente se obtienen numerosos datos que proporcionan poca información de interés.

  - Procedimientos de escalera: este método comienza con una serie ascendente o descendente de estímulos hasta que el sujeto comience o deje de oír dicho estímulo a una cierta intensidad. En ese momento se invierte la escalera, y estas reversiones son las que se promedian después para hallar el umbral. Existen numerosos tipos de algoritmos y diseños de escaleras. Los valores de umbral obtenidos de las escaleras pueden fluctuar mucho, por lo que se debe tener cuidado en su diseño.

    Uno de los diseños de escaleras más comunes (con tamaños de escalones fijos) es la escalera 1 arriba-N-abajo. Si el participante da la respuesta correcta N veces seguidas, la intensidad del estímulo se reduce en un tamaño de paso. Si el participante da una respuesta incorrecta, la intensidad del estímulo aumenta en un tamaño. Finalmente se estima un umbral a partir del punto medio de todas las ejecuciones.

  - Procedimientos bayesianos: Funcionan de manera similar a los de escalera, variando en la elección del nivel de intensidad posterior. Tras cada respuesta del observador, se calcula la probabilidad de dónde se encuentra el umbral a partir del conjunto de los pares de estímulo/respuesta anteriores. La mejor estimación del umbral corresponde al punto de máxima probabilidad. Estos procedimientos se consideran más robustos que los de escalera, aunque requieren mayor tiempo de implementación.

### 2.2.3 Tareas

Las tareas en los test psicofísicos sirven para determinar la cantidad de un estímulo necesaria para ser detectada por el sujeto. Pueden elegirse diferentes tareas para obtener la respuesta del sujeto en un proceso o método dado.

#### 2.2.3.1 Tarea Yes-No

Los participantes se someten a una serie de ensayos en los que deben juzgar la presencia (sí) o ausencia (no) de una señal. Generalmente se utiliza para estimar el umbral absoluto.

#### 2.2.3.2 2AFC

El método 2AFC (Two-alternative forced choice) se utiliza para medir la sensibilidad de un sujeto a dos versiones de un estímulo de entrada. Como su nombre indica, a los participantes se les presenta únicamente dos opciones y deben dar una respuesta.

Se puede utilizar para estimar tanto los umbrales absolutos en las tareas de detección como los umbrales de diferencia en las tareas de discriminación. Hay tareas espaciales (más apropiadas para test visuales) y temporales (adecuadas para test auditivos). En esta última, los estímulos son presentados al sujeto uno por uno con una pausa entre ambos y se estima la diferencia en JND.

Generalmente 2AFC es más utilizado para detectar los umbrales que la tarea yes-no ya que produce un mayor nivel de rendimiento.

#### 2.2.3.3 ABX

También denominado 3I-2AFC (three interval 2AFC), es un método muy utilizado en test auditivos. Es un tipo de método 2AFC con 3 estímulos, que examina la detectabilidad de la diferencia entre estímulos. Compara dos opciones de estímulos sensoriales para identificar diferencias detectables entre ellos.

A un sujeto se le presentan dos muestras conocidas seguidas de una muestra desconocida (X) que se selecciona aleatoriamente entre A o B. Si X no se puede identificar de manera confiable con un p-valor bajo en un número predeterminado de ensayos, entonces la hipótesis nula no se puede rechazar y no se puede probar que existe una diferencia perceptible entre A y B.

## 2.4 Elección del test en la prueba

En el listening test realizado en el presente proyecto (5), la opción escogida ha sido un grading test de comparación multiestímulo, en concreto un MUSHRA con 7 estímulos.

Son más sencillos de implementar que los test psicofísicos, y para evaluar a la misma vez las diferentes técnicas de upmixing (apartado 3) es conveniente utilizar estímulos múltiples, de manera que el propio sujeto pueda cambiar entre estímulos libremente y escucharlos a elección tantas veces como quiera.

Se ha escogido esta técnica ya que no es necesario realizar el test a un gran número de oyentes a diferencia de la prueba MOS, y porque los estímulos no tienen por qué emitirse necesariamente en el idioma nativo de los oyentes, siendo estos voces emitidas en inglés y el lenguaje nativo de los oyentes el castellano.

Además, este método tiene la ventaja de visualizar muchos estímulos al mismo tiempo, de forma que el sujeto puede verificar cualquier comparación entre ellos directamente.

También, en este caso resulta más eficiente que los de único estímulo porque el número de estímulos a comparar no es muy elevado, ya que esto puede ocasionar imprecisión en los datos obtenidos.

# 3. Técnicas de upmixing

## 3.1 Introducción

En los últimos años la cantidad de contenido grabado con cámaras de 360º ha ido en aumento. Como resultado, el audio Ambisonics ha encontrado un nuevo propósito, proporcionando bandas sonoras inmersivas para acompañar el video de 360º. Además, hay un número cada vez mayor de experiencias de realidad virtual, juegos y aplicaciones que requieren un audio envolvente en el que Ambisonics puede desempeñar un papel importante.

La reproducción con auriculares de escenas de sonido espacial grabadas es cada vez más relevante para las aplicaciones audiovisuales inmersivas. Existen herramientas que permiten trabajar con grabaciones estéreo de dos formas. La primera es colocar el sonido estéreo en algún lugar dentro del campo de sonido, que sigue los movimientos de la cabeza del espectador. Este método funciona bien para sonidos adjuntos a un objeto, pero no proporciona un sonido completamente envolvente. Para la música o los sonidos ambientales puede resultar interesante, pero para una completa inmersión en la escena sonora queda algo incompleto. La segunda es colocar música y otros elementos en la pista estéreo bloqueada, pero no responden al movimiento de la cabeza del espectador, por lo que carece de la experiencia de inmersión total que se obtiene al usar las grabaciones de Ambisonics.

Es por eso por lo que existen técnicas o algoritmos que tienen opciones para mezclar a Ambisonics de primer, segundo o tercer orden, atendiendo las cada vez más solicitadas necesidades de audio de postproducción para videos de 360º y realidad virtual.

## 3.2 Algoritmos

Existen dos tipos de métodos de reproducción: los paramétricos y los no paramétricos. Los métodos de reproducción no paramétricos populares, como los Ambisónicos de primer orden (FOA), ahora pueden superarse mediante el uso de métodos de reproducción paramétricos basados en la percepción. Los métodos paramétricos, como la codificación de audio direccional (DirAC) o HARPEX, han ganado notoriedad recientemente por ser capaces de lograr una nitidez o envolvente más allá de la reproducción tradicional de Ambisonics de primer o menor orden, utilizando las mismas señales de Ambisonics de menor orden.

En este proyecto se van a analizar a fondo los diferentes métodos paramétricos de upmixing.

### 3.2.1 DirAC

#### 3.2.1.1 Introducción

La codificación de audio direccional (DirAC) es un método paramétrico para la reproducción y mezcla ascendente (upmixing) de sonido espacial, que opera en un dominio de transformación tiempo-frecuencia y extrae un parámetro de dirección de llegada (DoA) y un parámetro de difusión en cada slot de tiempo-frecuencia [11, 14, 15, 16]. Se basa en el formato de audio 3D B-format y consigue una reproducción perceptiva flexible y eficaz para altavoces o auriculares.

Su función básica es la mejora de la reproducción FOA extrayendo una dirección de llegada (DoA) y un parámetro de difusión. Esto se consigue dividiendo esencialmente la escena sonora en un único flujo de origen y un flujo difuso isotrópico, reproducido entonces a través de altavoces o auriculares.

En DirAC, se supone que en un instante de tiempo y en una banda crítica la resolución espacial del sistema auditivo es limitada para decodificar una señal para la dirección y otra para la coherencia interaural. Además, se asume que, si la dirección y la difusión del campo de sonido se miden y se reproducen correctamente, un oyente humano percibirá correctamente las señales direccionales y de coherencia. En la práctica, el procesamiento de DirAC se realiza en dos fases: el análisis de metadatos direccionales y la síntesis de sonido, donde los metadatos direccionales se utilizan activamente en la reproducción (*figuras 19 y 20*).

DirAC analiza en ventanas cortas de tiempo el espectro de sonido junto con la dirección y la difusión en las bandas de frecuencia de la audición humana, y utiliza esta información en síntesis. Tiene aplicaciones en la captura, codificación y resíntesis de sonido espacial, en teleconferencias, filtrado direccional y en entornos auditivos virtuales.

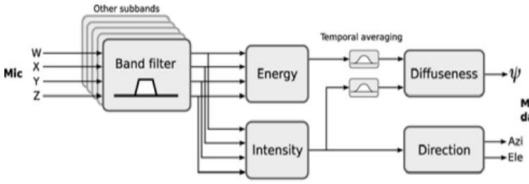



*Figura 19 – Análisis DirAC*

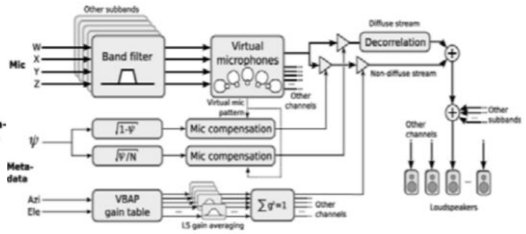



*Figura 20 – Síntesis DirAC*

Teniendo en cuenta la información de la matriz de grabación de los cuatro canales que genera las señales de B-format (W, X, Y y Z), es posible mejorar tanto el análisis de la escena de sonido como la reproducción. Después estos resultados se amplían para varias configuraciones de grabación procurando conseguir una mayor generalización de DirAC en un dominio de transformación espacial (dominio armónico esférico o SHD), con señales B-format de orden superior.

### 3.2.1.2 División en bandas de frecuencia

Las señales de entrada para la codificación DirAC son señales B-format. Primero, las señales se dividen en bloques en tiempo y frecuencia. Hay dos enfoques posibles para realizar esto: banco de filtros en cuadratura (QMF) y transformada de Fourier de corta duración (STFT). Además de estos, hay total libertad para diseñar un banco de filtros con filtros arbitrarios que estén optimizados para un propósito específico. Independientemente de la transformación tiempo-frecuencia seleccionada, el objetivo del diseño es imitar la resolución de la audición espacial humana.

En las primeras implementaciones de DirAC, se utiliza un banco de filtros con filtros subbanda arbitrarios alternativamente con STFT con ventanas de tiempo de 20 ms (figura 21). La resolución de tiempo uniforme en todas las frecuencias es un inconveniente para la implementación de STFT, que puede producir algunos errores a altas frecuencias con algunas señales críticas debido a ventanas temporales demasiado largas.

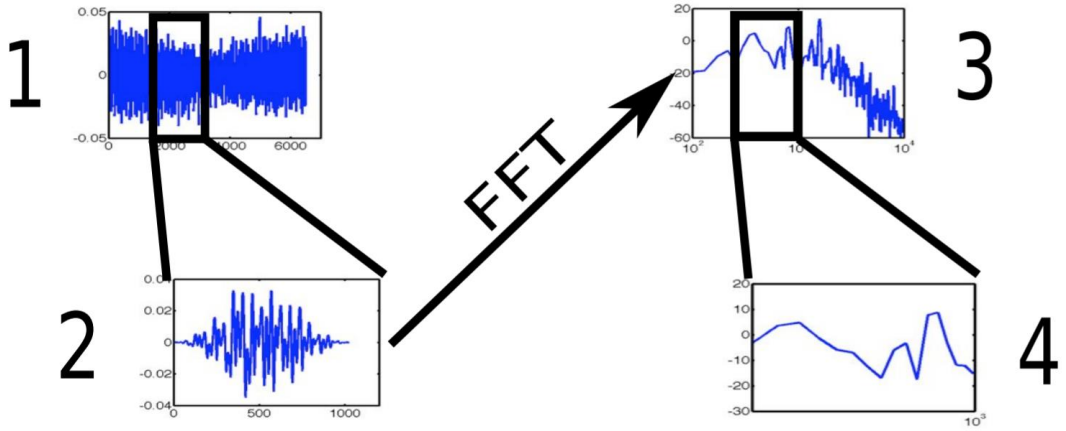



*Figura 21 - División de la señal en bloques en tiempo y frecuencia utilizando el método STFT. En la parte 1 de la figura hay una señal continua en el dominio del tiempo. Se recorta un bloque de cierto tamaño para su análisis y se abre una ventana de Hann dando como resultado la parte 2 de la figura. La FFT se calcula a partir de este bloque (parte 3). Los valores correspondientes a una determinada banda ERB (ancho de banda rectangular equivalente) se cortan para el análisis. En la parte 4 se puede observar un bloque de la señal que representa cierto instante de tiempo y banda crítica de audición.*

La implementación del banco de filtros resuelve este problema, sin embargo, puede tener limitaciones en la complejidad computacional. Se ha utilizado tanto un banco de filtros de fase lineal como una versión de STFT de resolución múltiple, donde el sonido de entrada se divide en pocos canales de frecuencia y procesados con diferentes STFT que tienen longitudes de ventana adecuadas para cada banda de frecuencia. La elección de la transferencia de tiempo-frecuencia no tiene un impacto importante en la calidad del audio en la reproducción de DirAC. Las diferencias son típicamente audibles solo con un material de entrada que tenga modulaciones de aproximadamente de 100Hz a 1kHz en envolventes de señal a altas frecuencias.

### 3.2.1.3 Análisis direccional

El objetivo del análisis direccional es estimar la dirección de llegada del sonido en cada banda de frecuencia, junto con una estimación si el sonido llega desde una o varias direcciones al mismo tiempo.

En principio, esto se puede realizar con una serie de técnicas, sin embargo, haciendo el análisis energético del campo sonoro resulta suficientemente adecuado. El análisis energético se puede realizar cuando la señal de presión y las señales de velocidad en 1-3 dimensiones se capturan desde una sola posición. En las señales de B-format de primer orden, la señal omnidireccional (W) se ha reducido en $\sqrt{2}$. La presión sonora se puede estimar como $P = \sqrt{2}W$, en el dominio SFTF. Los canales X, Y y Z tienen el patrón direccional de un dipolo dirigido a lo largo del eje cartesiano, que forman juntos un vector U = [X, Y, Z]. El vector estima el vector de velocidad del campo sonoro y también se expresa en el dominio STFT.

La energía E del campo sonoro se puede calcular siguiendo la siguiente expresión:

$$E = (\rho_0/4)||U||^2 + (1/4\rho_0 c^2)||P||^2$$

donde $\rho_0$ es la densidad media del aire en $kg/m^3$ y $c$ es la velocidad del sonido (344 m/s). La captura de señales de B-format se puede obtener colocando de manera coincidente los micrófonos direccionales o con un conjunto de micrófonos omnidireccionales poco espaciados. En algunas aplicaciones, las señales de micrófono pueden formarse en el dominio computacional, es decir, simuladas. El análisis se repite con tanta frecuencia como sea necesario para la aplicación, normalmente con la frecuencia de actualización del valor 100Hz-1000Hz.

La dirección del sonido se define como la dirección opuesta del vector de intensidad $I = \underline{P}U$, donde $\underline{(\cdot)}$ es una conjugación compleja. La dirección se indica como los valores de elevación y azimut angular correspondientes en los metadatos transmitidos. La difusión del campo de sonido $\Psi$ se calcula como indica la *ecuación 5*.

$$\Psi = 1 - \frac{||E\{I\}\,||}{cE\{E\}}$$

*Ecuación 5*

donde E denota el operador estadístico esperanza, donde $||\cdot||$ denota la norma del vector y {} denota el tiempo promedio. El resultado de esta ecuación es un número real entre cero y uno, que indica si la energía del sonido llega de una sola dirección o de todas las direcciones. Esta ecuación es apropiada en el caso de que la información de velocidad 3D completa esté disponible. Si la configuración del micrófono ofrece velocidad solo en 1D o 2D, se utiliza la *ecuación 6*, que da una estimación que está más cerca de la difusión real del campo de sonido.

$$\Psi_{cv} = \sqrt{1 - \frac{||E\{I\}||}{E\{||I||\}}}$$

*Ecuación 6*

#### 3.2.1.4 Transmisión DirAC

En muchas aplicaciones, el sonido espacial debe transmitirse de un lugar a otro. En DirAC, esto se puede realizar con enfoques diferentes. Una técnica sencilla es transmitir todas las señales de B-format. En ese caso, no se necesitan metadatos y el análisis se puede realizar en el extremo receptor. Sin embargo, en la versión de baja tasa de bits, solo se transmite un canal de audio, lo que proporciona una gran reducción en la tasa de datos, y el inconveniente es una ligera disminución en la calidad tímbrica del sonido reverberante y una disminución en la precisión direccional en escenarios de múltiples fuentes.

En algunos casos, resulta ventajoso fusionar varios flujos DirAC mono o estéreo. Esta no es una tarea trivial, ya que no existe una forma sencilla de fusionar metadatos direccionales. No

obstante, existen dos métodos propuestos que proporcionan una fusión eficiente y sin distorsiones o errores explicados en el estudio [13].

#### 3.2.1.5 Síntesis DirAC con altavoces

La versión de alta calidad de la síntesis DirAC, recibe todas las señales B-format, a partir de las cuales se calcula una señal de micrófono virtual para cada dirección de altavoz. El patrón direccional utilizado es típicamente un dipolo. Después las señales del micrófono virtual se modifican de forma no lineal, según los metadatos. En la versión de baja tasa de bits de DirAC sólo se transmite un canal de audio. La diferencia en el procesamiento es que todas las señales de micrófono virtual serían reemplazadas por el único canal de audio recibido. Las señales del micrófono virtual se dividen en dos flujos; difuso y no difuso, que se procesan por separado.

##### 3.2.1.5.1 Sonido no difuso

La parte no difusa tiene cierta dirección y puede consistir, por ejemplo, en sonido directo y las primeras reflexiones en una habitación. Se reproduce como fuentes puntuales mediante el uso de paneo en amplitud VBAP (Basics Vector Base Amplitude Panning) que sigue el esquema de la *figura 22*. En este paneo, se aplica una señal de sonido monofónica a un subconjunto de altavoces después de la multiplicación con factores de ganancia específicos del altavoz. Los factores de ganancia se calculan utilizando la información de la configuración del altavoz y la dirección de paneo especificada. En la versión de baja tasa de bits, la señal de entrada simplemente se desplaza en las direcciones implícitas en los metadatos. En la versión de alta calidad, cada señal de micrófono virtual se multiplica por el factor de ganancia correspondiente, lo que produce el mismo efecto con el paneo, sin embargo, es menos propenso a los artefactos no lineales. En muchos casos, la dirección de los metadatos está sujeta a cambios temporales abruptos. Para evitar artefactos, los factores de ganancia para los altavoces calculados con VBAP se suavizan mediante la integración temporal con una constante de tiempo dependiente de la frecuencia equivalente a aproximadamente 50 períodos de ciclo en cada banda. Esto elimina efectivamente los artefactos, sin embargo, los cambios de dirección no se perciben como más lentos que sin promediar en la mayoría de los casos.

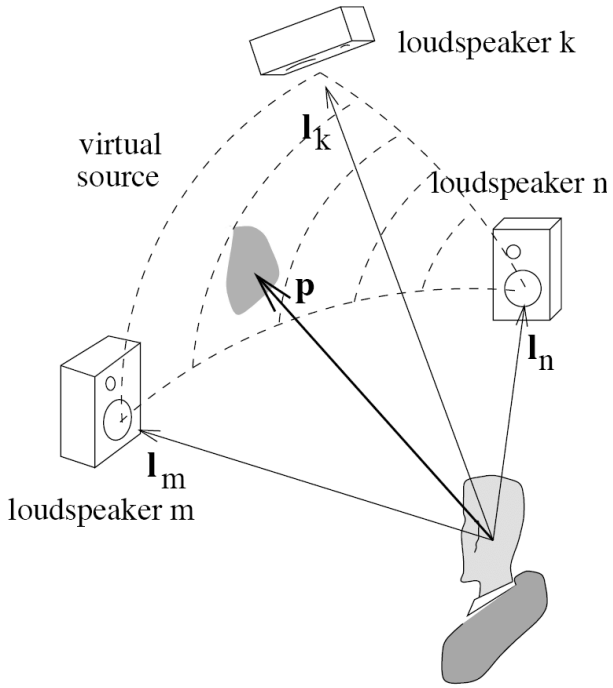



*Figura 22 - VBAP. El sonido se desplaza en la dirección correcta utilizando conjuntos de tres altavoces.*

#### 3.2.1.5.2 Sonido difuso

Sonido difuso: La parte difusa no tiene dirección y puede consistir, por ejemplo, en sonido difuso en una habitación. Consiste en crear una percepción del sonido que envuelve al oyente. La forma más sencilla de crear un sonido difuso es agregar la parte difusa del sonido a cada altavoz. Al utilizar este método, el sonido difuso es coherente en todos los altavoces, por lo que se coloca en la dirección en la que está el altavoz más cercano. El sonido de diferentes altavoces llega en diferentes momentos. Esto provoca un filtrado peine, ver *figura 23*. Si el oyente mueve la cabeza, los retrasos cambian y también se mueven las muescas en la respuesta de frecuencia. Esto se escucha como un efecto similar al flanger o phaser.

