ITUPublicaciones
Recomendaciones

Unión Internacional de Telecomunicaciones
Sector de Radiocomunicaciones

# Recomendación UIT-R BS.1387-2

**(05/2023)**

Serie BS: Servicio de radiodifusión (sonora)

# Método para mediciones objetivas de la calidad de audio percibida

## Prólogo

El Sector de Radiocomunicaciones tiene como cometido garantizar la utilización racional, equitativa, eficaz y económica del espectro de frecuencias radioeléctricas por todos los servicios de radiocomunicaciones, incluidos los servicios por satélite, y realizar, sin limitación de gamas de frecuencias, estudios que sirvan de base para la adopción de las Recomendaciones UIT-R.

Las Conferencias Mundiales y Regionales de Radiocomunicaciones y las Asambleas de Radiocomunicaciones, con la colaboración de las Comisiones de Estudio, cumplen las funciones reglamentarias y políticas del Sector de Radiocomunicaciones.

## Política sobre Derechos de Propiedad Intelectual (IPR)

La política del UIT-R sobre Derechos de Propiedad Intelectual se describe en la Política Común de Patentes UIT-T/UIT-R/ISO/CEI a la que se hace referencia en la Resolución UIT-R 1. Los formularios que deben utilizarse en la declaración sobre patentes y utilización de patentes por los titulares de las mismas figuran en la dirección web http://www.itu.int/ITU-R/go/patents/es, donde también aparecen las Directrices para la implementación de la Política Común de Patentes UIT-T/UIT-R/ISO/CEI y la base de datos sobre información de patentes del UIT-R sobre este asunto.

### Series de las Recomendaciones UIT-R

(También disponible en línea en https://www.itu.int/publ/R-REC/es)

| Series | Título |
|---|---|
| **BO** | Distribución por satélite |
| **BR** | Registro para producción, archivo y reproducción; películas en televisión |
| **BS** | **Servicio de radiodifusión (sonora)** |
| **BT** | Servicio de radiodifusión (televisión) |
| **F** | Servicio fijo |
| **M** | Servicios móviles, de radiodeterminación, de aficionados y otros servicios por satélite conexos |
| **P** | Propagación de las ondas radioeléctricas |
| **RA** | Radioastronomía |
| **RS** | Sistemas de detección a distancia |
| **S** | Servicio fijo por satélite |
| **SA** | Aplicaciones espaciales y meteorología |
| **SF** | Compartición de frecuencias y coordinación entre los sistemas del servicio fijo por satélite y del servicio fijo |
| **SM** | Gestión del espectro |
| **SNG** | Periodismo electrónico por satélite |
| **TF** | Emisiones de frecuencias patrón y señales horarias |
| **V** | Vocabulario y cuestiones afines |

***Nota**: Esta Recomendación UIT-R fue aprobada en inglés conforme al procedimiento detallado en la Resolución UIT-R 1.*

*Publicación electrónica*
Ginebra, 2023

# RECOMENDACIÓN UIT-R BS.1387-2

## Método para mediciones objetivas de la calidad de audio percibida

(1998-2001-2023)

**Cometido**

En esta Recomendación se especifica un método para la medición objetiva de la calidad de audio percibida.

**Palabras clave**

Calidad de audio percibida, medición objetiva, codificación a baja velocidad binaria

La Asamblea de Radiocomunicaciones de la UIT,

*considerando*

*a)* que los métodos objetivos convencionales (por ejemplo, los utilizados para medir la relación señal/ruido y la distorsión) no son ya adecuados para medir la calidad de audio percibida de sistemas que utilizan esquemas de codificación a baja velocidad binaria o que emplean procesamiento analógico o digital de la señal;

*b)* que se están desarrollando y empleando rápidamente esquemas de codificación a baja velocidad binaria;

*c)* que no todas las realizaciones conformes a una especificación o norma garantizan la obtención de una calidad óptima con dicha especificación o norma;