La solución es decorrelacionar las señales de sonido difusas a cada altavoz. La fase en cada altavoz debe ser aleatoria. Esto se puede hacer, por ejemplo, filtrando la señal con una ráfaga de ruido aleatorio que tiene una respuesta de magnitud de la unidad. El ruido debe ser diferente para cada altavoz. El resultado es una señal con la misma magnitud de respuesta que la señal original, pero con una respuesta de fase aleatoria (*figura 23*). Las respuestas de fase de las señales para cada altavoz son aleatorias, lo que significa que están decorrelacionadas. Otra solución es usar retardos aleatorios para diferentes bandas de frecuencia y diferentes altavoces o aleatorizar la fase en el dominio de la frecuencia. El resultado de reproducir señales no relacionadas de múltiples altavoces alrededor del oyente es que el sonido no tiene una dirección prominente y el sonido rodea al oyente. Moverse ligeramente no posiciona el sonido al altavoz más cercano.

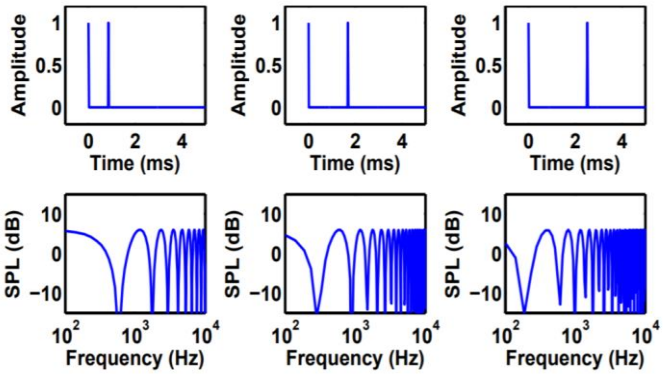



*Figura 23 – Filtros peine. La parte de arriba de la figura muestra las señales que tienen el impulso en el instante 0 y otro después de un retraso. La parte de abajo muestra sus correspondientes transformadas de Fourier.*

### 3.2.1.6 Síntesis DirAC con auriculares

Para reproducir audio espacial a través de auriculares se implementan altavoces virtuales a través de las funciones de transferencia HRTF.

#### 3.2.1.6.1 Sonido no difuso

La manera más sencilla de reproducir un sonido no difuso es utilizar paneo en amplitud entre dos canales de auriculares. Computacionalmente es una forma muy eficiente de posicionar objetos auditivos. El sonido paneado en amplitud reproducido con auriculares generalmente da la impresión de que el sonido proviene del interior de la cabeza. El principal problema de este método es que el sonido sólo se mueve a lo largo del eje entre los dos oídos (lateralización) en la dimensión izquierda-derecha.

Otro enfoque es utilizar las HRTF para sonido no difuso. De esta manera se pueden sintetizar todas las direcciones posibles en tres dimensiones, aunque es computacionalmente pesado. Este es el método más utilizado y el que se va a explicar con más detalle en los próximos párrafos.

La tercera opción consiste en utilizar algún modelo simplificado de HRTF, como modelar el efecto de la cabeza y el torso utilizando una cabeza y un torso esféricos. Los HRTF de este modelo se aproximan mediante dos retardos de tiempo y dos IIR de primer orden.

En el método de las HRTF, las señales de salida se crean para los altavoces mediante paneo en amplitud y luego se filtran con los HRTF correspondientes (*figura 24*). Las salidas filtradas para cada altavoz se suman para obtener señales para los oídos izquierdo y derecho. La ventaja de este método es que también funciona bien con cardioides virtuales en la versión de micrófono virtual de DirAC, porque las señales de entrada al filtrado HRTF son señales de altavoz. En la versión de micrófono virtual, los cardioides virtuales se crean para corresponder a la dirección de los altavoces virtuales. Estos cardioides virtuales se procesan para obtener las señales de los altavoces.

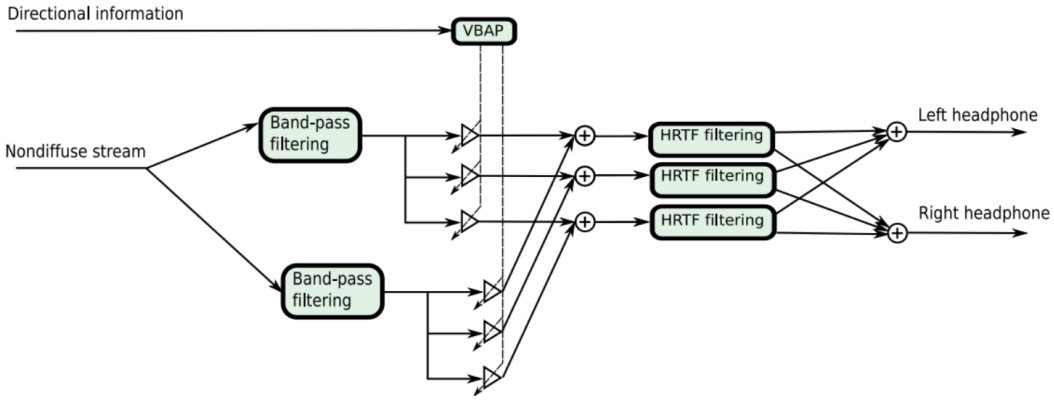



*Figura 24 – Síntesis del sonido no difuso mediante altavoces virtuales HRTF*

Otro enfoque computacionalmente más eficiente es el de interpolación, que consiste en crear un par de funciones de transferencia que coloquen diferentes bandas de frecuencia en diferentes direcciones (*figura 25*). Cada banda de frecuencia puede tener una dirección diferente, por lo que para cada una las HRTF deben interpolarse por separado según la dirección de esa banda de frecuencia. Las HRTF para diferentes bandas se filtran con un filtro de paso de banda para obtener la función de transferencia solo de esa banda. Estas HRTF filtradas mediante paso de banda se suman para obtener la HRTF resultante. Una ventaja de este enfoque es que las HRTF se pueden interpolar utilizando cualquier método que se desee, y que se puede lograr una mayor precisión que con el método de las HRTF, pero los problemas con los artefactos deben resolverse primero.

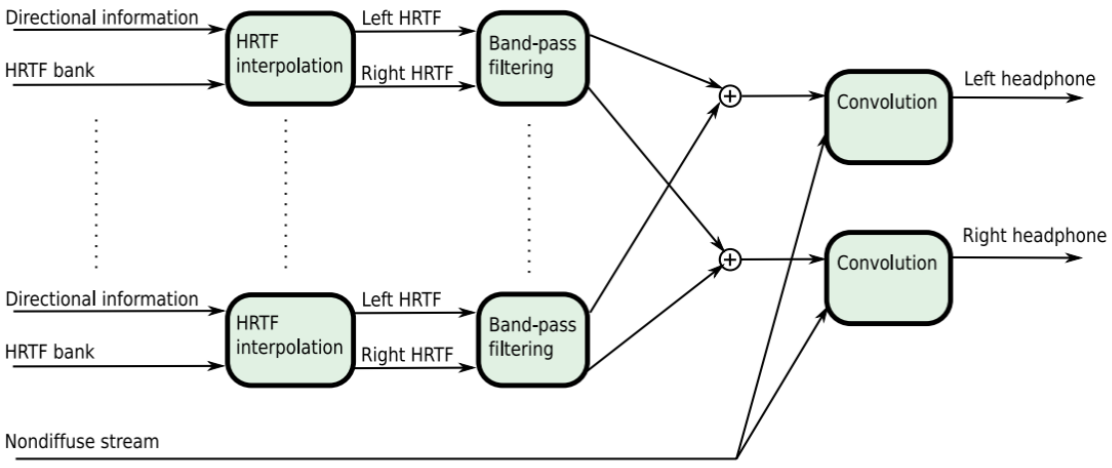



*Figura 25 – Síntesis del sonido no difuso mediante interpolación de HRTF*

#### 3.2.1.6.2 Sonido difuso

La reproducción binaural tiene dos salidas, señales para los oídos izquierdo y derecho. Por tanto, la forma más sencilla de crear un sonido difuso es añadir señales decorrelacionadas a cada oído. Estas dos señales no se pueden utilizar directamente para el sonido difuso, porque la forma espectral de estas señales es diferente de la forma espectral de la presión sonora en el canal auditivo causada por el sonido difuso. La respuesta de magnitud de estas dos señales debe ser similar a la de las HRTF (*figura 26*).

Esto se puede hacer filtrando con HRTF de campo difuso. Las HRTF de campo difuso corresponden a las HRTF medidas en campo difuso. No coloca la señal en ninguna dirección, pero hace que el timbre del sonido sea similar al de los HRTF. Las HRTF de campo difuso se pueden medir en una habitación ecoica (con eco) o promediando las HRTF de diferentes direcciones.

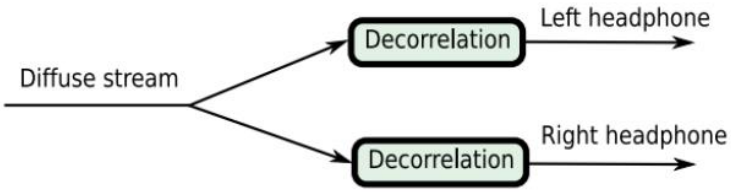



*Figura 26 – Síntesis del sonido difuso con auriculares*

Otra forma es crear señales decorrelacionadas y filtrarlas con HRTF (*figura 27*). Las HRTF deben seleccionarse de modo que cubran toda la esfera alrededor del oyente. De esta manera, el sonido difuso no parece provenir de ninguna dirección distinta, sino de todas partes alrededor del oyente. De esta forma se reproduce la sensación de estar en otro espacio.

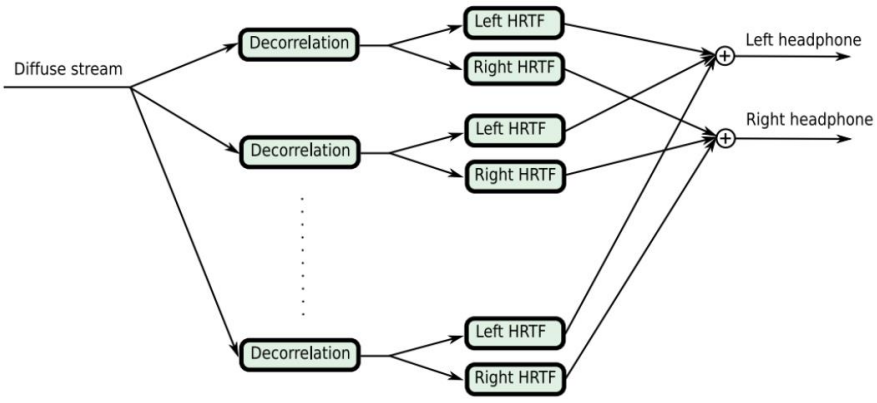



*Figura 27 – Síntesis del sonido difuso con auriculares mediante señales decorrelacionadas*

#### 3.2.1.7 Plugin

[14] DirAC no cuenta con un plugin propio, pero sí con una implementación en Matlab, la cual se utiliza en el apartado 5 para obtener los audios en DirAC.

### 3.2.2 Harpex

#### 3.2.2.1 Introducción

Generalmente FOA tiene baja resolución angular y un punto óptimo pequeño, y la decodificación de muchos altavoces es un problema que los métodos de decodificación paramétrica tratan de resolver, aunque con el riesgo añadido de introducir errores [21]. El método HARPEX (*High Angular Resolution Planewave EXpansion*), trata de expandir la resolución angular, y combina la nitidez espacial de los métodos paramétricos con la corrección física de la decodificación lineal sin introducir errores audibles.

En un test de audio, un decodificador que utiliza este método para decodificar señales B-format de primer orden puntúa mucho más alto que la decodificación max-$R_E$ (1.2.3) de las mismas señales, y de manera similar a la decodificación max-$R_E$de versiones de tercer orden de las mismas señales.

La resolución angular y el tamaño del punto óptimo de una señal B-format dependen del número de altavoces y del número de canales de entrada. Aumentar el número de altavoces sin aumentar también el orden del B-format no mejora la resolución angular con los métodos de decodificación lineal.

El método propuesto en [21] descompone cada componente de frecuencia del campo sonoro en dos ondas planas y después reconstruye esas ondas planas con los altavoces disponibles. El resultado es una reconstrucción físicamente correcta del campo de sonido en el punto central (u óptimo) y una expansión de la onda plana de alta resolución fuera de él. En las frecuencias en las que solo una o dos ondas planas afectan significativamente a la señal grabada, es una expansión físicamente correcta. En frecuencias con dos ondas planas, determinar las direcciones correctas de llegada impulsa la capacidad del oído humano. Con más de dos ondas planas a la misma frecuencia, es más probable que los errores direccionales pasen desapercibidos.

#### 3.2.2.2 Descomposición paramétrica

Este método opera en el dominio de la frecuencia y el primer paso es aplicar funciones de ventana superpuestas, *cero padding* y FFT. Después, 8 números representan cada intervalo de tiempo/frecuencia de una señal 3D de primer orden: la parte real e imaginaria de cada canal. Esto se descompone en dos ondas planas, cada una representada por 4 números

independientes; la parte real e imaginaria de la amplitud y un vector unitario de tres elementos que apunta en la dirección de llegada.

Sólo se considerará un componente de frecuencia, asumiendo que se aplica el mismo método a todos los componentes de frecuencia. Siendo X la señal de valor complejo, entonces la descomposición puede expresarse de la siguiente manera:

$$\mathbf{X} = \begin{bmatrix} w_r + iw_i \\ x_r + ix_i \\ y_r + iy_i \\ z_r + iz_i \end{bmatrix} = \underbrace{\begin{bmatrix} 2^{-\frac{1}{2}} & 2^{-\frac{1}{2}} \\ x_1 & x_2 \\ y_1 & y_2 \\ z_1 & z_2 \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}}_{\mathbf{A}}$$

donde las 3 filas inferiores de V contienen vectores unitarios de valor real que apuntan en las direcciones de llegada y A contiene las amplitudes complejas de esas ondas. Para calcular la descomposición se hallan primero las fases de $a_1$ y $a_2$ después su magnitud y finalmente se hace la matriz inversa para encontrar V.

#### 3.2.2.3 Experimentos

El método está basado en modelos y no es lineal, por lo que no es trivial predecir su comportamiento cuando los supuestos del modelo no son válidos. Por ello en [21] se realizan diferentes experimentos numéricos para estudiar el comportamiento bajo la influencia de varios tipos de ruido.

En cada experimento, los resultados se visualizan en un diagrama de dispersión donde se varía un solo parámetro a lo largo del eje horizontal. El azimut de las estimaciones de dirección se traza a lo largo del eje vertical. Para cada píxel horizontal, se realizan 300 experimentos y se trazan dos puntos para cada experimento, a menos que el método no descomponga la señal. La opacidad de cada punto es proporcional a la amplitud de la onda plana correspondiente.

Para llevar a cabo dichos métodos, se parte de la división de la señal en bandas de frecuencia y se utiliza la correlación a corto plazo entre el canal W y cada uno de los canales direccionales para calcular una estimación de la dirección de llegada (la cual estima la presencia de ruido) y la difusión del sonido en cada banda de frecuencia, que a su vez se ha utilizado para dirigir parte de la señal a los altavoces más cercanos a esa dirección.

Bajo **ruido no direccional**, se sintetiza una señal a partir de dos fuentes puntuales de ruido blanco y se agrega ruido blanco a cada uno de los cuatro canales. Como resultado, las estimaciones de dirección son generalmente correctas a niveles de ruido por debajo de 0 dB y se degradan gradualmente a niveles de ruido por encima de 10 dB (*figura 28*).

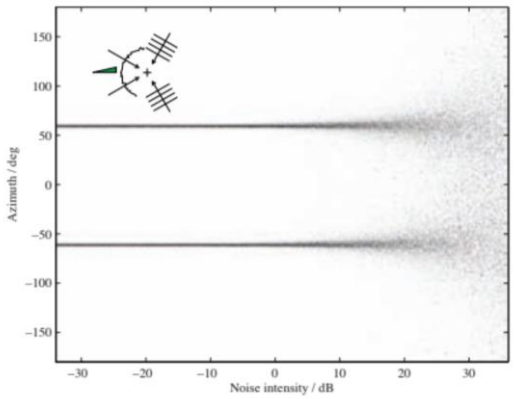



*Figura 28 – Ruido no direccional*

Bajo **ruido direccional** (*figura 29*), se sintetiza una señal que consta de tres fuentes puntuales de ruido blanco. La potencia de dos de las fuentes se mantiene constante mientras que la potencia de la tercera fuente interferente varía en un rango de –35 dB a +35 dB en relación con cada una de las otras fuentes. Se observa que las estimaciones de dirección son prácticamente inmunes a las interferencias por debajo de –20 dB y que en una región de transición donde las tres fuentes están en el mismo rango de potencia de 10 dB, las estimaciones de dirección caen en una región amplia en el plano que contiene las tres fuentes. Cuando la fuente interferente es más de 20 dB más fuerte que las otras fuentes, la estimación de una dirección corresponde a la dirección de la fuente interferente sola. También se produce una segunda estimación de dirección en esta región, ampliamente distribuida sobre el plano, cuya amplitud es insignificante.

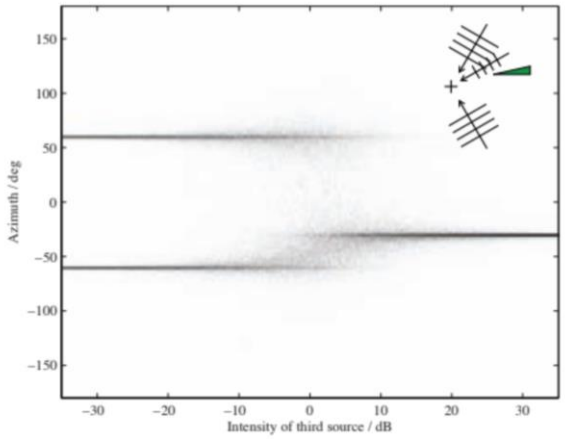



*Figura 29 – Ruido direccional*

En un tercer experimento **(degeneración de fase)** se sintetiza una señal que consta de dos fuentes puntuales sinusoidales coherentes cuya fase relativa varía en un rango de –180° a 180°. Se agrega ruido blanco a cada canal y las estimaciones de dirección son correctas, excepto cuando la fase de las dos ondas difiere en menos de 10° o más de 170°, donde las estimaciones

de dirección divergen. Otras ejecuciones del mismo experimento muestran que las regiones de divergencia se estrechan a medida que disminuye el nivel de ruido, mientras que la cantidad de divergencia permanece constante (*figura 30*).

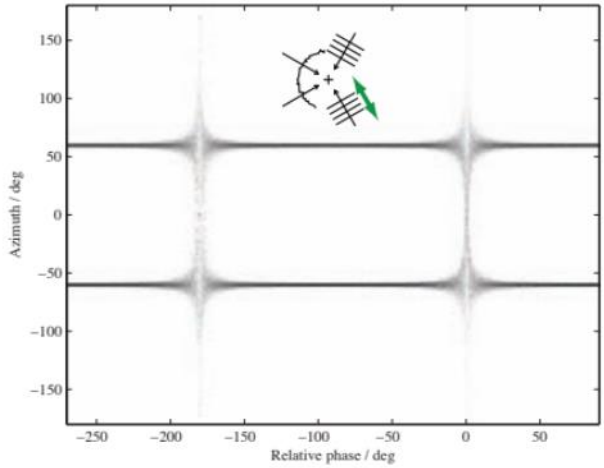



*Figura 30 – Degeneración de fase*

En el último experimento **(degeneración direccional)** se sintetiza una señal a partir de dos fuentes puntuales de igual potencia, una en línea recta y la otra moviéndose en círculo alrededor del plano horizontal. El ruido se agrega a un nivel igual a 0 dB y las estimaciones de dirección son correctas para ángulos mayores de 30° entre las dos ondas planas. Para ángulos más pequeños, las estimaciones de dirección se amplían hasta que las dos fuentes están muy cerca (<5°). En esta separación, las fuentes se fusionan en una estimación de dirección con una amplitud grande (*figura 31*).

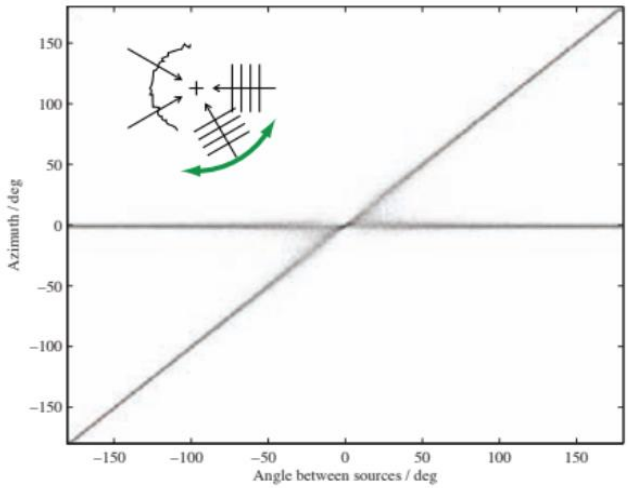



*Figura 31 – Degeneración direccional*

### 3.2.2.4 Decodificación

La implementación más sencilla de un decodificador que usa HARPEX consiste en enviar cada una de las dos estimaciones de dirección en una función de paneo, que devuelve un peso para cada uno de los altavoces de salida. Luego, cada peso se multiplicaría por la amplitud compleja de la onda plana correspondiente para generar señales de altavoz (*figura 32*), sin mostrar ventanas, FFT e IFFT, que también son necesarias.

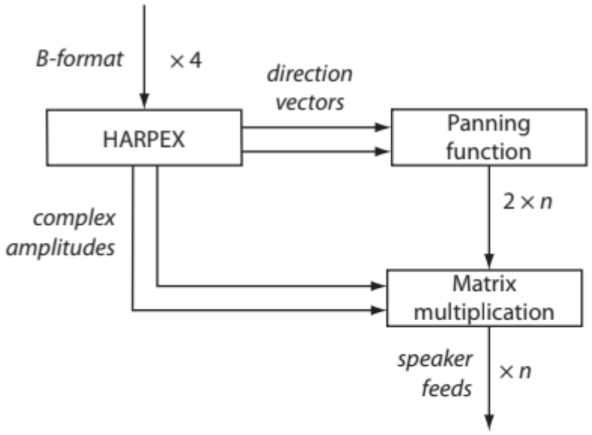



*Figura 32 – Decodificador simple*

El problema de esta implementación está en que el método HARPEX no siempre devuelve una solución y debe ir acompañado de un método alternativo para usar en tales casos, pero ninguno proporciona más de una estimación de dirección única.

Además, los vectores de dirección pueden cambiar rápidamente de un fotograma al siguiente, provocando artefactos en el dominio del tiempo relacionados con el período del fotograma (se puede solucionar suavizando los vectores de dirección o los pesos del paneo resultante).

Los vectores de dirección también pueden diferir significativamente de un intervalo de frecuencia al siguiente dentro de una trama, provocando dispersión y transitorios indeseablemente suaves. Este problema se puede solventar suavizando los vectores de dirección a lo largo del eje de frecuencia o suavizando los pesos del paneo.

Un efecto del suavizado es introducir leakage para fuentes puntuales que se han separado. La cantidad de suavizado representa una compensación entre la nitidez de la localización y la audibilidad de los artefactos (para fuentes difusas, el leakage es deseable).

Si el suavizado se realiza antes o después del paneo, la descomposición en dos ondas planas ya no es válida porque se altera la dirección de las ondas. Para que una descomposición sea válida, se pueden agregar otras dos ondas planas y la señal debe descomponerse en esta nueva base. En los casos en que HARPEX no devuelve ninguna solución, se deben agregar tres nuevas ondas planas, de modo que la segunda descomposición siempre devuelva 4 ondas.

Para asegurar un buen acondicionamiento de la matriz de decodificación, las ondas planas adicionales deben colocarse lo más lejos posible de las ondas planas originales y también entre sí.

Como la descomposición y la resíntesis se dividen en dos operaciones independientes, se puede utilizar cualquier función de paneo. La opción más obvia para diseños de altavoces horizontales es un paneo por pares, usando una ley de panoramización en el rango de 3 a 6 dB. Una ley de panoramización es un principio de grabación que establece que cualquier señal de igual amplitud y fase que se reproduzca en los dos canales de un sistema estéreo aumenta en volumen hasta 6.02 dB. Esto puede admitir fácilmente diseños irregulares y puede extenderse a diseños con altura utilizando paneo en amplitud basada en vectores.

Otras funciones interesantes son las funciones de paneo equivalentes de Ambisonics, la síntesis de campo de ondas y los armónicos esféricos. En este último caso, la salida del decodificador no es la alimentación de los altavoces, sino una mezcla en B-format de primer orden a B-format de orden superior. Esta función de paneo tiene la propiedad de reconstruir el campo de sonido en el punto central (óptimo), si se combina con un decodificador HOA adecuado.

La decodificación para la reproducción binaural utilizando HRTFs presenta otros desafíos derivados del hecho de que estas contienen términos de fase que codifican el retardo de tiempo interaural. Estos términos de fase pueden provocar artefactos audibles.

#### 3.2.2.5 Test de audio

Dado que el método propuesto en [21] tiene como objetivo mejorar la calidad del sonido percibido fuera de la región donde se puede reproducir el campo sonoro físico, la forma de evaluar si funciona de acuerdo con las expectativas es realizar test de audio. En él se realiza un experimento siguiendo la configuración del estudio [23], el cual funciona de acuerdo con las recomendaciones MUSHRA.

En la prueba se colocan 12 altavoces Genelec 1030A en un círculo formando un octágono regular ($\pm 45º, \pm 135º, \pm 90º, 0º, 180º$) más un diseño estándar ITU 5.0 ($\pm 30º, \pm 110º, 0º$), con el altavoz central perteneciente a ambos conjuntos (*figura 33*). Además, se utilizan 6 seis escenas de sonido y una séptima únicamente para entrenamiento. Dado que una posible debilidad de cualquier método no lineal es la reproducción de escenas con múltiples sonidos superpuestos, en todas las escenas hay 3 o más fuentes de sonido superpuestas y tienen una duración entre 10 y 17 segundos.

Como señales de referencia se utiliza una fuente de sonido por altavoz, enrutada a un subconjunto de los doce altavoces disponibles. En las 6 escenas, aproximadamente la mitad de las fuentes se colocan en los altavoces del octágono y la otra mitad en altavoces pertenecientes al diseño 5.0.

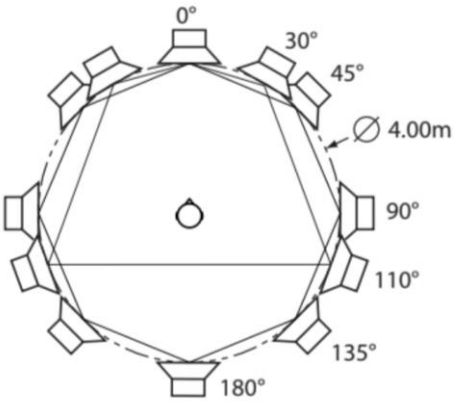



*Figura 33 – Diseño de altavoces 5.0 y octágono regular [21]*

Se prueban los siguientes sistemas comparándolos con una referencia:

1. 1-8: primer orden, decodificado con max-$R_E$ a octágono
2. 3-8: tercer orden, decodificado con max-$R_E$ al octágono
3. H-3-8: primer orden, mezclado a tercer orden usando HARPEX, tercer orden decodificada con max-$R_E$ a octágono
4. H-8: primer orden, decodificado con HARPEX y paneo por pares de 3 dB a octágono
5. H-5: primer orden, decodificado con HARPEX y paneo por pares de 3 dB a ITU 5.0
6. REF: Referencia oculta

La evaluación consiste en pedir a una serie de participantes una calificación para cada una de las 6 señales en cada escena en una escala de 0 a 100 de manera que de 0 a 20 se califica como “malo”, de 20 a 40 como “pobre”, de 40 a 60 como “justo”, de 60 a 80 como “bueno” y de 80 a 100 como “excelente”. En [21] participaron 19 sujetos con una duración total de la prueba de 35 minutos.

### 3.2.2.6 Resultado de la prueba

La diferencia en la puntuación entre los sistemas H-5, 3-8 y H-8 no es estadísticamente significativa pero las diferencias entre otros sistemas sí. La diferencia entre los sistemas H-5 y H-8 es casi significativa, con un p-valor de 0,062 en una prueba ANOVA de medidas repetidas (ver 2.1.2.2 a) (*figura* 34).

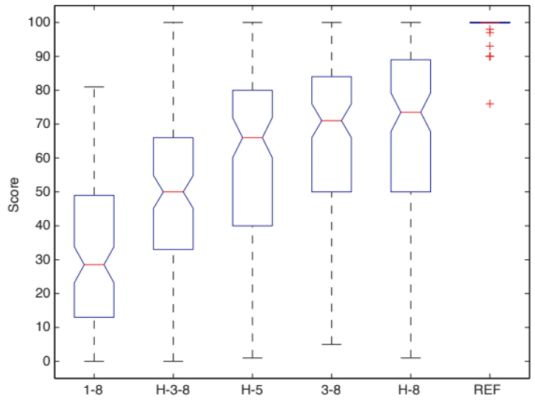



*Figura 34 – Resultados en las 5 escenas y 19 participantes*

Al igual que con cualquier prueba de comprensión auditiva, puede cuestionarse la generalidad de los resultados, dado que solo se pueden presentar a los oyentes unos pocos sonidos de prueba. La debilidad más probable de los métodos no lineales es su reproducción de múltiples fuentes de sonido superpuestas, y las escenas que se probaron estaban considerablemente más "ocupadas" que la típica escena de sonido que se encuentra en la radiodifusión o las telecomunicaciones. Por tanto, los resultados de los sistemas que utilizan HARPEX deben ser estimaciones conservadoras.

También se puede argumentar que el sonido ambisónico suena diferente al paneo por pares, y que la señal de referencia, al ser un caso especial de paneo por pares, sesga la prueba a favor de esa función de paneo. Esta fue la razón para la inclusión del sistema H-3-8, donde la función de panoramización es idéntica a la del decodificador max-$R_E$ de tercer orden. El sistema H-3-8 obtuvo una puntuación más baja que el sistema H-8. Se puede especular que la decodificación max-$R_E$y HARPEX introducen cada una sus propias diferencias con respecto a la señal de referencia, y que estas se acumulan en una mayor diferencia general en el sistema H-3-8. Esta teoría podría probarse con una prueba de comparación por pares.

Probablemente el resultado más sorprendente es la alta puntuación del sistema H-5 ya que las señales de referencia utilizan excesivamente los altavoces del conjunto 5.0. Este efecto es difícil de cuantificar, ya que diferentes fuentes de sonido atraen cantidades distintas de atención, no solo por diferencias en amplitud y espectro, sino también por diferencias en propiedades extrínsecas. Una comparación rigurosa de diferentes diseños de altavoces requiere un mayor número de altavoces, de modo que la referencia puede reproducirse tanto en otros altavoces como en cualquiera de los sistemas.

Se puede concluir que el método propuesto en [21] proporciona un medio para reproducir material de primer orden en configuraciones de altavoces grandes con una definición espacial mejorada y un punto óptimo más grande de lo que es posible con el otro método que se probó. Además, en una configuración 5.0 se obtienen resultados sorprendentemente buenos en comparación con una configuración de ocho altavoces.

Por último, los errores o artefactos que son perceptibles en un decodificador sencillo usando HARPEX se pueden suprimir a niveles inaudibles de forma segura sin perder cantidades notables de nitidez en la escena.

#### 3.2.2.7 Plugin

[21] HARPEX tiene su propio plugin (Harpex-X) para la mezcla ascendente de audios disponible en VST, AAX y formato AU para 32- y 64-bit Windows y Mac OS X. Con los plugins de Harpex, los creadores de contenido utilizan esta tecnología para transformar grabaciones de campo de sonido en A-format, B-format o AmbiX en los formatos envolvente estándar y envolvente 3D (utilizados en cine y televisión), estéreo coincidente, no coincidente y binaural, y AmbiX de orden superior para aplicaciones de realidad virtual y realidad aumentada (*figura 35*). Harpex procesa el audio con matrices para grabaciones de campo sonoro y esto hace que pueda producir decodificaciones envolventes con una gran separación de canales y un punto central (óptimo) mayor. Esto permite que las grabaciones se puedan utilizar en lugares como cines, Dolby Atmos, IMAX y Auro 3D. También, gracias a la separación de canales, se puede eliminar las fases de forma virtual, es decir, evitar la coloración desagradable del sonido que cambia a medida que los oyentes se mueven dentro del espacio de escucha.

Otra funcionalidad del plugin es que se pueden generar retrasos dependientes de la dirección como lo hacen los micrófonos casi coincidentes. Además, con el modo de decodificación binaural, Harpex puede producir estéreo adaptado a auriculares similar a las grabaciones de cabeza artificial. En aplicaciones de realidad virtual donde el campo de sonido debe rotarse para que coincida con la orientación de la cabeza del usuario, el formato de audio estándar es AmbiX. Harpex puede producir hasta una salida AmbiX de tercer orden a partir de grabaciones de campo sonoro.

Si necesita aislar una sola fuente de sonido o dividir una grabación en pistas de un solo instrumento, hay un modo que permite hacerlo en la posproducción (modo *shotgun*) seleccionando las fuentes de sonido que se quieran aislar.


*Figura 35 – Upmixing en HARPEX-X*

## 3.2.3 COMPASS

### 3.2.3.1 Introducción

La grabación, el procesamiento y la reproducción de sonido espacial modernos se están alejando de los formatos de canal basados en la configuración de reproducción, para pasar a sistemas flexibles y capaces de distribuir adecuadamente la grabación de sonido espacial a configuraciones arbitrarias. En Ambisonics, la localización de los sonidos en FOA puede ser imprecisa, pero HOA tiene un coste mayor en cuanto a ancho de banda [22].

Para mejorar las limitaciones existentes en FOA, tal y como se explica en los apartados anteriores, se han desarrollado algunos métodos dependientes de la señal, todos ellos operando en el dominio de la frecuencia temporal y diferenciándose en su modelo de campo sonoro y en la estimación de los parámetros (DirAC (3.2.1) y HARPEX (3.2.2)). Estos métodos han resultado eficaces en una gran variedad de escenarios sonoros, permitiendo además modificaciones espaciales flexibles útiles de la escena y el upmixing de FOA a HOA.

Existe un nuevo enfoque para el análisis y la síntesis de las señales Ambisonics, denominado COMPASS (Coding and Multidirectional Parametrization of Ambisonic Sound Scenes). A diferencia de los métodos anteriores, utiliza un modelo acústico general de la escena sonora de múltiples señales de fuentes en primer plano y un componente ambiental de fondo que no es necesariamente isotrópico o difuso. Es un marco para el procesamiento de audio espacial paramétrico de escenas de sonido capturadas en el formato Ambisonics.

La *tabla 4* muestra la comparación del modelo COMPASS en con otras técnicas paramétricas (siendo M es el número de canales y los sectores angulares divisiones del campo de sonido capturado dentro de los cuales se estiman los parámetros del modelo de onda plana):

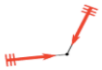
| Método | Entrada | Canales | Modelo | |
|---|---|---|---|---|
| **DirAC** | FOA | 4 | 1 componente origen +<br>1 componente iso. difusa |  |
| **HARPEX** | FOA | 4 | 2 componentes origen |  |
| **HO-DirAC** | HOA | 9+ | M sectores fuente +<br>M componentes difusos del sector |  |
| **COMPASS** | FOA/HOA | 4+ | ≤ M/2 componentes de origen +<br>componente espacial ambiental |  |

*Tabla 4 - Comparación de métodos de representación paramétrica de Ambisonics*

Como se puede observar, el método no se limita sólo a FOA, y es más apropiado para la modificación espacial de los componentes de la escena sonora que DirAC, ya que los flujos DirAC de orden superior no corresponden necesariamente a los componentes reales de la fuente en la escena.

COMPASS no requiere preprocesamiento en un escenario no disperso, y es computacionalmente eficiente y capaz de operar en tiempo real, lo que es especialmente importante para la reproducción con auriculares donde se puede emplear el seguimiento de cabeza (head-tracking). Opera sobre señales Ambisonics en lugar de señales de micrófono, y pretende preservar todos los componentes de la escena sin rechazar interferencias, ambiente y reverberación. Además, debido a la generalidad del formato de los armónicos esféricos de la señal, puede aplicarse tanto al audio Ambisonics generado a partir de un software de mezcla, como a las grabaciones de sonido espacial.

### 3.2.3.2 Método COMPASS

El método COMPASS (*figura 36*), se basa en el modelo general de la escena sonora como una combinación de múltiples señales de fuentes direccionales K<M, capturadas en $a_s$ y un componente adicional sin direccionalidad clara capturada en $a_d$ que incluye el sonido ambiental, fuentes distribuidas de manera incoherente y reverberación tardía. Es decir, el método estima múltiples componentes de sonido directo en cada bloque de tiempo-frecuencia, y un componente ambiental (que también puede ser espacial y tener direccionalidad) que captura la reverberación y otros sonidos difusos.

$$a\,(t,f) = a_s\,(t,f) + a_d\,(t,f) = Y_s\,(t,f)\;s\,(t,f) + a_d\,(t,f),$$

donde $Y_s\,(t,f)\;s\,(t,f)$ es la codificación de un conjunto de K señales $s$ de fuente de onda plana.

COMPASS tiene como objetivo estimar los parámetros de estas dos componentes y explotarlos durante la síntesis y la reproducción. Al igual que la mayoría de los métodos de codificación de

audio espacial, COMPASS opera en señales transformadas en tiempo-frecuencia (STFT, banco de filtros…).

Se asume tanto que la parte ambiental y direccional están decorrelacionadas, como las señales fuente entre sí. Contrariamente a la mayoría de los métodos de análisis, no se asumen necesariamente condiciones difusas isotrópicas en la componente ambiental.

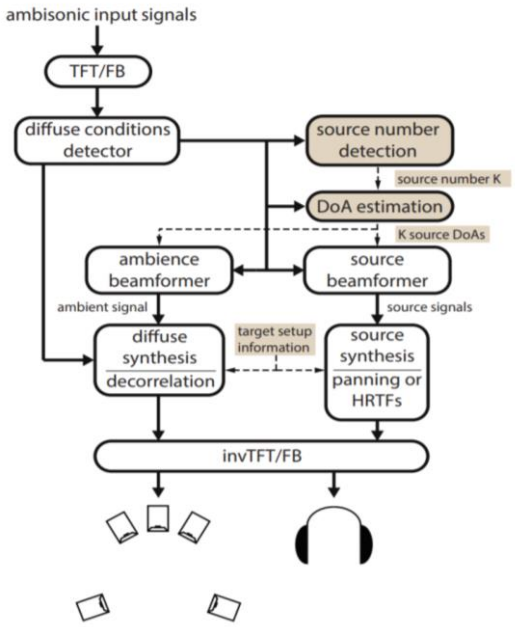



*Figura 36 – Diagrama de bloques de COMPASS*

Mediante un proceso de análisis se estima la DoA (dirección de llegada) de origen, la fuente, las señales ambientales y la potencia de las señales (Pd). Con unas funciones de distribución calculadas a partir de las DoA de las componentes, se hallan las HRTF y con ello se espacializan las señales.

### 3.2.3.3. Evaluación

Para evaluar el rendimiento de COMPASS, en [22] se realiza una prueba MUSHRA, para la reproducción con auriculares simulando 5 escenas variando el número de fuentes para condiciones tanto anecoicas como reverberantes. Las escenas anecoicas sólo incluyen retardos de propagación y codificación direccional para las distancias de la fuente y DoA, mientras que las reverberantes introducen una reverberación completa.

Para obtener los tiempos de reverberación objetivo que van de 0,6 a 1,2 segundos (por banda de frecuencia), se ajustan los perfiles de absorción. Para cada fuente, con una convolución del filtro de propagación con la HRTF apropiada se genera la SRIR binaural (*spatial room impulse*

*response*). Al convolucionar cada SRIR con las señales de la fuente, se genera la versión binaural de referencia de la escena de sonido.

Después los filtros de propagación se codifican para Ambisonics de tercer orden o TOA, obteniéndose la SRIR Ambisonics. La convolución con las señales de la fuente da como resultado una codificación TOA de la escena sonora general. Para obtener una versión FOA se mantienen sólo los primeros 4 canales de las señales TOA.

Mediante una serie de oyentes y tras una recreación de escenas de sonido, estas se decodifican de manera binaural tanto para FOA como para TOA, y además se procesan por el método COMPASS.

Según las puntuaciones obtenidas, COMPASS con entrada FOA tiene una calidad general similar a la decodificación Ambisonics lineal con entrada TOA. Además, COMPASS tiene un timbre más cercano a la referencia en comparación con la decodificación Ambisonic, tanto para la entrada FOA como para la TOA. Los resultados indican que el método mejora la calidad percibida espacial, tímbrica y general en comparación con Ambisonics con el mismo orden de entrada.

#### 3.2.3.4. Plugins

[12] COMPASS tiene una colección de complementos de audio VST flexibles para la producción, manipulación y reproducción de audio espacial, que fue desarrollado por el Dr. Anchontis Politis con contribuciones del Dr. Sakari Tervo y Leo McCormack [22]. El componente ambiental también es espacial y puede tener direccionalidad, contrariamente a modelos anteriores que lo obligan a ser isotrópico. Los complementos VST aplican este marco a diferentes tareas de producción de audio espacial. Todos los plugins cumplen con la convención de orden ACN (Ambisonic Channel Number) (1.1) y ofrecen soporte para escalas tanto ortonormalizadas (N3D) como semi-normalizadas (SN3D) (AmbiX utiliza ACN/SN3D). El orden máximo de Ambisonics para estos plugins es 3.

A continuación, se explican brevemente los tipos de plugins de COMPASS.