*d)* que los métodos formales de evaluación subjetiva no son adecuados para la supervisión continua de la calidad de audio, por ejemplo en condiciones operacionales;

*e)* que la evaluación objetiva de la calidad de audio percibida puede acabar complementando o sustituyendo a los métodos de prueba objetivos convencionales en todo tipo de medición;

*f)* que la evaluación objetiva de la calidad de audio percibida puede complementar en forma útil a los medios de evaluación subjetiva;

*g)* que, para algunas aplicaciones, es necesario utilizar un método que pueda aplicarse en tiempo real,

*recomienda*

**1** que para cada una de las aplicaciones indicadas en el Anexo 1 se utilice el método indicado en el Anexo 2 para la medición objetiva de la calidad de audio percibida.

## Preámbulo

Esta Recomendación especifica un método para la medición objetiva de la calidad de audio percibida de un dispositivo sometido a prueba, por ejemplo un codificador-decodificador (códec) a baja velocidad binaria. Se divide en dos Anexos. El Anexo 1 ofrece al usuario una perspectiva general del método e incluye cuatro Adjuntos. El Adjunto 1 describe aplicaciones y señales de pruebas. El Adjunto 2 indica las variables de salida del modelo y analiza las limitaciones de utilización y de exactitud. El Adjunto 3 describe el modelo, y el Adjunto 4 enuncia los principios y describe, en general, las características de los métodos objetivos de medición de la calidad de audio percibida.

El Anexo 2 proporciona al realizador una descripción detallada del método; se utilizan dos versiones del modelo sicoacústico que fueron desarrolladas en la fase de integración, en la que se combinaron seis modelos. En el Adjunto 1 al Anexo 2 se describe el proceso de validación del método de medición objetiva. El Adjunto 2 al Anexo 2 presenta una visión de conjunto de todas las bases de datos que se utilizaron en el desarrollo y la validación del método.

# ÍNDICE

*Página*

*Página*

*Página*

*Página*

*Página*

# Anexo 1

# Visión de conjunto

## 1 Introducción

La calidad de audio es uno de los factores esenciales en el diseño de un sistema digital para radiodifusión. La rápida introducción de diversos esquemas de reducción de la velocidad binaria ha conducido a importantes actividades destinadas a crear y mejorar procedimientos para evaluaciones subjetivas, simplemente porque las pruebas de escucha formales han sido el único método relevante para juzgar la calidad de audio. La experiencia adquirida sirvió de base para la formulación de la Recomendación UIT-R BS.1116, la cual a su vez, sirvió de base para la mayor parte de las pruebas de escucha de este tipo.

Dado que las evaluaciones subjetivas de la calidad toman mucho tiempo y son costosas, es conveniente desarrollar un método de medición objetiva para obtener una estimación de la calidad de audio. Los métodos tradicionales de medición objetiva, como la relación señal ruido (*S/N*) o la distorsión armónica total (THD, *total-harmonic-distorsión*) nunca han demostrado ser realmente capaces de dar resultados que estén relacionados de una manera fiable con la calidad de audio percibida. Los problemas se hacen aún más palpables cuando estos métodos se aplican en códecs modernos que son no lineales y no estacionarios.

En los últimos diez años se han introducido varios métodos para efectuar mediciones perceptuales objetivas de la calidad de audio percibida. Sin embargo, ninguno de estos métodos fue totalmente validado y, en consecuencia, ninguno de ellos fue normalizado ni ampliamente aceptado. En 1994, el UIT-R reconoció la urgente necesidad de formular una norma en este sector, y se comenzó a trabajar en ella. Se convocó a la presentación de propuestas y se recibieron respuestas relativas a los seis métodos de medición siguientes: índice de perturbación (DIX, *disturbance index*), relación ruido-enmascaramiento (NMR, *noise-to-mask ratio*), medida de calidad de audio perceptual (PAQM, *perceptual audio quality measure*), evaluación perceptual (PERCEVAL), medida objetiva perceptual (POM, *perceptual objective measure*) y el enfoque *Toolbox* (caja de herramientas). Estos métodos se describen en el Adjunto 4 al Anexo 1.