##### 3.2.3.4.1 Decodificador

El plugin decodificador COMPASS *(figura 37*) es un decodificador paramétrico para Ambisonics de primer, segundo y tercer orden para configuraciones de altavoces arbitrarias. Tiene funcionalidades de ángulos de altavoz especificados por el usuario, control de equilibrio entre los componentes directo y ambiental, control de mezcla entre decodificación paramétrica y ambisónica, control “diffusse-to-direct” para dar más protagonismo a un componente u otro, y el control “Linear-to-parametric”, que permite al usuario mezclar la salida entre la decodificación Ambisonic lineal estándar y la decodificación paramétrica COMPASS.

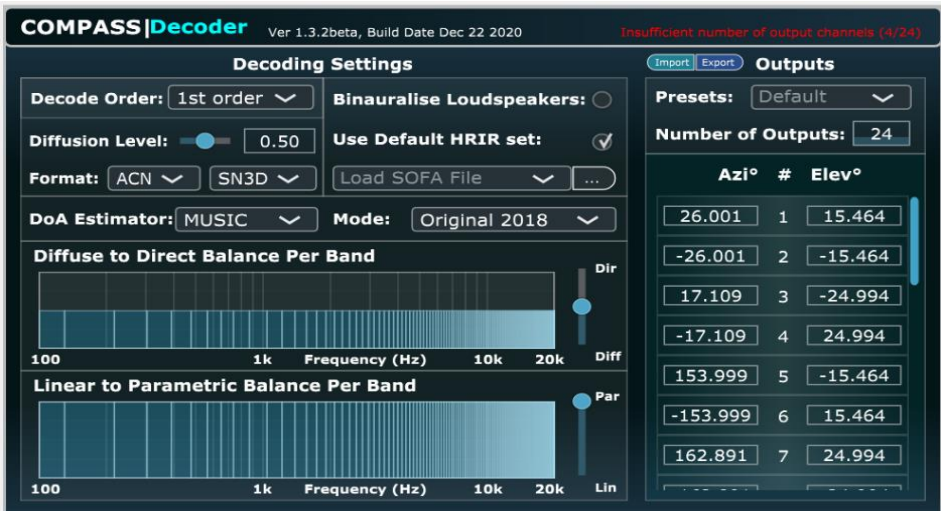



*Figura 37 – Plugin COMPASS Decoder*

### 3.2.3.4.2 Binaural

Ésta es una versión optimizada del decodificador COMPASS para reproducción binaural desarrollada por Leo McCormack y Archontis Politis. Utiliza filtros binaurales (HRTF), que pueden ser proporcionados por el usuario y personalizados con el formato SOFA (*figura 38).*

Además, es posible añadir ángulos de rotación de un head tracker (dispositivo que hace el seguimiento de cabeza) en un puerto especificado por el usuario.

Esta versión está pensada principalmente para la reproducción binaural con seguimiento de la cabeza de contenido Ambisonics a velocidades de actualización interactivas, generalmente junto con una pantalla montada en la cabeza (HMD). Los parámetros de promediado se pueden utilizar para hacer que el análisis paramétrico y la síntesis respondan, proporcionando al usuario un medio para ajustarlos de manera óptima para una escena de sonido en particular.

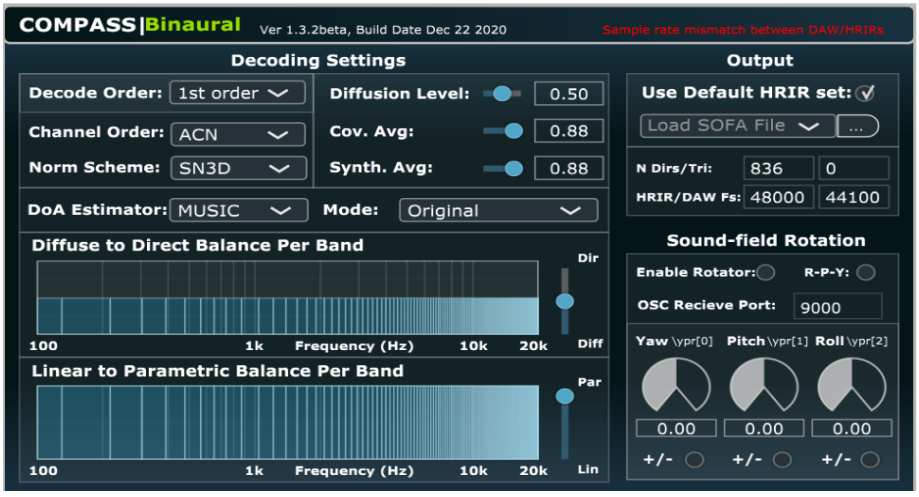



*Figura 38 – Plugin COMPASS Binaural*

### 3.2.3.4.3 Tracker

Este VST se basa en el análisis espacial realizado por el marco de codificación y parametrización multidireccional de escenas de sonido ambisónico (COMPASS), pero en lugar de utilizar la información para sintetizar señales de altavoz o binaurales, se emplea un rastreador de fuentes múltiples para asociar las direcciones estimadas con sus fuentes u objetivos correspondientes. Por lo tanto, este VST se puede utilizar para visualizar la trayectoria de las fuentes de sonido presentes en una escena de sonido Ambisonic (*figura 39*).

Opcionalmente, un formador de haz puede dirigirse a cada dirección de destino y emitirse como señales individuales, o como una binauralización de estos "tallos" individuales (espacializados hacia sus respectivas direcciones de destino).

Como el seguimiento de múltiples fuentes ha sido un tema de investigación activo durante varias décadas, todavía se considera una tarea difícil y por ello existe una curva de aprendizaje bastante grande para ajustar de manera efectiva los parámetros para una escena o distribución de sonido específica.


*Figura 39 - Plugin COMPASS Tracker*

### 3.2.3.4.4 Upmixer

Este VST emplea COMPASS para la tarea de mezclar una grabación Ambisonic de orden inferior con una grabación Ambisonic de orden superior. Está destinado a usuarios que ya están

trabajando con un flujo de trabajo de decodificación Ambisonic lineal preferido de contenido Ambisonic de orden superior y desean combinar material Ambisonic de orden inferior con una mayor resolución espacial (*figura 40*). Se puede mezclar material de primer, segundo o tercer orden (4, 9,16 canales) hasta material de séptimo orden (64 canales).

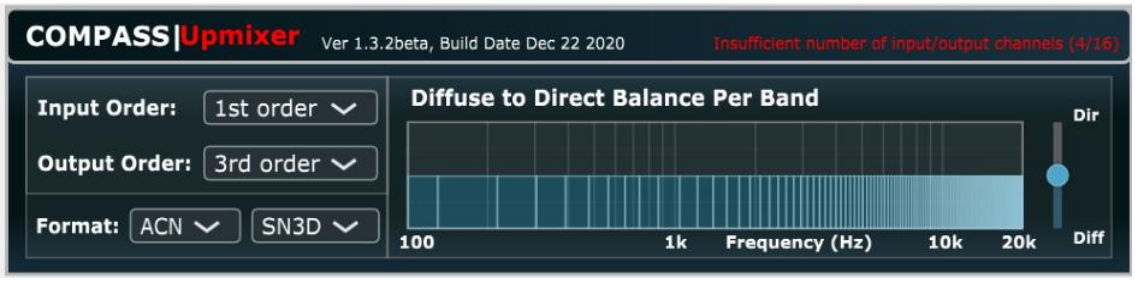



*Figura 40 – Plugin COMPASS Upmixer*

### 3.2.3.4.5 Sidechain

En este VST se aplica el análisis COMPASS a dos escenas de sonido diferentes (escena A, con canales comprendidos del 1 al 16 en escena B, con canales del 17 al 32), y utiliza los parámetros espaciales estimados de una escena para manipular las señales de la otra escena (*figura 41*). Si las escenas A y B son iguales, entonces el complemento es funcionalmente idéntico al plugin de upmixer (3.2.3.4.4)

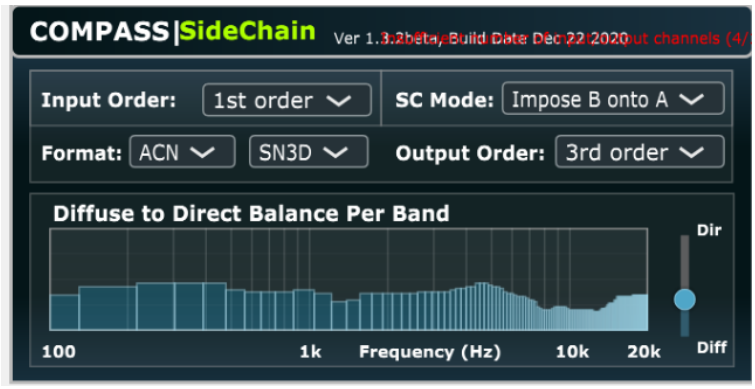



*Figura 41 – Plugin COMPASS Sidechain*

# 4. JAULAB

Jaulab, inicialmente denominado *SONORIZART3,* es un proyecto impulsado por la Universidad Pública de Navarra en 2020 en colaboración con el Conservatorio Superior de Música de Navarra y la Escuela de Arte y Superior de Diseño de Pamplona, con el fin de investigar, experimentar y conocer las características del sonido tridimensional (*figura 42*).

El objetivo principal de este proyecto es generar una esfera de sonido mediante la utilización de altavoces DIY para sumergir al oyente en una experiencia de sonido totalmente envolvente y tridimensional. De esta manera, pretende mostrar el trabajo que realiza el Laboratorio de Acústica de la UPNA a un público menos habituado a este tipo de sonido y acercarlo a la experiencia 3D mediante la recreación de escenas grabadas o creadas. Dicha esfera, instalada en abril del 2021, se encuentra en el laboratorio de acústica de la Universidad y está construida mediante el cableado y anclaje de 24 altavoces DIY (*figura 43*). Tiene un radio de 1,45 metros y mide 2,70 metros de alto. Es una herramienta óptima para simular el campo sonoro y las técnicas de upmixing comentadas en los apartados anteriores.

*Figura 42 – Esfera de 24 altavoces*

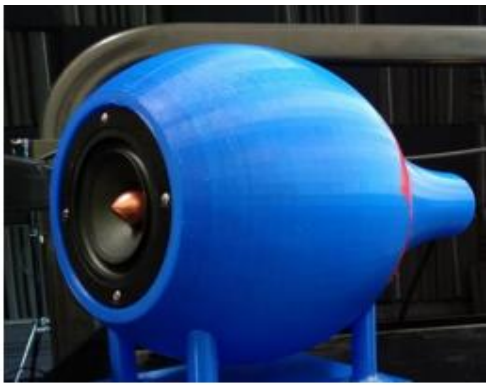


*Figura 43 – Altavoz DIY*

La *tabla 5* contiene las posiciones en detalle de cada uno de los altavoces instalados en la esfera, donde los ángulos azimut y elevación están expresados en grados.

| Canal | Azimut | Elevación |
|---|---|---|
| 1 | -34,4 | -25,0 |
| 2 | -25,5 | 15,5 |
| 3 | -19,3 | 60,0 |
| 4 | 6,2 | -60,0 |
| 5 | 12,5 | -15,5 |
| 6 | 21,3 | 25,0 |
| 7 | 55,6 | -25,0 |
| 8 | 64,5 | 15,5 |
| 9 | 70,7 | 60,0 |
| 10 | 96,2 | -60,0 |
| 11 | 102,5 | -15,5 |
| 12 | 111,3 | 25,0 |
| 13 | 145,6 | -25,0 |
| 14 | 154,5 | 15,5 |
| 15 | 160,7 | 60,0 |
| 16 | -173,8 | -60,0 |
| 17 | -167,5 | -15,5 |
| 18 | -158,7 | 25,0 |
| 19 | -124,4 | -25,0 |
| 20 | -115,5 | 15,5 |
| 21 | -109,3 | 60,0 |
| 22 | -83,8 | -60,0 |
| 23 | -77,5 | -15,5 |
| 24 | -68,7 | 25,0 |

*Tabla 5 – Posición de los altavoces en la esfera*

La *figura 44* representa la posición de los altavoces en la esfera de una manera más visual.

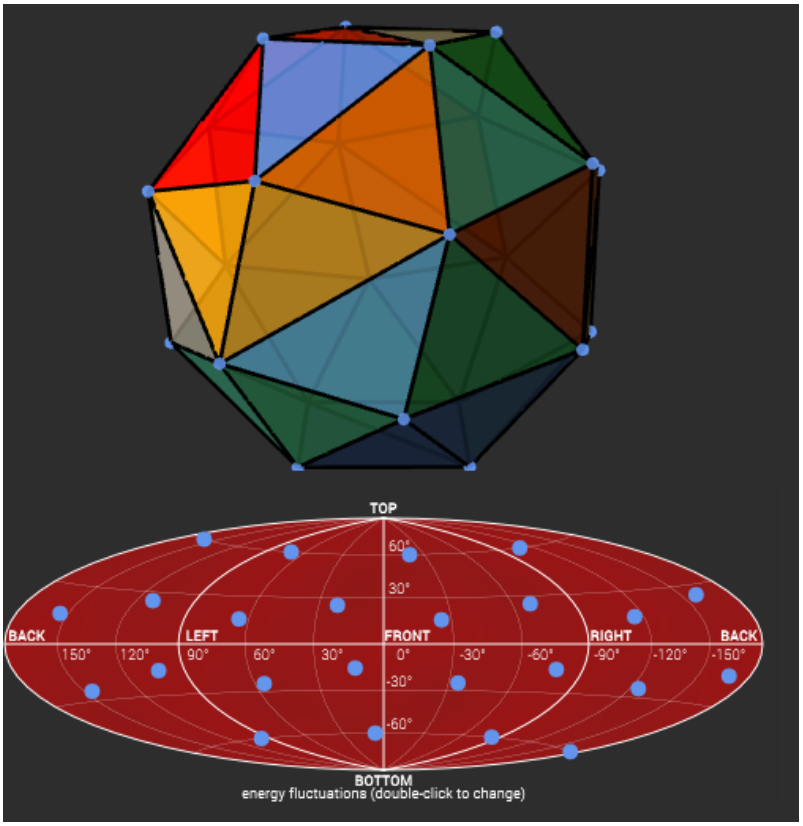



*Figura 44 – Colocación de los altavoces en la esfera*

# 5. Listening test

## 5.1 Introducción

El experimento en el que se basa este proyecto consiste en realizar un conjunto de test subjetivos a una serie de participantes para detectar las diferencias o prestaciones de cada algoritmo de upmixing: DirAC, HARPEX y COMPASS. También se va a valorar el grado de diferenciación entre estímulos y algoritmos dependiendo de si el participante es músico profesional, aficionado o sin formación musical.

El test consiste en dos pruebas. La primera se realiza en ordenador mediante un programa diseñado para este proyecto, donde los estímulos se presentan mediante auriculares en formato binaural. En la segunda prueba, como escenario se utiliza la esfera de 24 altavoces explicada en (4), de manera que los participantes realizan la prueba físicamente en el recinto.

Para la realización de la prueba, se imita la técnica de calificación y recopilación de datos empleada en HARPEX (3.2.2.5) y se someten a 39 oyentes a un test MUSHRA por comparación de 7 estímulos. Se ha escogido esta técnica ya que no es necesario realizar el test a un gran número de oyentes a diferencia de la prueba MOS, y porque los estímulos no tienen por qué emitirse necesariamente en el idioma nativo de los oyentes (las voces serán emitidas en inglés). Además, este método tiene la ventaja de visualizar muchos estímulos al mismo tiempo, de forma que el sujeto puede verificar cualquier comparación entre ellos directamente.

Al sujeto se le presentan 7 estímulos de manera aleatoria, una referencia escondida y un ancla. La referencia es el audio que van a comparar con el resto de estímulos. Como referencia se utilizan dos y cuatro audios de voces cantadas en inglés (fever y demode respectivamente) en unas posiciones específicas. Dichas posiciones se encuentran al detalle en las *tablas 6 y 7*. A los participantes se les plantea la cuestión de cómo de diferente perciben cada estímulo con respecto a la referencia, en términos de desvío o deslocalización de las fuentes.

Además, se dispone de un “ancla”, algo que se parece poco a la referencia y esta sirve para graduar cómo de similares son el resto de estímulos. El propósito del ancla es hacer que la escala esté más cerca de una "escala absoluta", asegurándose de que los artefactos menores no se califiquen como de muy mala calidad. El ancla consiste en la misma señal de audio emitida por los 24 altavoces de forma decorrelacionada (3.2.1.6.2), tanto en el audio de fever como en el de demode.

| **FEVER** | | |
|---|---|---|
| Canales | Azimut | Elevación |
| 1 5 | -34,4<br>12,5 | -25<br>-15,5 |
| 5 6 | 12,5<br>21,3 | -15,5<br>25 |
| 6 17 | 21,3<br>-167,5 | 25<br>-15,5 |
| 9 22 | 70,7<br>-83,8 | 60<br>-60 |

*Tabla 6 – Posiciones de cada escena audios fever*

| DEMODE | | |
|---|---|---|
| Canales | Azimut | Elevación |
| 1 2 5 6 | -34,4 | -25 |
| | -25,5 | 15,5 |
| | 12,5 | -15,5 |
| | 21,3 | 25 |
| 3 7 14 22 | -19,3 | 60 |
| | 55,6 | -25 |
| | 154,5 | 15,5 |
| | -83,8 | -60 |

*Tabla 7 – Posiciones de cada escena audios demode*

Los 7 estímulos que se utilizan en el experimento son audios de voces cantadas en FOA, TOA (Ambisonics de tercer orden), HARPEX, DirAC, COMPASS, la referencia y el ancla. Para que dispongan de 24 canales, dichos audios deben estar codificados y decodificados previamente. Para ello, se dispone del siguiente material:

- Aplicación Reaper
- Plugin SPARTA (binauraliser)
- Plugin IEM (Multi encoder, decoder).
- Plugin HARPEX
- Plugin COMPASS
- Librería DirAC (Matlab)
- 4 audios base (2 voces cantadas y 2 habladas)
- Matlab (Loudness)

Mediante dicho material, el método empleado para procesar los estímulos y obtener su formato en 24 canales y binaural en FOA, TOA, HARPEX, DirAC y COMPASS se refleja en la *figura 45*.

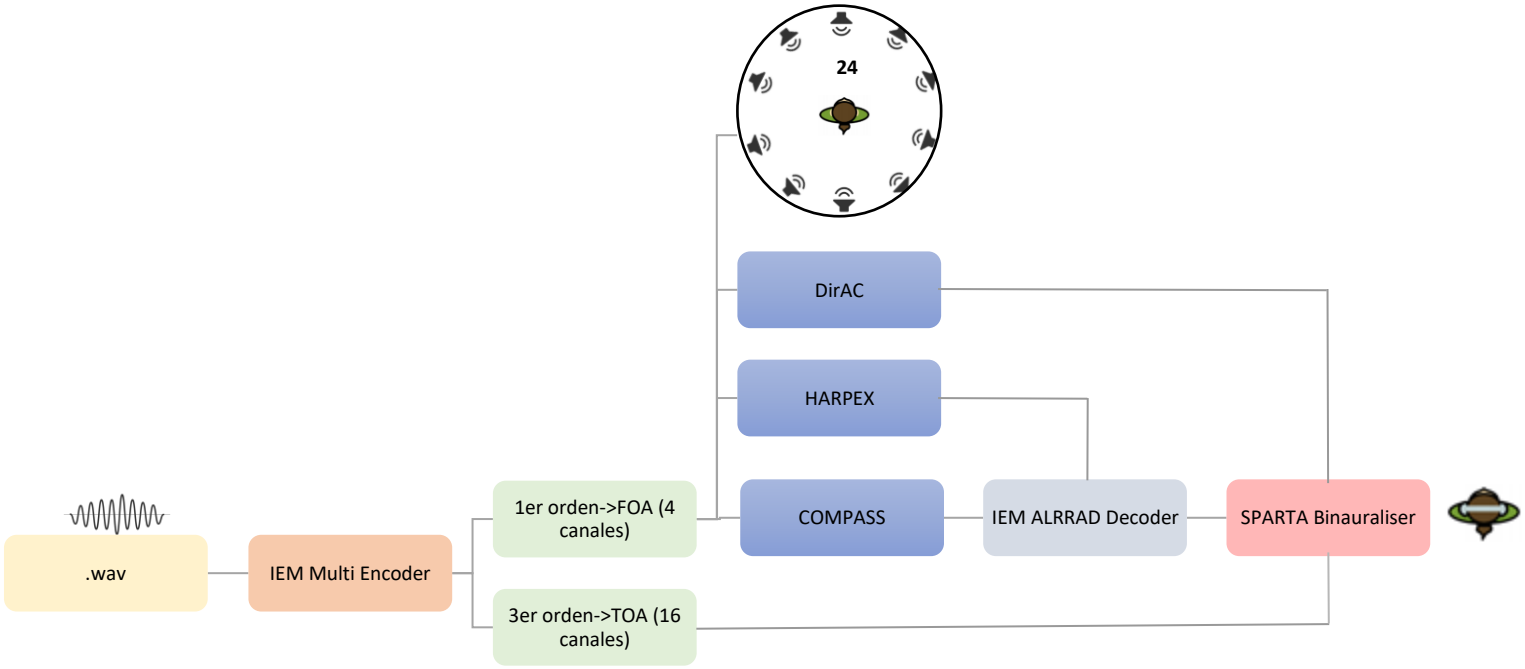



*Figura 45 – Método empleado para obtener los estímulos*

El procesado de los audios en HARPEX y COMPASS para obtener 16 canales se realiza mediante sus correspondientes plugins, siendo finalmente decodificados a 24 canales y binauralizados para escuchar por auriculares con el plugin de SPARTA Binauraliser. DirAC, sin embargo, requiere de una librería independiente en Matlab, la cual crea directamente los audios finales decodificados en 24 canales y después se binauralizan del mismo modo que el resto de los audios. Con todo ello, se obtienen los estímulos tanto para FOA, TOA y la referencia como para los 3 tipos de upmixers (DirAC, COMPASS y HARPEX).

Los estímulos procesados son audios de voces cantadas en inglés:

- Fever: dos voces cantadas a capella emitidas, para cada caso, por los altavoces de las posiciones indicadas en la *tabla 6*.
- Demode: cuatro voces cantadas a capella emitidas, para cada caso, por los altavoces de las posiciones indicadas en la *tabla 7*.

Las referencias para cada posición se crean en Reaper mediante la localización de los audios en los correspondientes canales que figuran en *las tablas 6 y 7* y con su posterior renderización a 24 canales.

El ancla, por su parte, se obtiene de una forma más sencilla emitiendo cada audio de igual manera por los 24 canales.

Para que se produzca una sensación de percepción del nivel igual en todos los estímulos, se ha modificado la sonoridad o loudness en todos los casos. Esta es una medida subjetiva de la intensidad con la que un sonido es percibido por el ser humano y depende de las propiedades de la fuente, del medio y del oyente.

Para que el promedio del sonido final en cada intervalo (izquierdo y derecho) sea igual en todos los estímulos, se aplica la función *integratedLoudness(audioIn,Fs)* en Matlab tras modificar el audio mediante el plugin SPARTA Binauraliser. Se ha determinado una sonoridad promedio para cada estímulo de 22.3±0.1 sones de manera que la sensación de intensidad sonora sea igual todos los casos.

La prueba consta de dos partes diferenciadas; el test binaural y el test en la esfera.

## 5.2 Test binaural

Al comienzo del test los participantes deberán posicionarse, según la relación que tengan con la música, en uno de los siguientes grupos:

1) Profesional
    a) Más de 10 años
    b) Menos de 10 años
2) Aficionado
    a) Más de 10 años
    b) Menos de 10 años
3) Lego: sin formación.

De los 39 sujetos que participan, 13 son profesionales, 13 aficionados y 13 lego, por lo que en el apartado 5.5 se evalúan los 3 grupos por igual.

También se hace distinción entre los participantes que han tenido o no problemas de audición.

Esta primera prueba consiste en la realización, por parte de los oyentes, de un test MUSHRA creado específicamente para este proyecto mediante auriculares con 6 escenas diferentes.

En cada escena tiene cabida un estímulo de referencia, que es el audio de las voces emitido por los 24 canales, y los diferentes estímulos binaurales en FOA, TOA, DirAC, HARPEX, COMPASS y el ancla, tal y como muestra la *figura 46*. Del total de las 6 escenas, las 4 primeras se realizan escenas con las voces de fever y 2 últimas con las de demode.

Los 7 estímulos se lanzan en cada escena de manera aleatoria, no pudiendo repetirse el mismo estímulo una escena. El oyente tiene que calificar en una escala del 0 al 100 cuánto se parecen los diferentes estímulos a la referencia, bajo el siguiente criterio:

- Del 0 al 20 - Malo: voces completamente fuera de su posición original o imposible localizarlas.
- Del 20 al 40 - Pobre: desvío sustancial, ensanchamiento claro y/o dificultad para localizarlas.
- Del 40 al 60 - Justo: Las voces se desvían de su posición original y/o se ensanchan ligeramente.

- Del 60 al 80 - Bueno: Ligero cambio en la posición original de las voces.
- Del 80 al 100 - Excelente: No percibo la diferencia.

Para puntuar los estímulos, el oyente puede hacerlo tanto con una consola mezcladora de audio como en el propio programa mediante el ordenador.

El participante registra su evaluación de la calidad utilizando los cursores de la visualización electrónica (*figura 46*), asignando una puntuación por cada estímulo. Se permite el ajuste individual del nivel de audición por un sujeto en una sesión, pero debe limitarse a la gama de ±4 dB respecto al nivel de referencia definido en la Recomendación UIT-R BS.1116.

Al terminar de calificar cada escena, el participante debe pulsar el botón de "siguiente", hasta finalizar el total de 6 escenas.

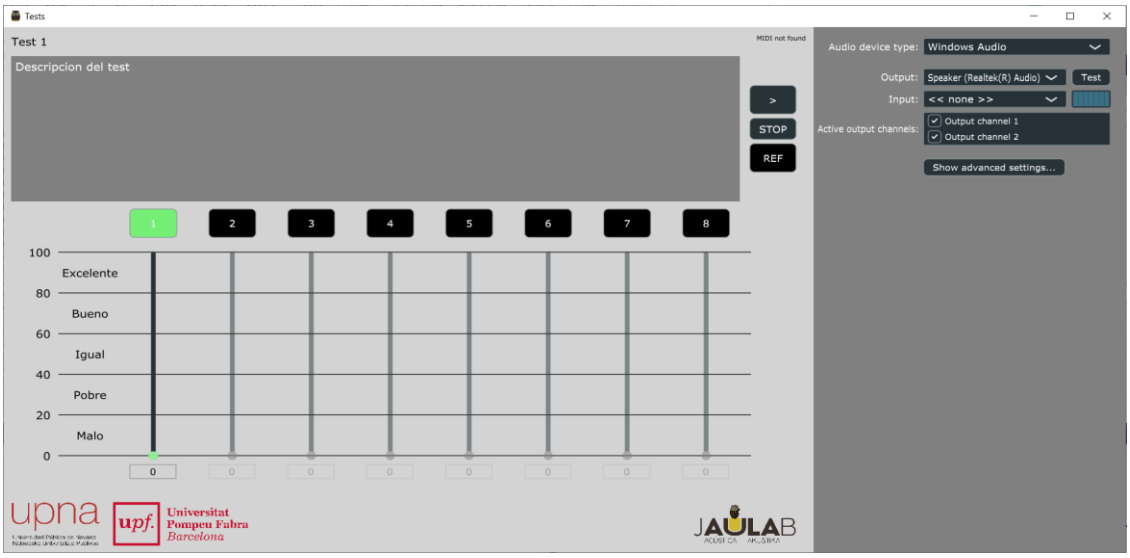



*Figura 46 – Interfaz del test binaural*

Una vez realizadas las 6 escenas, el programa finaliza y se guardan los datos obtenidos en un archivo de extensión .xml. De esta manera, es más sencillo ordenar y analizar los datos obtenidos en ambos test de cada participante.

Al finalizar el test, los sujetos pueden reflejar de manera opcional sus impresiones y valoraciones personales de la prueba.

## 5.3 Test esfera

En esta segunda parte de la prueba, los oyentes deben posicionarse en el interior de la esfera de 24 altavoces y se les realiza exactamente la misma prueba que con auriculares (ver *figura 47*). Los estímulos de cada escena se lanzan de manera aleatoria, por lo que el sujeto no puede basarse en los resultados obtenidos en el test binaural.

De igual manera que en el test binaural y con el mismo criterio, los oyentes deben valorar del 0 al 100 cómo de parecidos son los estímulos emitidos y el audio origen de referencia.

Los estímulos se puntúan uno a uno mediante la consola mezcladora de audio, visualizando en todo momento la interfaz del programa mediante un proyector.

Al finalizar el test los sujetos pueden reflejar de manera opcional sus impresiones y valoraciones personales de la prueba.

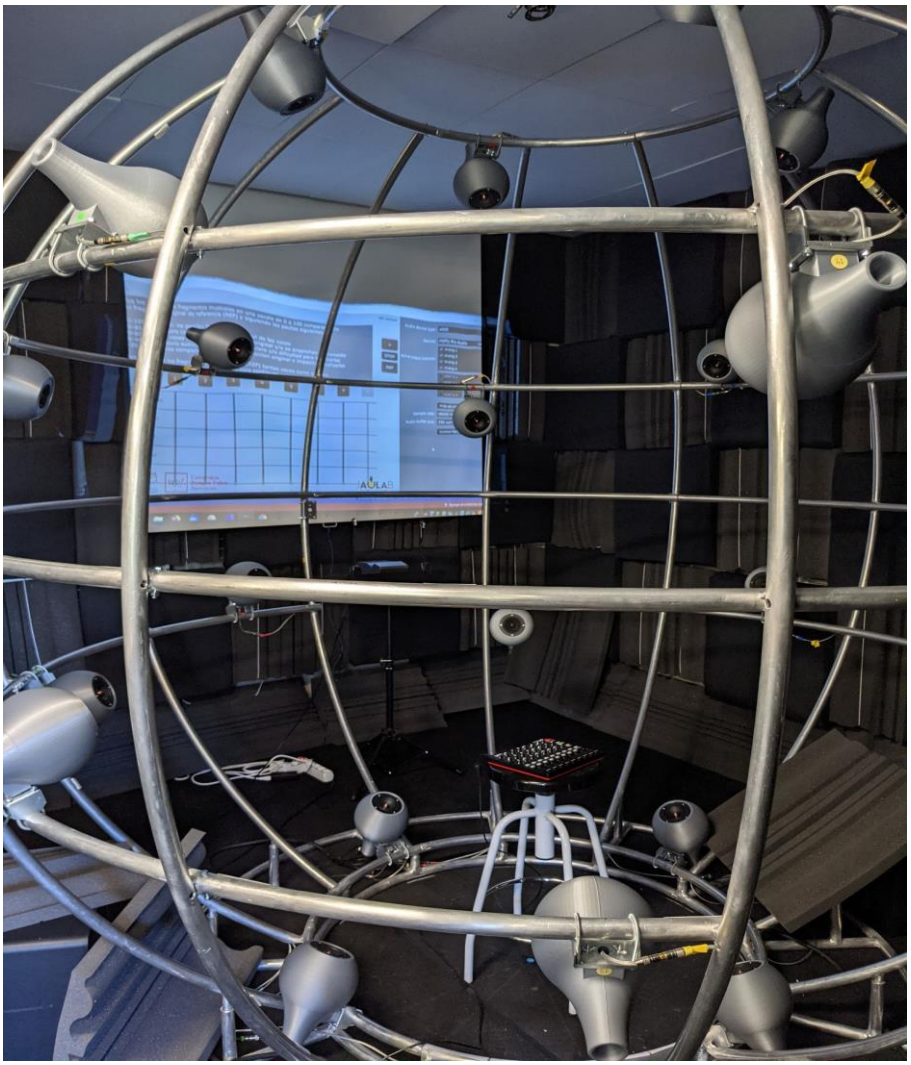


*Figura 47 – Escenario del test esfera*

## 5.4 Material empleado

Para la realización del test mediante auriculares, el material utilizado consiste en el siguiente:

- Auriculares Sennheiser HD600.
- Ordenador.
- Programa JAULAB creado para el proyecto.
- Controlador MIDI AKAI Midmix.

El material utilizado para el test mediante la esfera de altavoces es el siguiente:

- Esfera JAULAB de 24 altavoces.
- Controlador MIDI AKAI Midmix.
- Tarjeta sonido MOTU24AO
- Amplificadores DaytonAudio MA1240a
- Ordenador.
- Proyector.

## 5.5 Resultados obtenidos

Los resultados obtenidos se representan mediante diagramas de cajas y bigotes tal y como se hace en 3.2.2.6., los cuales muestran además las medianas y los cuartiles de los datos.

El diagrama de cajas y bigotes, también llamado boxplot, representa la siguiente información:

- El bigote superior corresponde al valor máximo de los datos.
- El bigote inferior se corresponde con el valor mínimo de los datos.
- La primera parte de la caja representa el primer cuartil, que el el valor mayor que el 25% de los valores de la distribución.
- El segundo cuartil es el valor de la variable que ocupa el valor central del conjunto, siendo además la mediana de la distribución.
- La parte final de la caja representa el tercer cuartil, que es el valor que sobrepasa al 75% de los valores de la distribución.

En los siguientes apartados se realizan varios análisis del experimento con enfoques diferentes, para llegar a las conclusiones pertinentes.

### 5.5.1 Análisis global por estímulo

En la *figura 48* se muestran los resultados globales del test binaural por cada tipo de estímulo en todas las escenas mediante boxplot.

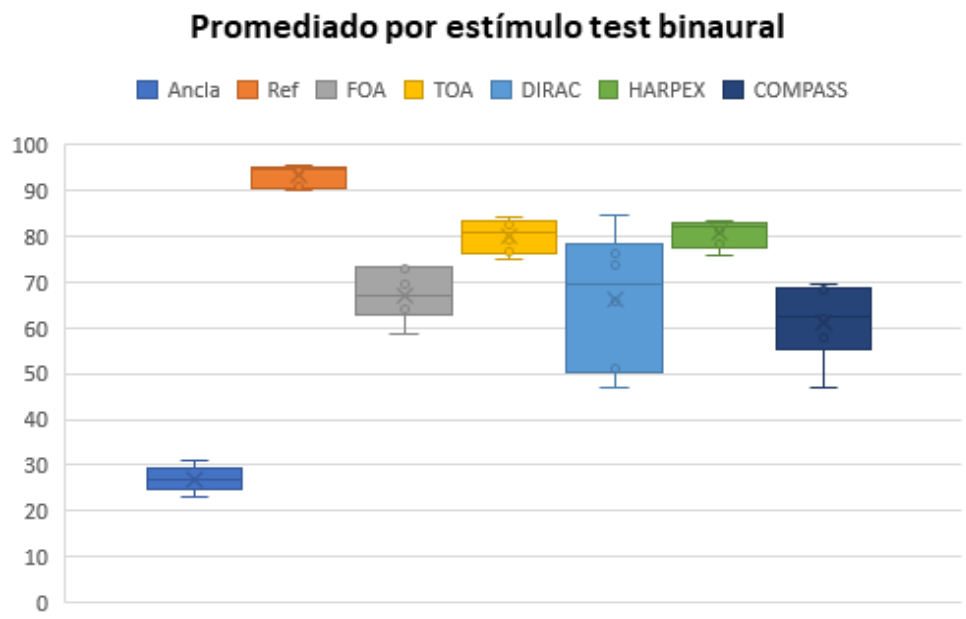



*Figura 48 – Resultados del test binaural por estímulo*

Tal y como era de esperar, el ancla es el estímulo que menor calificación obtiene de manera general, con una puntuación promedio de 27,2. Los oyentes perciben este audio como pobre, cuyas voces sufren un desvío sustancial, ensanchamiento claro y/o dificultad para localizarlas, ya que es el que se emite por igual por los 24 canales.

Por otro lado, según lo previsto, la referencia obtiene el mayor valor promedio en todas las escenas, con una puntuación media de 93,40. Los oyentes califican este estímulo como excelente (no perciben la diferencia), ya que es el mismo que el de referencia.

Las técnicas de upmixing obtienen resultados bastante diferentes, siendo HARPEX calificada con una mayor puntuación, con un promedio de 80,81. Sorprenden los resultados de COMPASS, el cual obtiene las peores valoraciones con una media de 61,25, la cual es incluso menor que la de FOA.

Como se puede observar en la *figura 48*, DirAC es la técnica que presenta más variaciones en la impresión subjetiva de los participantes. Con una puntuación media de 66,7, pero resultados muy cambiantes en cada escena, obtiene puntuaciones máximas y mínimas de 85 y 47 respectivamente dependiendo del tipo de escena.

En el test binaural, los audios que obtienen una mayor puntuación, y por tanto, los que se perciben de una manera más similar a la referencia, son los procesados mediante TOA y HARPEX, obteniendo puntuaciones medias de 80,34 y 81,81 respectivamente. Además, son los datos menos cambiantes por lo que por lo general todos los participantes les asignan buena puntuación.

En la *figura 49* se muestran los resultados globales del test en la esfera por cada tipo de estímulo en todas las escenas.

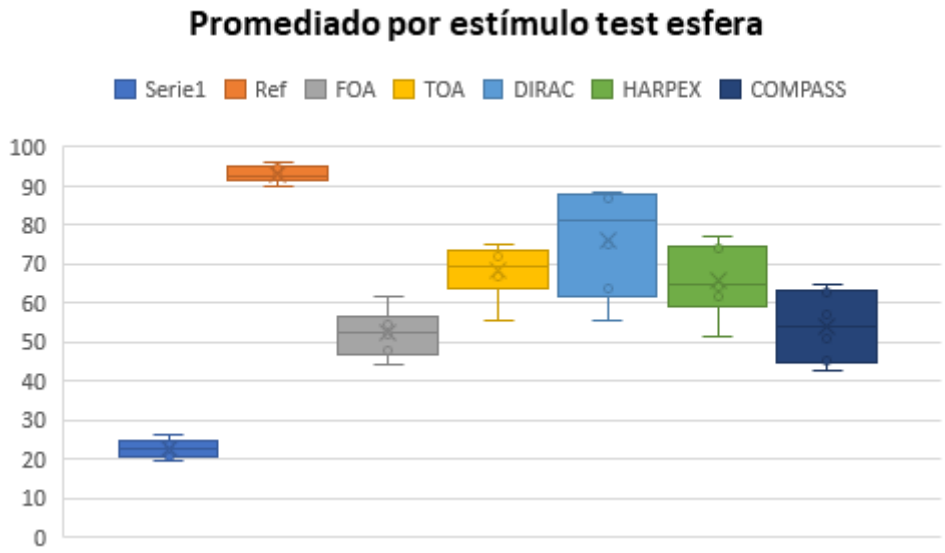



*Figura 49 – Resultados del test esfera por estímulo*

En el test realizado en la esfera de 24 altavoces, se puede observar que hay cambios significativos en las valoraciones globales con respecto al test mediante auriculares, ya que se detecta más fácilmente la localización de la procedencia (altavoces) de los estímulos.

En el test en la esfera, los audios que obtienen una mayor puntuación, y por tanto, los que se perciben de una manera más similar a la referencia, son los procesados mediante DirAC y TOA, obteniendo puntuaciones medias de 76,16 y 68,2 respectivamente.

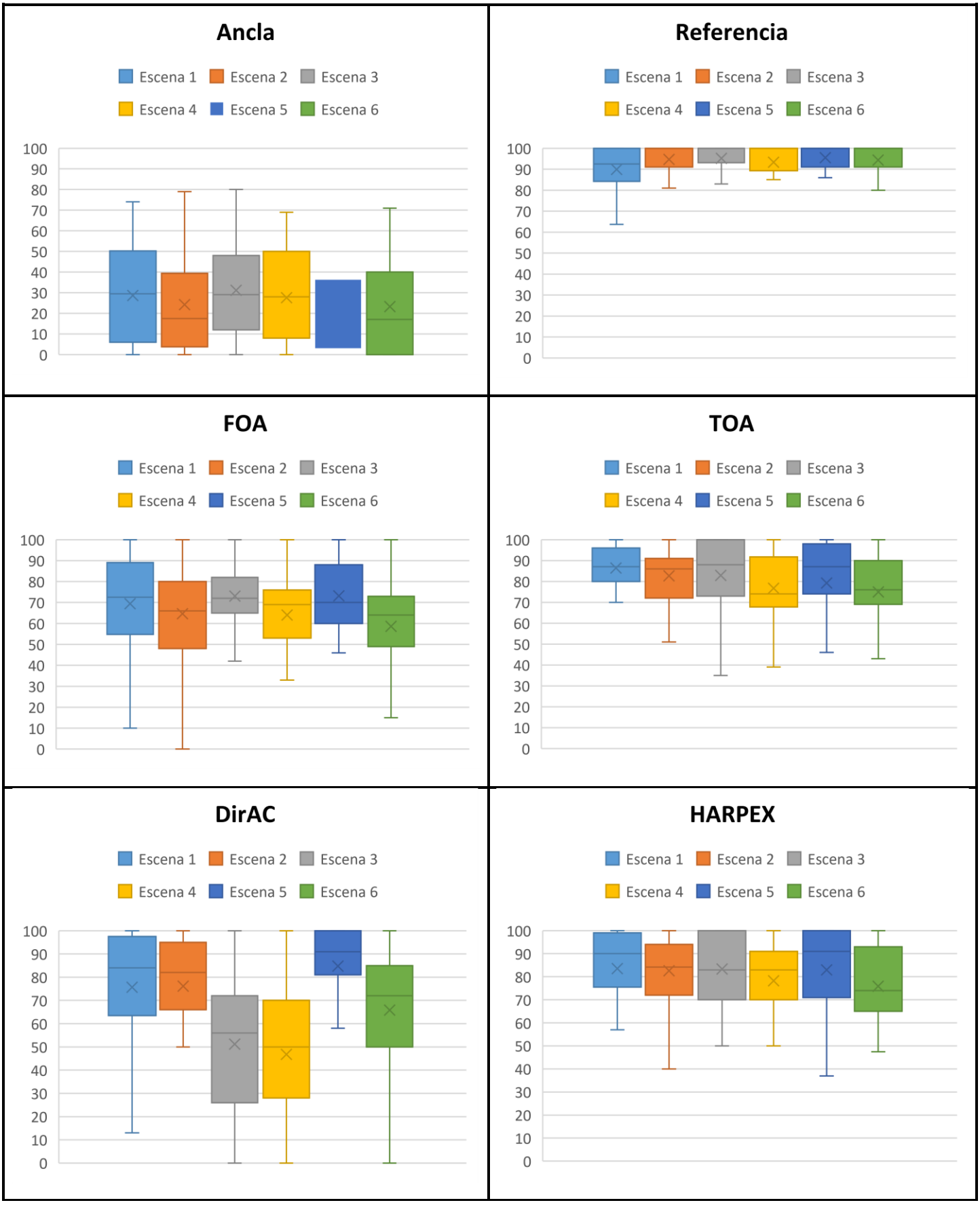
Respecto a las técnicas de upmixing, COMPASS obtiene los peores resultados, con una valoración media de 53,89. Le sigue HARPEX, con una valoración de 65,67 y DirAC es quien obtiene resultados más variables pero mejores en promedio. Este es el único estímulo con mayor puntuación en la esfera que en el test binaural, al posicionar la fuente mediante VBAP.