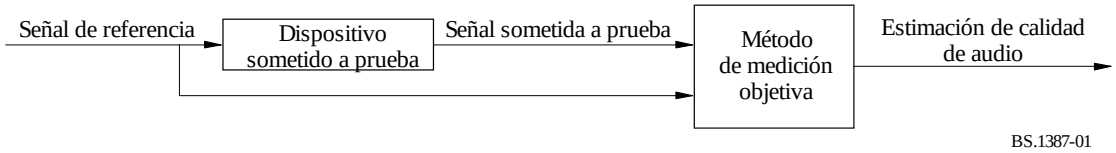
El método de medición especificado en esta Recomendación es el resultado de un proceso en el que se estudió el comportamiento de los seis métodos mencionados, y las herramientas más prometedoras de dichos métodos se integraron en un solo método. El método recomendado ha sido cuidadosamente validado en diversos lugares de prueba. Ha quedado probado que el método genera información fiable y útil para varias aplicaciones. No obstante, se debe tener presente que el método de medición objetiva de esta Recomendación no es, en general, un sustituto de la prueba de audición formal.

## 2 Aplicaciones

El concepto básico para realizar mediciones objetivas con el método recomendado se ilustra en la Fig. 1 que sigue.

FIGURA 1

**Diagrama de principio para efectuar mediciones objetivas**




El método de medición especificado en esta Recomendación es aplicable a la mayor parte de los tipos de equipo digitales y analógicos de procesamiento de señales de audio. Sin embargo, se espera que muchas aplicaciones empleen fundamentalmente códecs de audio.

Se han identificado las siguientes ocho clases de aplicaciones:

CUADRO 1

**Aplicaciones**

| | Aplicación | Breve descripción | Versión |
|---|---|---|---|
| 1 | Evaluación de realizaciones | Procedimiento que se emplea para caracterizar diferentes realizaciones de equipos de procesamiento de audio, en muchos casos códecs de audio | Básica/Avanzada |
| 2 | Ajuste de calidad perceptual | Procedimiento rápido que se emplea antes de poner en servicio una unidad de equipo o un circuito | Básica |
| 3 | Supervisión en línea | Proceso continuo para la supervisión de una transmisión de audio en curso | Básica |
| 4 | Estado del equipo o de la conexión | Análisis detallado de una unidad de equipo o de un circuito | Avanzada |
| 5 | Identificación de códec | Procedimiento para identificar el tipo y realización de un determinado códec | Avanzada |
| 6 | Desarrollo de códec | Procedimiento que caracteriza el comportamiento del códec lo más detalladamente posible | Básica/Avanzada |
| 7 | Planificación de red | Procedimiento para optimizar el coste y el comportamiento de una red de transmisión en presencia de determinadas exigencias | Básica/Avanzada |
| 8 | Ayuda a evaluación subjetiva | Herramienta para el análisis de material crítico que habrá de incluirse en una prueba de audición | Básica/Avanzada |

## 3 Versiones

Para cumplir de una manera óptima los diferentes requisitos de coste y de funcionamiento, el método de medición recomendado en esta Recomendación presenta dos versiones. La versión básica está concebida para permitir una realización en tiempo real, eficaz con respecto al coste, mientras que la versión avanzada tiene por objetivo principal obtener la mayor exactitud posible. Según la realización, esta exactitud adicional aumenta la complejidad aproximadamente por un factor de cuatro, en comparación con la versión básica.

El Cuadro 1 da directrices sobre la versión que convendría utilizar para cada una de las aplicaciones.