### 5.5.2 Análisis por escenas

Se ha realizado un análisis por escenas de los resultados para cada estímulo, para ver si la posición por la que se emiten los audios afecta a la calificación media.

Primero se analizan estos resultados en el test binaural.

Mediante una tabla comparativa (*tabla 8*), se observan las diferencias por estímulo entre cada escena.


Ancla
Escena 1
Escena 2
Escena 3
Escena 4
Escena 5
Escena 6
100
90
80
70
60
50
40
30
20
10
0
Referencia
Escena 1
Escena 2
Escena 3
Escena 4
Escena 5
Escena 6
100
90
80
70
60
50
40
30
20
10
0
FOA
Escena 1
Escena 2
Escena 3
Escena 4
Escena 5
Escena 6
100
90
80
70
60
50
40
30
20
10
0
TOA
Escena 1
Escena 2
Escena 3
Escena 4
Escena 5
Escena 6
100
90
80
70
60
50
40
30
20
10
0
DirAC
Escena 1
Escena 2
Escena 3
Escena 4
Escena 5
Escena 6
100
90
80
70
60
50
40
30
20
10
0
HARPEX
Escena 1
Escena 2
Escena 3
Escena 4
Escena 5
Escena 6
100
90
80
70
60
50
40
30
20
10
0

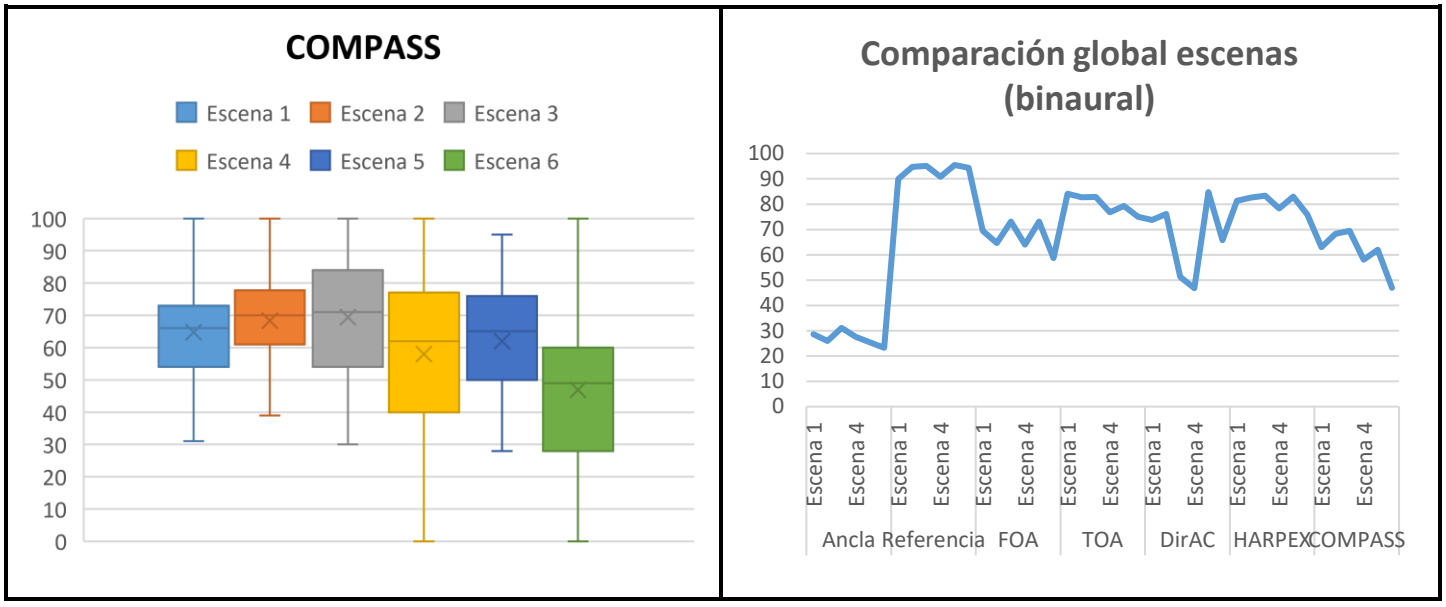



*Tabla 8 – Tabla comparativa de las diferencias en las escenas por estímulos (test binaural)*

En la *tabla 8* se observa que en el test binaural el estímulo más cambiante es el ancla, por lo que los participantes lo perciben en cada escena de manera muy distinta. Destaca la alteración en el diagrama de cajas de la escena 5 con el cambio de los audios a 4 voces (demode), donde el 75% de los valores son menores a 35 tal y como muestra el tercer cuartil del diagrama.

En el caso de la referencia, en todas las escenas excepto en la primera el 75% de los valores son mayores de 90 tal y como era de esperar. Además, más del 25% de los valores son 100 ya que el inicio de todas las cajas se sitúa en el máximo.

Los valores en TOA son los menos cambiantes después de HARPEX, con una puntuación media de 80,11. A excepción de la cuarta escena, el tercer cuartil en el resto es mayor de 60, por lo que el 75% de los valores de la distribución son mayores a 60. En FOA la puntuación media por escenas obtenida es de 67,16 con una variación máxima entre escenas del 19,81%.

Teniendo en cuenta que el orden de TOA es mayor al de FOA y que a mayor orden, mayor cantidad de armónicos, y mayor es la resolución espacial, los resultados obtenidos son los esperados, superando TOA a FOA en un 16,17%.

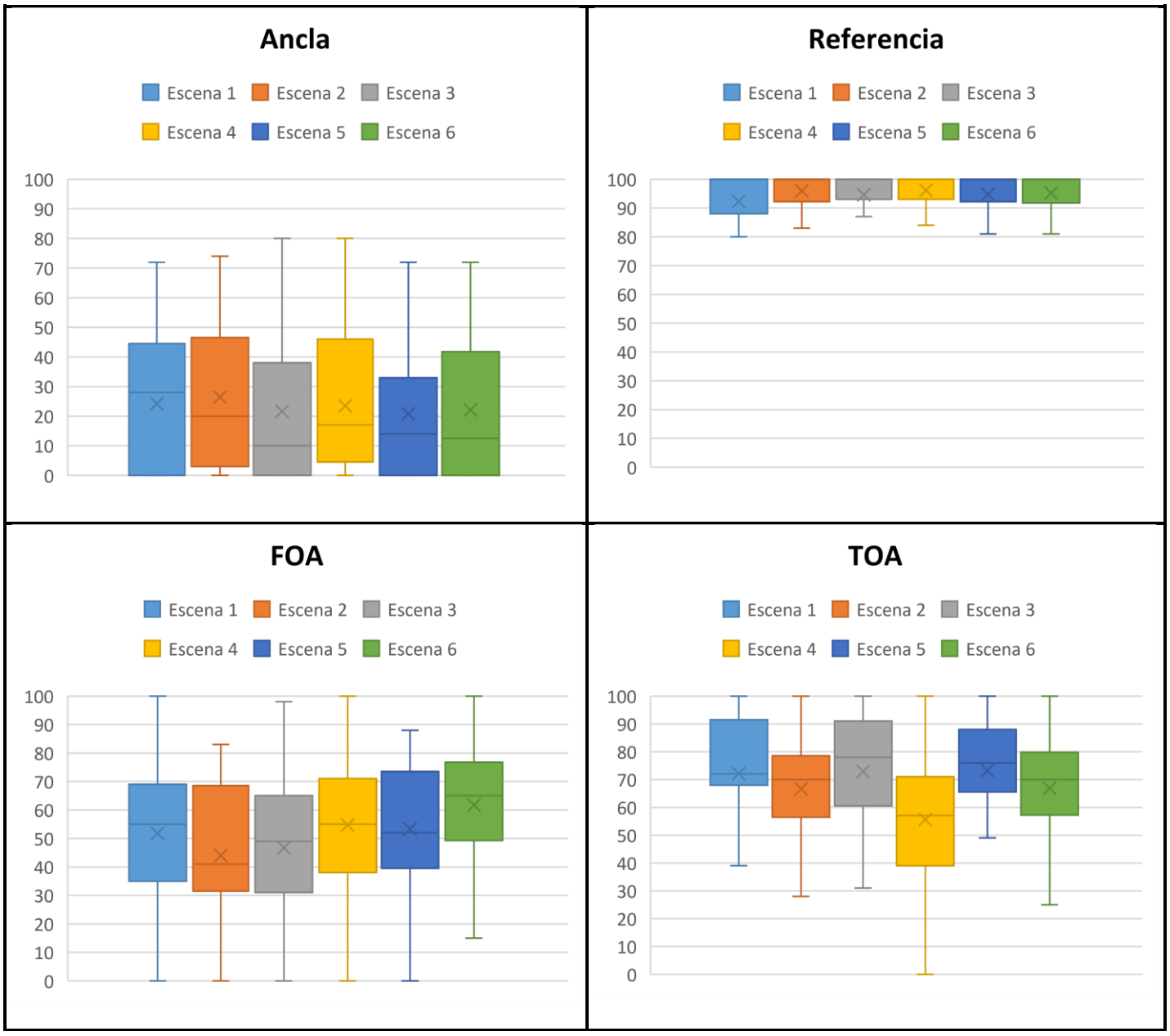
Respecto a las técnicas de upmixing, DirAC es el caso con valores más cambiantes, siendo las escenas 3 y 4 las que presentan valores más pequeños. Entre la escena 4 y la 5 se aprecia una variación media en los valores del 44,8% con el cambio de audios de fever (escena 4) a demode (escena 5). En esta última escena es donde los participantes califican de excelente los audios, con una puntuación media de 85,75 y sin apreciar prácticamente diferencias entre la referencia y el estímulo. Esto también puede observarse en la última gráfica de la tabla 8, donde se aprecia un abrupto cambio en los valores obtenidos en DirAC.

HARPEX, sin embargo, contempla valores poco cambiantes en las 6 escenas con una variación máxima del 8,96%. Tal y como indican los diagramas de cajas y bigotes, a excepción de la escena

6, en el resto el 75% de los valores son superiores a 70, por lo que se consiguen unos resultados medios excelentes.

Por último, COMPASS presenta resultados similares en todas las escenas con una variación máxima del 32,49% entre la tercera y la última. La calificación media de los audios en COMPASS por escenas es de 61,26, por lo que los participantes perciben los audios procesados mediante COMPASS como “bueno, pero ligero cambio en la posición original de las voces”.

De la misma manera, para el test mediante altavoces, la *tabla 9* muestra las diferencias obtenidas en cada escena por estímulos.



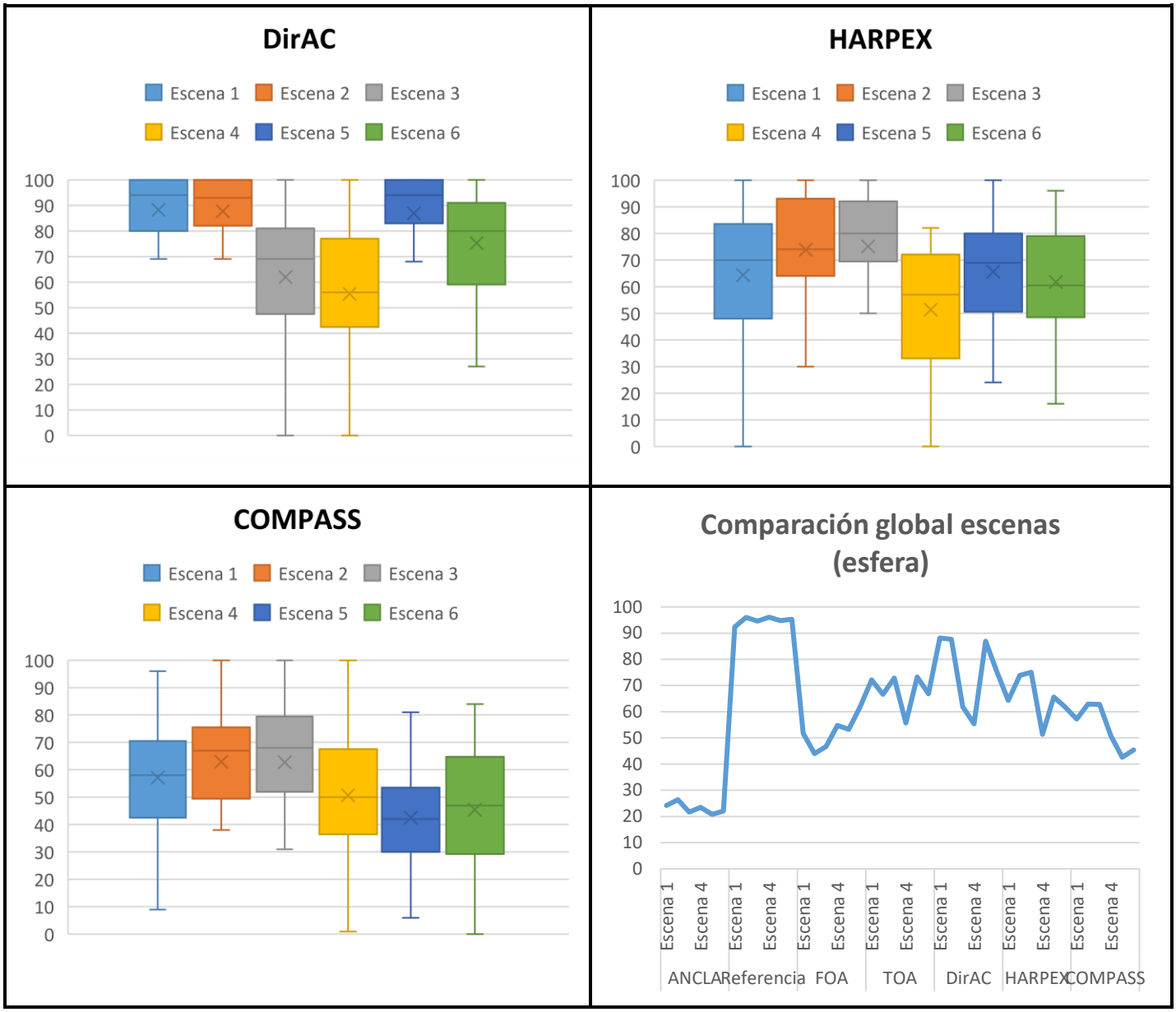



*Tabla 9 – Tabla comparativa de las diferencias en las escenas por estímulos (test esfera)*

En la *tabla 9,* se observa que en el test en la esfera el estímulo con menores puntuaciones es el ancla. Al igual que en el test en binaural, destaca el diagrama de cajas de la escena 5 con una puntuación promedio de 20,78, por debajo de la media del resto de escenas. Bajo criterio de cada participante, en general los valores del ancla son cambiantes, no superando bajo ningún caso más del 75% de los valores una puntuación de 50.

En el caso de la referencia, en todas las escenas excepto en la primera el 75% de los valores son mayores de 90 tal y como muestra el final de las cajas. Además, más del 25% de los valores son 100 ya que el inicio de todas las cajas se sitúa en el máximo.

Respecto a las técnicas de upmixing, todas tienen valores muy cambiantes para cada escena. DirAC obtiene los mejores resultados globales, con una puntuación media de 75,87. En esta técnica, se aprecia un cambio sustancial entre las primeras escenas y la tercera y cuarta, variando en un máximo del 37,15% entre escena 1 y la 4. Se aprecia un incremento considerable de los valores medios de la escena 4 del 56,82% con el cambio de audios de fever (escena 4) a demode (escena 5). En esta última escena es donde los participantes califican de excelente los audios, con una puntuación media de 86,89 y sin apreciar prácticamente diferencias entre las voces de la referencia y las del estímulo procesado. Se intuye que esto puede ser debido a las posiciones de las fuentes en cada escena. Las escenas 1, 2 y 5, las voces están "agrupadas", mientras que la 3, 4 y 6, que obtienen puntuaciones medias más bajas, están "separadas". Esto también puede observarse en la última gráfica de la *tabla 9*, donde se aprecia un abrupto cambio en los valores obtenidos en DirAC entre cada escena.

HARPEX le sigue con una puntuación media de 65,32, donde se aprecia un importante decremento del 46,41% entre las escenas 3 y 4. A excepción de esa variación, el resto de las escenas presentan resultados poco cambiantes. Destaca la tercera escena por ser la más constante y la que mayor puntuación obtiene con una valoración media de 75,11, y cuyos valores son superiores de 70 en más del 75% de los casos de la distribución.

De las 3 técnicas de upmixing, COMPASS es el que peores resultados presenta, con una puntuación media de 53,60 para todas las escenas. Los mejores resultados se encuentran en las escenas 2 y 3 con los audios de fever, y con una valoración media de 62,92 y 62,76 respectivamente. En la cuarta escena, se observa un decremento del 19,2% respecto a la escena 3. Los menores valores se encuentran en la escena 5 con una puntuación de 42,57 y una disminución en los valores de la escena 5 del 16,04% con respecto a la escena anterior, donde los participantes perciben que las voces se desvían de su posición original y/o se ensanchan ligeramente.

En general en todos los estímulos se produce un cambio en los valores medios entre las escenas 3 y 4 con un cambio en la posición de las voces emitidas por los altavoces 6 y 17 (ver *tabla 6*) a los altavoces 9 y 22.

Siguiendo con la dinámica del análisis por escena, se ha realizado una comparativa entre los audios FOA, TOA y los audios procesados para pasar de FOA a TOA (upmixers), con el objetivo de ver si realmente estos mejoran la sensación subjetiva de los oyentes.

Para ambos test, las gráficas resultantes se encuentran en la *tabla 10*.

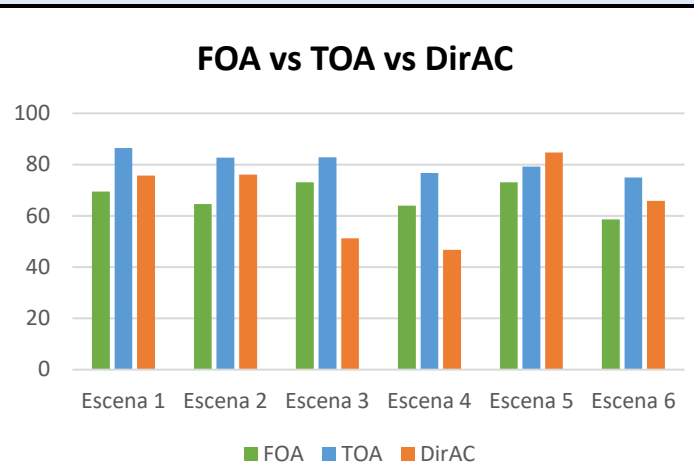
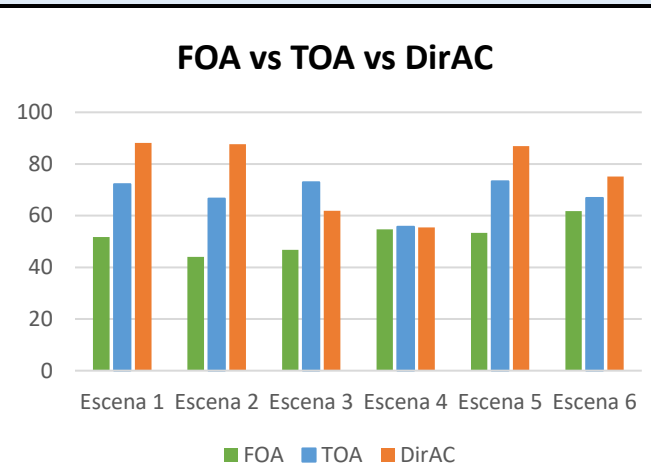
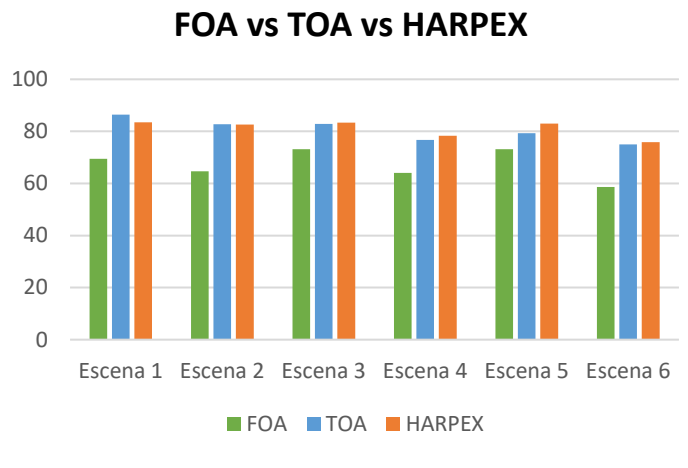
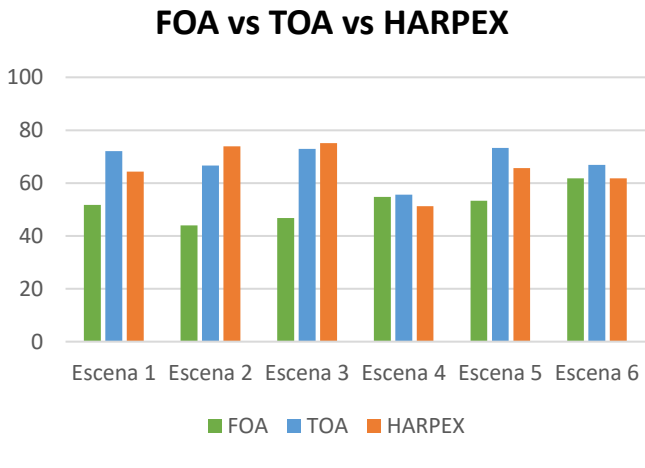
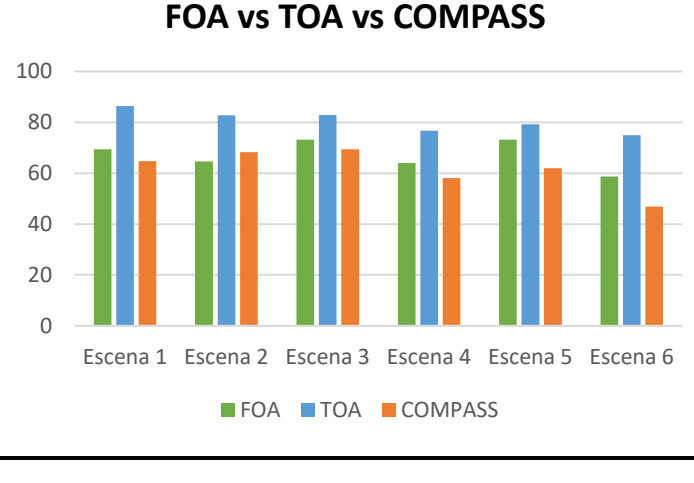
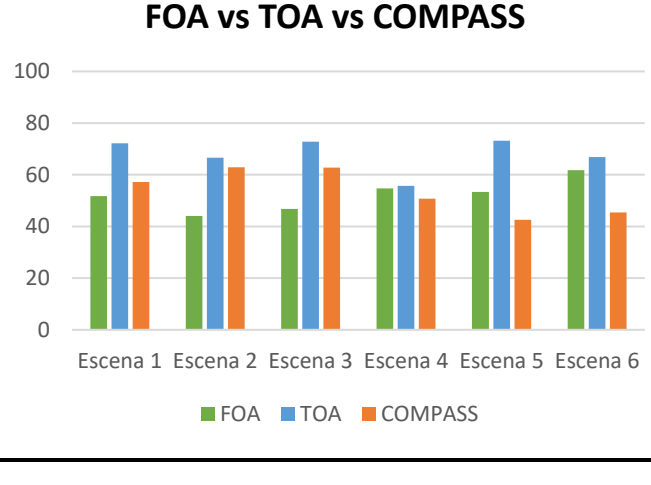
| BINAURAL | ESFERA |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Tabla 10 - Comparación FOA y TOA con las técnicas de upmixing en el test en la esfera*

En el test binaural, en todas las escenas TOA consigue mejores puntuaciones que FOA ya que son audios con un orden mayor, y por lo tanto se consigue más resolución. Los audios en TOA procesados mediante DirAC por lo general alcanzan mejores resultados que FOA excepto en las escenas 3 y 4, donde las voces se encuentran más separadas. No obstante, los resultados comparados con TOA son peores.

HARPEX obtiene resultados muy superiores a FOA (hasta un 22,73% superiores) pero prácticamente iguales a TOA, variando como máximo en un 4,6% en la escena 5.

COMPASS presenta peores resultados tanto frente a FOA como a TOA, siendo la técnica de upmixing peor valorada de manera global.

En el test mediante altavoces, DirAC ofrece resultados muy superiores a FOA sobre todo en las escenas 1, 2 y 5 donde las voces de los audios están más juntas. También, aunque con menor diferencia, supera a TOA en cuanto a puntuación en todas las escenas excepto en la tercera. Esto concluye que es técnica de upmixing más valorada por los participantes.

HARPEX obtiene mejores resultados que FOA en todas las escenas, pero ligeramente peores que TOA. Lo mismo sucede con COMPASS, aunque con una mayor diferencia con FOA y TOA.

Puede concluirse que las técnicas para pasar de Ambisonics de primer a tercer orden que mejor funcionan mediante altavoces son DirAC y HARPEX, las cuales ofrecen mejores resultados que el propio audio codificado en TOA. COMPASS, sin embargo, ofrece valores más pequeños en ambos test, siendo la técnica de upmixing peor valorada por los oyentes.

### 5.5.3 Análisis por participantes

Para ambos test, se ha realizado un análisis por estímulos para observar si el grado de relación de los participantes con la música afecta realmente a la manera de percibir los diferentes estímulos. Como se indica en el apartado 5.2, los participantes tienen que situarse en un grupo dependiendo de su relación con la música: profesional, aficionado o lego. Este análisis se lleva a cabo ya que la cantidad de participantes en cada grupo es la misma.

Agrupando los resultados para el test binaural, los resultados han sido los siguientes (*figura 50*):

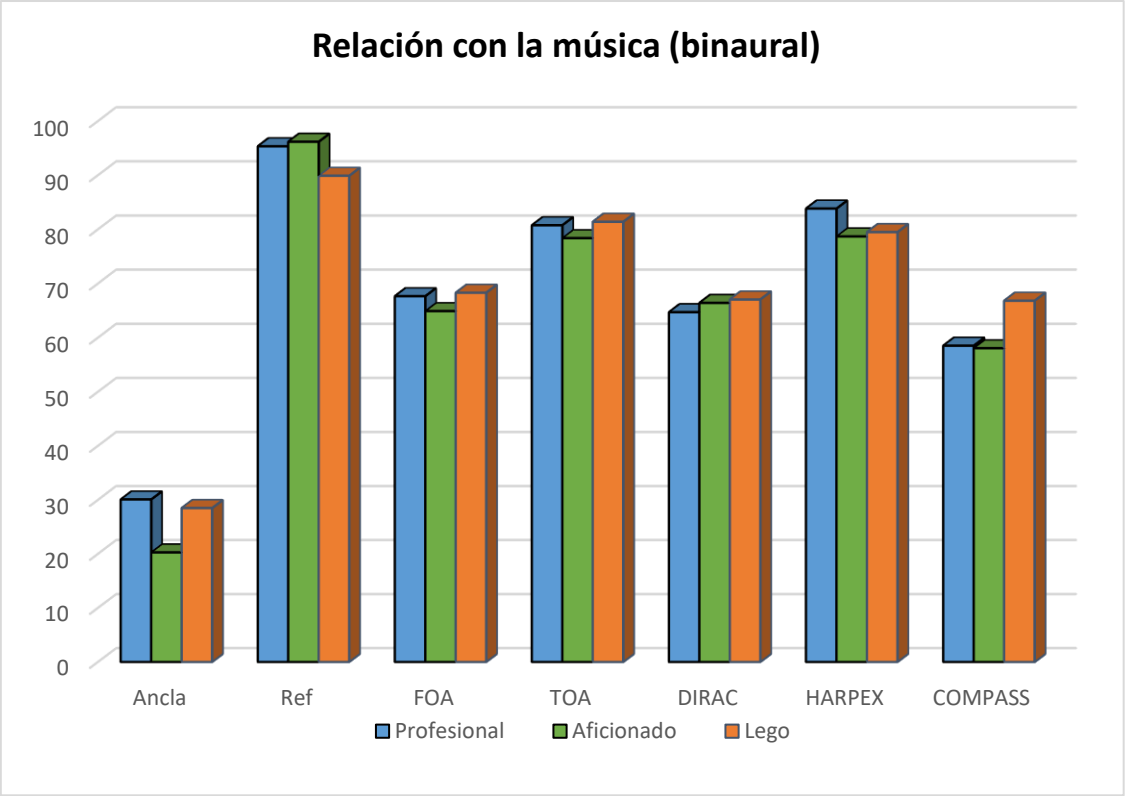



*Figura 50 – Comparación de la valoración de los estímulos por tipo de participante (binaural)*

En el test binaural, no se obtienen grandes diferencias de un grupo respecto a otro. El caso más cambiante es el ancla, la cual varía un máximo del 32,61% entre profesionales y aficionados, siendo estos últimos quienes la perciben más diferente a la referencia. Sin embargo, entre profesionales y lego la percepción de similitud del ancla con respecto a la referencia varía en menor medida, en un 5,3%.

Después del ancla, los casos más notorios se producen en los estímulos procesados mediante COMPASS, siendo un 15% mayor el valor en el caso de los participantes lego. Esto concluye que dichos participantes perciben menos diferencia en los estímulos en COMPASS con respecto a la referencia que el resto de los oyentes.

En los resultados destaca TOA, que es percibido por los tres grupos de forma muy similar con una puntuación media de 80,18 y con una máxima variación en los valores del 3,73% entre un aficionados y lego. Entre profesionales y lego la diferencia es mínima, siendo el porcentaje de variación entre ambos casos inferior al 1%.

DirAC también obtiene resultados muy parecidos en los tres grupos, siendo el menos cambiante de todos con una puntuación media de 66,05 y una variación máxima de los valores del 3,49%.

Con esto se puede concluir que en este caso la relación de los participantes con la música no afecta de manera significativa en los resultados del test en cuanto a la escucha mediante auriculares.

Sin embargo, sorprenden los resultados en el test de la esfera (*figura 51*), los cuales son un 43% más variables que en el test binaural y por lo tanto no contemplan valores tan uniformes como en la prueba mediante auriculares (*figura 50*).

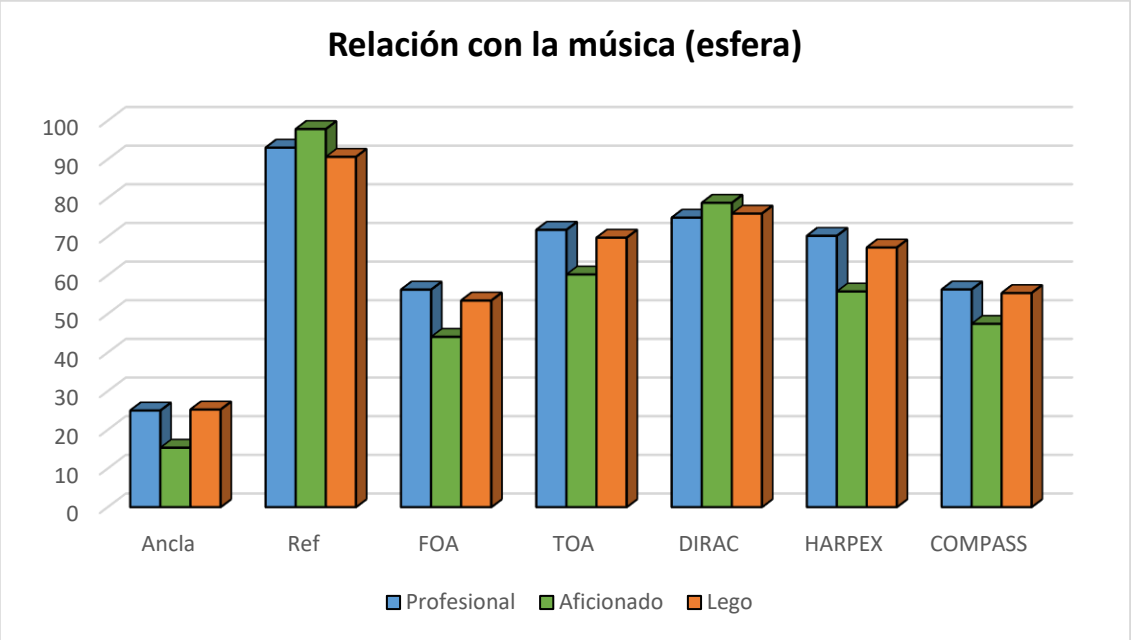



*Figura 51 – Comparación de la valoración de los estímulos por tipo de participante (esfera)*

En el test mediante altavoces (*figura 51*), el ancla es el estímulo que presenta más diferencias entre cada tipo de participante, variando en un 39% entre aficionados y legos.

Los audios en FOA son los siguientes con diferencias más evidentes, con una variación máxima de la percepción de los audios emitidos en FOA de 21,70%. TOA, sin embargo, obtiene una buena puntuación general, con un promedio de 67,22 y una variación máxima entre percepciones por participantes del 16,10%.

Respecto a las técnicas de upmixing, destaca la uniformidad de los valores obtenida en DirAC en los tres grupos, con una variación máxima entre profesionales y aficionados del 4,88% y una mínima del 1,41% entre profesionales y lego. Además, es la técnica que obtiene mayores valores en los tres grupos con una media de 76,52.

COMPASS, no obstante, presenta una variación máxima entre profesionales y aficionados del 15,78% y sin embargo una mínima del 1,63% entre profesionales y lego.

Del mismo modo que COMPASS, en HARPEX se producen grandes diferencias en cuanto a los estímulos percibidos por los aficionados y el resto de participantes. De las tres técnicas, los resultados en HARPEX son los más dispares, con una máxima variación del 20,53% entre profesionales y aficionados, frente a una variación mínima del 4,3% entre profesionales y lego.

Como se puede observar en la *figura 51*, en todos los estímulos los participantes aficionados son quienes obtienen el mayor porcentaje de variación respecto a las valoraciones medias de los

profesionales y los lego. Los aficionados presentan resultados que varían en una media del 18% respecto al resto de los grupos, quienes obtienen resultados muy similares (en torno al 2,67% de variación entre ambos).

Se ha creído conveniente hacer una comparación entre ambos test dependiendo de la relación de cada participante con la música.

La *figura 52* refleja las valoraciones obtenidas en ambos test por cada tipo de estímulo, para los participantes que mantienen una relación profesional con la música.

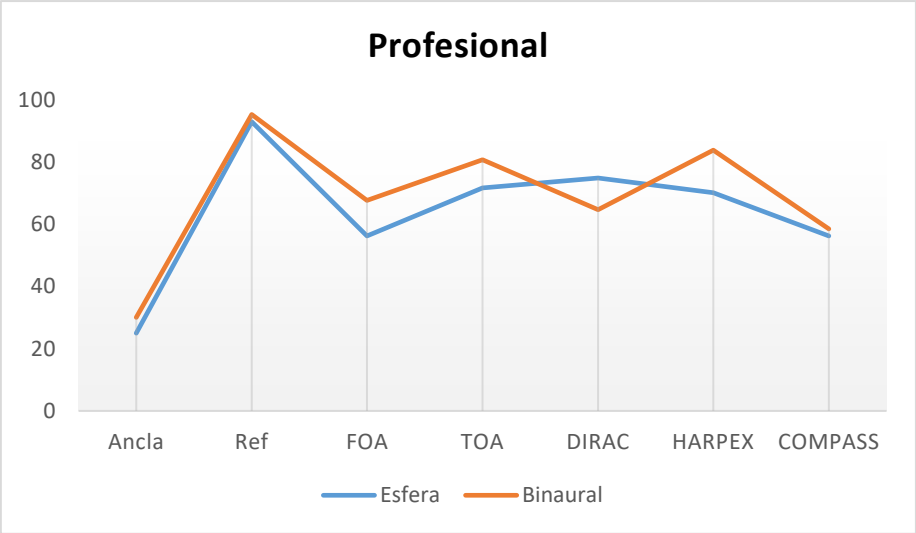



*Figura 52 – Valoración global en los estímulos de ambos test por participantes profesionales*

La *figura 53* contiene las valoraciones obtenidas de los aficionados en ambos test por cada tipo de estímulo.

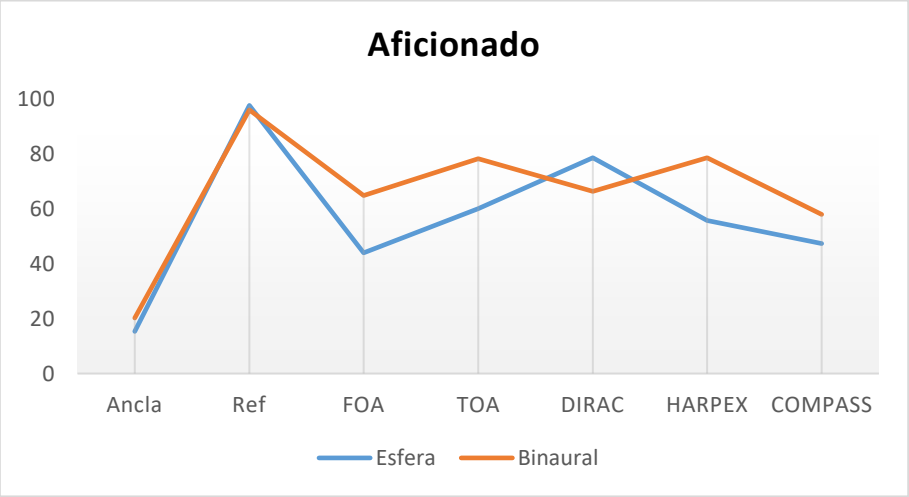



*Figura 53 - Valoración global en los estímulos de ambos* test *por participantes aficionados*

La *figura 54* muestra las valoraciones obtenidas en ambos test por cada tipo de estímulo, para los participantes no han recibido ninguna formación musical (lego).

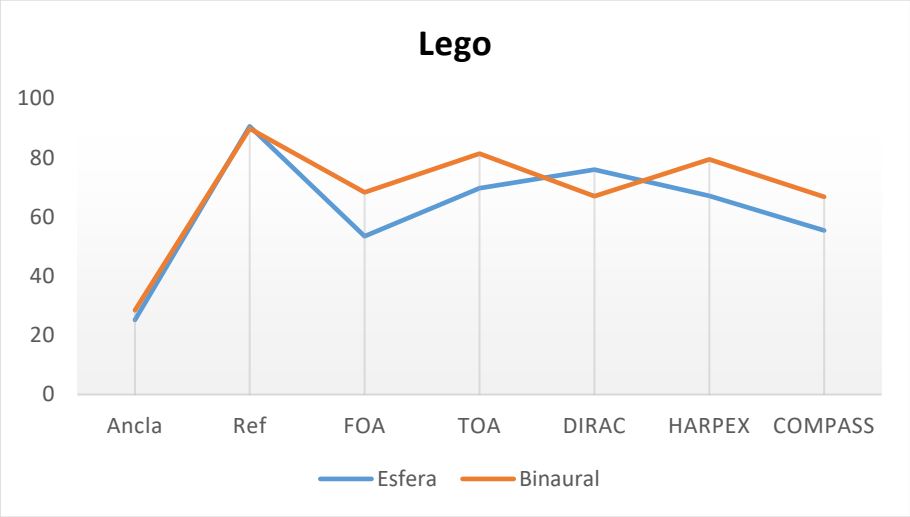



*Figura 54 - Valoración global en los estímulos de ambos* test *por participantes sin formación*

Los resultados en la gráfica de los participantes profesionales (*figura 52*) siguen una distribución más uniforme que en el resto. En el caso de los participantes aficionados y lego, ambas rectas tienen diferencias más abruptas a excepción del ancla y la referencia. Tanto en FOA y TOA como en las técnicas de upmixing, las sensaciones subjetivas en ambos test varían más que en el caso de los profesionales.

En todos los casos las diferencias percibidas han sido mayores para el test en la esfera, la cual obtiene menor puntuación en todos los estímulos excepto en DirAC.

A diferencia del resto de estímulos, los tres grupos de participantes coinciden en que perciben una mayor diferencia en los audios procesados mediante DirAC a través de auriculares que mediante altavoces, siendo el único caso en el que los valores obtenidos en la esfera superan a los valores obtenidos en el test binaural. A su vez es el estímulo que se percibe de manera más similar en los tres grupos, variando un máximo de 3,49% en el test binaural y un 4,88% en el test esfera. Los valores medios de DirAC en el test de la esfera decrementan en un 13,6% en el caso de participantes profesionales, 15,65% en el de aficionados y un 11,72% en el de los oyentes sin formación musical.

La mayor diferencia se produce en el caso de FOA y HARPEX en los aficionados. En FOA, la diferencia percibida en el test binaural es mayor, con un incremento en los valores del test mediante altavoces del 50,83%. En HARPEX se produce una variación creciente del 45,56% de las diferencias percibidas en el test binaural con respecto al test en la esfera. Con esto, se puede concluir que los participantes aficionados perciben la mayor diferencia en los estímulos emitidos en Ambisonics de primer orden (FOA) mediante el test en la esfera que a través del test binaural.

### 5.5.4 Análisis por participantes con problemas auditivos

Al inicio de las pruebas, los participantes deben indicar si alguna vez les han detectado problemas de audición o no. Respecto a esos participantes y al resto, se han realizado unas gráficas comparativas de ambos test (*figuras 55 y 56*).

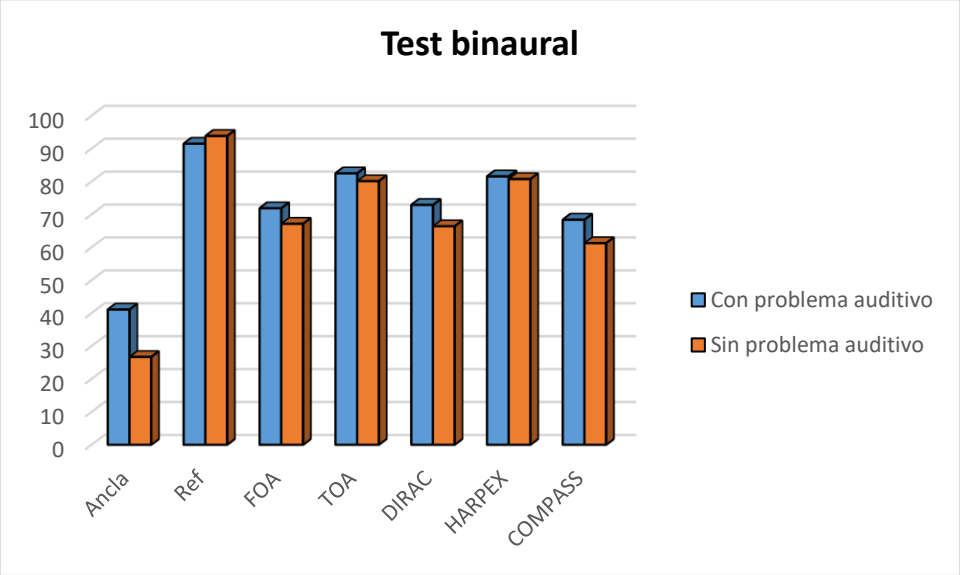



*Figura 55 - Comparativa entre los resultados de los participantes con problemas auditivos y los participantes sin problemas auditivos en el test binaural.*

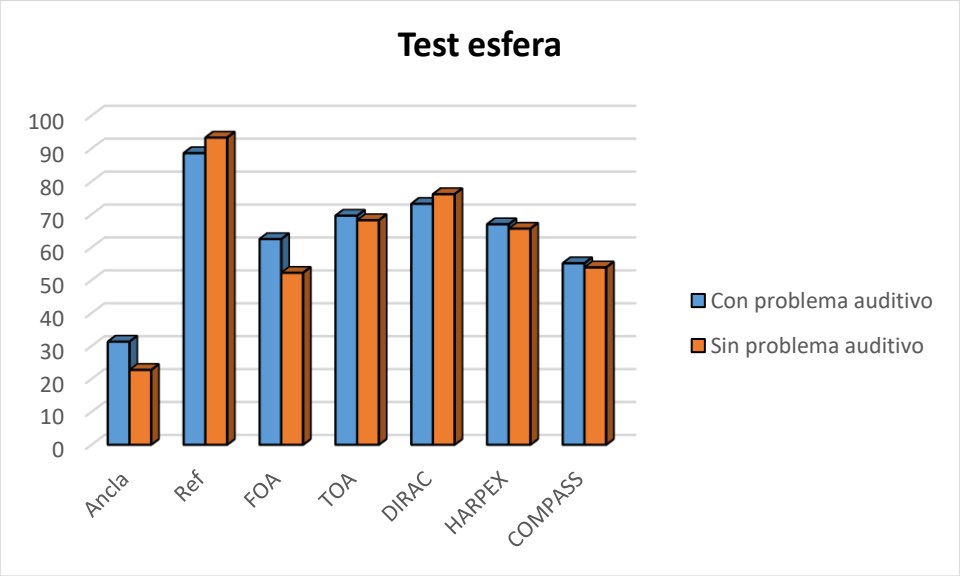



*Figura 56 - Comparativa entre los resultados de los participantes con problemas auditivos y los participantes sin problemas auditivos en el test en la esfera.*

No se aprecian cambios significativos en ninguno de los dos casos, por lo que en este experimento el hecho de haber padecido problemas auditivos no ha tenido un impacto relevante, aunque la muestra no resulta representativa al haber sido una minoría de los casos (3 personas). No se aprecian diferencias con respecto al resto de población al no ser una cantidad significativa.

### 5.5.5 Análisis por tipo de test

En cuanto al análisis por tipo de test, en la *figura 57* se obtiene una comparación del resultado en todas las escenas de ambos test por estímulos.

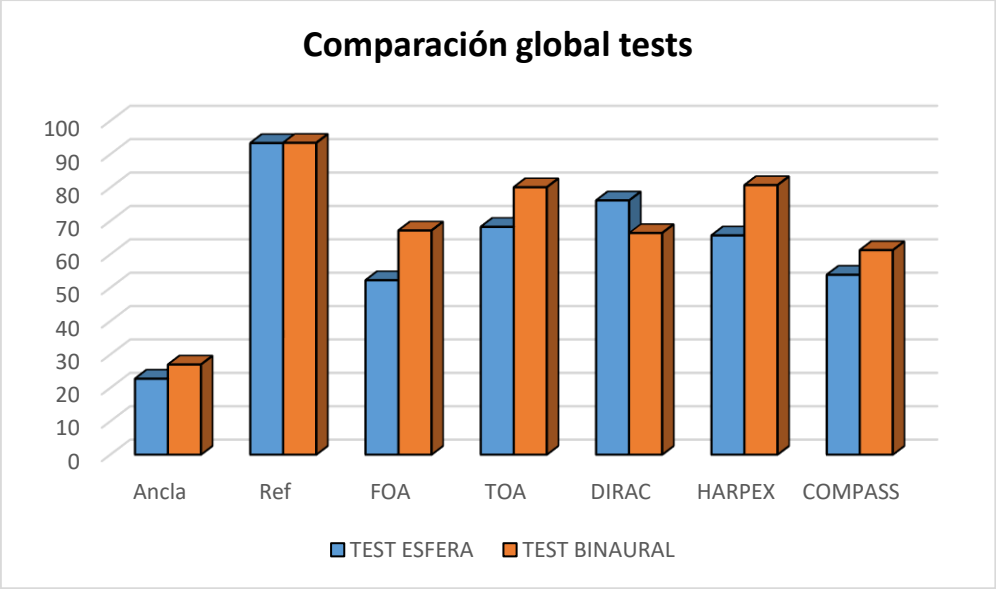



*Figura 57 - Comparación global por estímulos de los dos* test

Los resultados en todos los casos, excepto en DirAC, son mayores en el test realizado a través de auriculares que mediante altavoces. Esto confirma las impresiones reflejadas de manera generalizada por los oyentes en cuanto a percibir más diferencias en la esfera que en el test binaural, ya que el sonido es absolutamente tridimensional y la localización de las voces dentro de la esfera es más evidente que mediante auriculares.

En todos los estímulos los participantes perciben en el test binaural mayor deslocalización de las fuentes con respecto a la referencia que en el test en la esfera.

Como era de esperar, la referencia obtiene los mayores valores. Destaca su caso, el cual permanece prácticamente inalterable en ambos test, con una variación global del 0,44% y una puntuación media de 93,19.

DirAC es el único caso en el que los participantes perciben una mayor diferencia en los estímulos del test por auriculares que a través de altavoces, con un decremento del 12,82% en el segundo.

En FOA y HARPEX la diferencia global percibida es mayor. Para FOA, los resultados globales en el test esfera son de 52,26 y en el test binaural el resultado incrementa un 28,5%, con una valoración media de 67,16. En los estímulos procesados mediante HARPEX, la puntuación global en el test mediante altavoces es de 65,67 y en el test binaural la puntuación aumenta un 22,95%, obteniendo una calificación media de 80,74.

Por último, en las *figuras 58 y 59* se presentan los resultados con mejores puntuaciones en orden creciente en el test binaural y en el test en la esfera respectivamente.

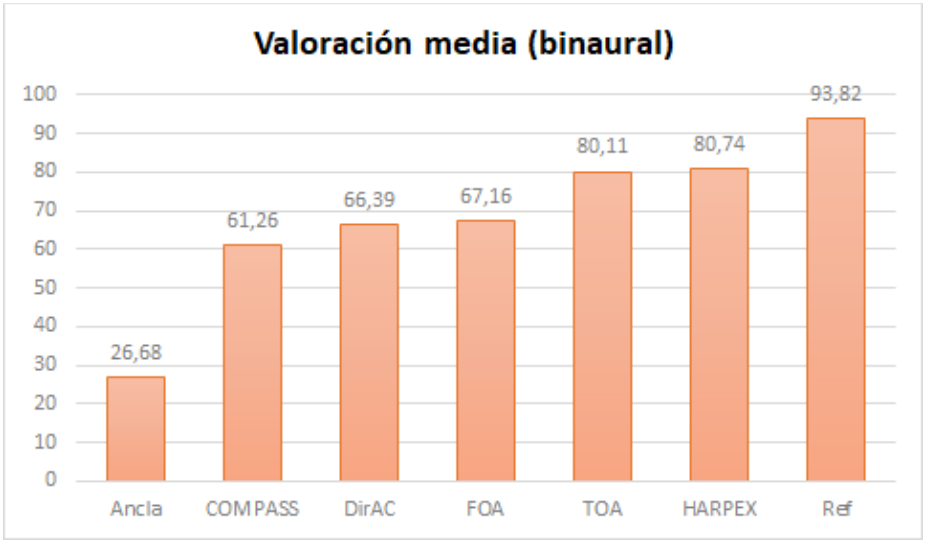



*Figura 58 - Valoraciones por orden de puntuación en el test binaural*

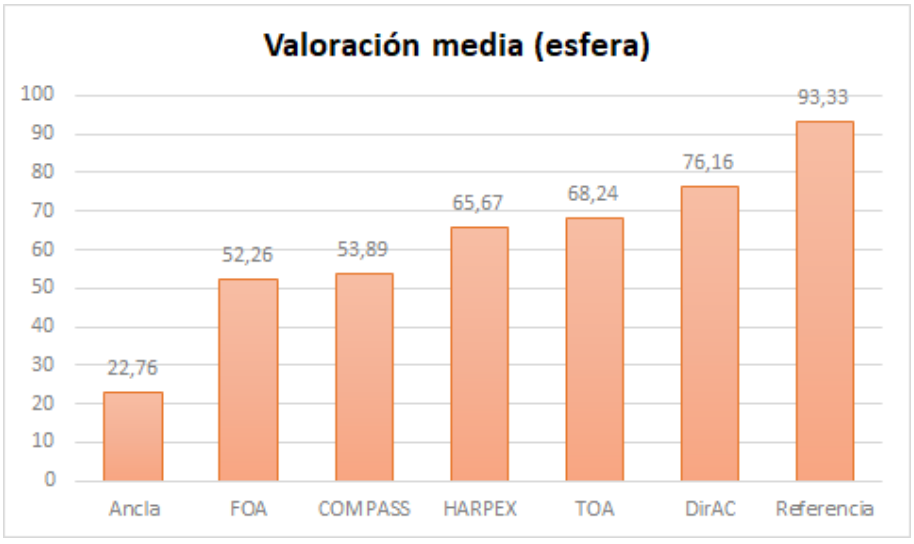



*Figura 59 - Valoraciones por orden de puntuación en el test esfera*

Se observa que en test binaural los mejores resultados, después de la referencia, los proporcionan HARPEX y TOA con puntuaciones de 80,74 y 80,11 respectivamente. Para conseguir un sonido envolvente mediante auriculares, los participantes califican como excelente la técnica de upmixing de HARPEX para convertir audios Ambisonics de primer orden en Ambisonics de tercer orden, con la cual apenas aprecian diferencias con respecto a la referencia.

Como era de esperar, TOA también arroja buenos resultados, ya que son los audios convertidos directamente a Ambisonics de tercer orden sin un paso previo por técnicas de procesado de audio.

DirAC y COMPASS son las técnicas de upmixing que presentan los peores resultados, con una puntuación de 66,39 y 61,26 respectivamente. No obstante, se mueven en el rango de 60 a 80, por lo que los participantes califican los audios con respecto a la referencia como “bueno, con un ligero cambio en la posición original de las voces“.

Inesperadamente, aunque no por mucha diferencia, las voces emitidas en Ambisonics de primer orden (FOA) proporcionan mejores resultados que las voces emitidas en Ambisonics de tercer orden mediante el procesado con DirAC y COMPASS, con una calificación media de 67,16.

Tal y como era de esperar el ancla obtiene los resultados más bajos, con una media de 22,76 y por lo tanto los participantes califican las voces del ancla respecto a las voces de la referencia como “pobre, desvío sustancial, ensanchamiento claro y/o dificultad para localizarlas”.

En el test a través de altavoces los mejores resultados los proporciona DirAC con una valoración media de 76,16, donde los participantes perciben un ligero cambio en la posición de las voces procesadas mediante dicha técnica respecto a las voces de referencia.

A diferencia del caso anterior, en este test las voces en TOA procesadas mediante DirAC y COMPASS logran mejores resultados que las voces emitidas en Ambisonics primer orden (FOA). Estas obtienen la peor valoración con un 52,26, donde los oyentes perciben que las voces se desvían de su posición original y/o se ensanchan ligeramente. HARPEX continúa arrojando resultados notables con una puntuación de 65,67, los cuales superan en un 17,93% a los audios procesados mediante COMPASS pero son inferiores en un 15,97% a las calificaciones medias en DirAC.

El ancla obtiene peores resultados que en el test binaural con una puntuación de 22,76 al ser un audio emitido por los 24 altavoces, por lo que las voces sufren una clara deslocalización y los participantes perciben un desvío sustancial, ensanchamiento claro y/o dificultad para localizarlas respecto a las voces de referencia.

# 6. Conclusiones

- Se han realizado dos test MUSHRA de 7 estímulos y 6 escenas a 39 participantes para valorar tres técnicas de upmixing (DirAC, HARPEX COMPASS), a través de auriculares y de altavoces.

- Los estímulos empleados consisten en audios de dos y cuatro voces en FOA y TOA, y los mismos audios procesados de FOA a TOA mediante las tres técnicas de upmixing. Los dos estímulos restantes son el ancla y la propia referencia.

- Se han agrupado a los participantes conforme a su relación con la música en profesionales, aficionados y lego (sin formación).

- Se han hecho distinción entre los participantes que han padecido problemas de audición y los que no.

- Para el test binaural se han obtenido resultados excelentes en los audios procesados mediante HARPEX con un promedio de 80,74.

- Para el test esfera se han obtenido buenos resultados en los audios procesados mediante DirAC, con una valoración media de 76,16, donde DirAC funciona mejor al posicionar la fuente directa mediante VBAP.

- Para ambos test se ha conseguido la segunda mejor puntuación en los audios en TOA, con una puntuación media de 68,24 para el test en la esfera y 80,11 para el test binaural.

- En los dos test, en DirAC, las escenas en las que las voces están agrupadas (escenas 1, 2 y 5) obtienen mayores puntuaciones que las escenas en las que las voces están más separadas (escenas 3, 4 y 6), lo cual tendrá que ver con cómo esta técnica procesa la señal para el upmixing. En el resto de técnicas esto también pasa, aunque no de forma tan exagerada, mientras que en TOA o REF este fenómeno no se da.

- En el análisis por grupos de participantes, en el test binaural los resultados son uniformes sin grandes cambios de un respecto a otro.

- En el análisis por grupos de participantes, en el test esfera se obtienen resultados cambiantes para cada grupo. Destacan los valores en los aficionados, los cuales en todos los estímulos excepto en DirAC y en la referencia son significativamente menores que en el resto de oyentes.

- No se han encontrado diferencias significativas entre los participantes con problemas auditivos y el resto.

- Al finalizar ambos test los participantes han reflejado sus impresiones y valoraciones y todos coinciden en que la claridad y la distinción de los estímulos con respecto a la referencia es mucho más notoria en la esfera de altavoces que mediante auriculares.

- Los oyentes coinciden en que el sonido es más nítido en la esfera de altavoces y la sensación de profundidad mayor.

# 7. Valoraciones y líneas de futuro

Para escuchar el sonido de manera tridimensional mediante auriculares, los participantes prefieren audios procesados con HARPEX o emitidos en Ambisonics de tercer orden (TOA). Sin embargo, para una escucha mediante altavoces, las técnicas que mejores resultados proporcionan son DirAC y TOA.

Tal y como era de esperar, TOA ofrece buenos resultados para ambos test con un promedio de, ya que es el estímulo en tercer orden al que se quería llegar mediante el procesado de FOA con los diferentes tipos de técnicas de upmixing.

Para el test a través de auriculares, las técnicas de upmixing DirAC y COMPASS para pasar de Ambisonics de primer orden (FOA) a Ambisonics de tercer orden (TOA) ofrecen peores resultados que los propios audios en FOA. No obstante, si el sonido se emite mediante altavoces estas técnicas consiguen resultados óptimos, siendo DirAC la técnica con mayor puntuación. DirAC funciona mejor en la esfera, al posicionar la fuente directa mediante VBAP. Al binauralizar los audios esto se pierde.

Tanto para escuchar sonido a través de auriculares como a través de altavoces, HARPEX logra buenos resultados, por lo que es la técnica que mejor se adapta a ambos tipos de formato.

Las técnicas de upmixing existentes todavía proporcionan resultados de peor calidad que mediante una grabación directa en Ambisonics de tercer orden. Ambisonics y el sonido tridimensional tienen cabida en múltiples líneas a futuro, ya que es un sistema relativamente novedoso que puede jugar un papel muy interesante en realidad virtual, entre otros.

Si se desarrolla la investigación en los procesos de upmixing y se consiguen resultados sin necesidad de grabar en TOA, el sonido inmersivo podrá convertirse en algo cada vez más habitual.

Este experimento ha logrado entidad suficiente gracias al número de participantes que ha realizado el test, pero como línea de futuro queda por analizar mediante un test ANOVA la significancia estadística de los datos.

# 7. Bibliografía


[1] Daniel Arteaga (2018) - Introduction to Ambisonics

[2] Franz Zotter - Matthias Frank (2019) - Ambisonics A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality

[3] Victor Perales (2015) - Ambisonics como alternativa de sonido inmersivo

[4] Michael Chapman, Winfried Ritsch , Thomas Musil , IOhannes Zmölnig , Hannes Pomberger , Franz Zotter , Alois Sontacchi (2009) - A STANDARD FOR INTERCHANGE OF AMBISONIC SIGNAL SETS Including a file standard with metadata

[5] Michael A Gerzon (1992) - General metatheory of auditory localisation.

[6] mh acoustics https://mhacoustics.com/products

[7] HULTI-GEN: https://zenodo.org/record/4302080#.YCpm5WhKiXI

[8] Recomendación ITU-T P.911

[9] Recomendación ITU-R BS.1116

[10] Recomendación ITU-R BS. 1534-2

[11] V. Pulkki, M-V Laitinen1, J Vilkamo, J Ahonen, T Lokki and T Pihlajamaki (2009) - Directional audio coding - perception-based reproduction of spatial sound

[12] Aalto University - Coding and Multidirectional Parameterisation of Ambisonic Sound Scenes (COMPASS) http://research.spa.aalto.fi/projects/compass_vsts/plugins.html

[13] G. D. Galdo, V. Pulkki, F. Kuech, M.-V. Laitinen, R. Schultz-Amling, and M. Kallinger, 2009) - Efficient methods for high quality merging of spatial audio streams in directional audio coding.

[14] DirAC https://www.dirac.com

[15] Mikko-Ville Laitinen (2008) - Binaural Reproduction for Directional Audio Coding

[16] Archontis Politis, Leo McCormack, Ville Pulkki (2017) - Enhancement Of Ambisonic Binaural Reproduction Using Directional Audio Coding With Optimal Adaptive Mixing

[17] P. Felgett (1974) - Ambisonic reproduction of directionality in surround-sound systems. Nature 252,534–538

[18] M.A. Gerzon (1975) - The design of precisely coincident microphone arrays for stereo and surround sound, prepr. L-20 of 50th Audio Eng. Soc. Conv.

[19] P. Craven, M.A. Gerzon (1977) - Coincident microphone simulation covering three dimensional space and yielding various directional outputs, U.S. Patent, no. 4,042,779

[20] D.H. Cooper, T. Shiga (1972) - Discrete-matrix multichannel stereo. J. Audio Eng. Soc. 20(5), 346–360

[21] Svein Berge Berges Allmennd (2010) - HIGH ANGULAR RESOLUTION PLANEWAVE EXPANSION

[22] Politis, A., Tervo S., and Pulkki, V. (2018) COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[23] Stéphanie Bertet, Jérôme Daniel, Etienne Parizet, Olivier Warusfel (2009) - INFLUENCE OF MICROPHONE AND LOUDSPEAKER SETUP ON PERCEIVED HIGHER ORDER AMBISONICS REPRODUCED SOUND FIELD

[24] Harpex https://harpex.net/