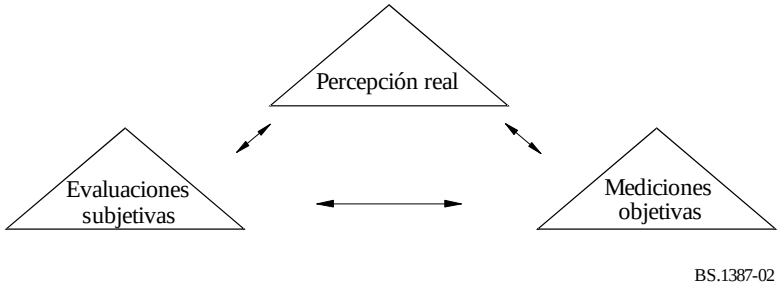
## 4 El dominio subjetivo

Las pruebas de escucha subjetivas formales, por ejemplo las basadas en la Recomendación UIT-R BS.1116, han sido cuidadosamente diseñadas para proporcionar una estimación lo más fiable posible del juicio sobre la calidad de audio. Sin embargo, no cabe esperar que el resultado de una prueba de audición subjetiva refleje plenamente la percepción real. La Fig. 2 muestra las imperfecciones que son inherentes tanto al dominio subjetivo como al dominio objetivo.

Es evidente que no es posible validar directamente un método objetivo. Para validar métodos objetivos hay que compararlos con resultados obtenidos en pruebas de escucha subjetivas.

FIGURA 2

**Conceptos de validación**



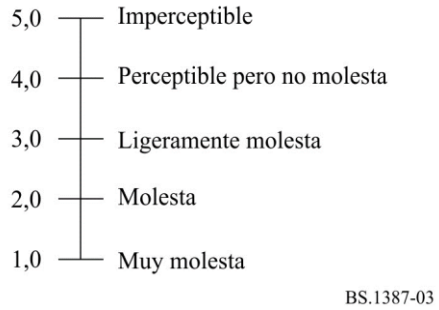

El método de medición objetiva especificado en esta Recomendación se ha concentrado en aplicaciones que normalmente se evalúan en el dominio subjetivo aplicando la Recomendación UIT-R BS.1116. El principio básico de este método de prueba puede describirse brevemente como sigue: el oyente puede elegir entre tres fuentes («A», «B», «C»). La señal conocida de referencia siempre se presenta como fuente «A». La señal de referencia oculta y la señal sometida a prueba se presentan simultáneamente, pero son designadas «al azar» como «B» y «C», según el ensayo de que se trate.

Se pide al oyente que evalúe las degradaciones en «B» en comparación con «A» y las de «C» en comparación con «A», de acuerdo con la escala continua de cinco puntuaciones de degradación. Una de las fuentes, «B» o «C», no deberá poder distinguirse de la fuente «A»; la otra podrá mostrar ciertas degradaciones. Toda diferencia percibida entre la fuente de referencia y la otra fuente ha de interpretarse como una degradación. Normalmente, sólo se utiliza un atributo, «calidad de audio básica». Éste se define como un atributo global que incluye cualquiera y todas las diferencias detectadas entre la señal de referencia y la señal sometida a prueba.

La escala de puntuación calificativa de la degradación se tratará como una escala continua con puntos preestablecidos, derivados de la escala de degradación de cinco puntuaciones del UIT-R especificada en la Recomendación UIT-R BS.1284, indicada a continuación.

FIGURA 3

**Escala de degradación de cinco puntuaciones del UIT-R**




El análisis del resultado de una prueba de audición subjetiva se basa, en general, en el grado de diferencia subjetiva (SDG, *subjective difference grade*), que se define como sigue:

$$SDG = Grade_{Signal\ Under\ Test} - Grade_{Reference\ Signal}$$

Solamente los valores SDG deben estar comprendidos entre 0 y −4, donde 0 corresponde a una degradación imperceptible y −4 a una degradación juzgada como muy molesta.

## 5 Resolución y exactitud

El grado de diferencia objetiva (ODG, *objective difference grade*) es la variable de salida del método de medición objetiva y corresponde al SDG en el dominio subjetivo. La resolución del ODG está limitada a una cifra decimal. Sin embargo, se debe proceder con precaución y no esperar, de una manera general, que una diferencia de un décimo de grado, entre cualquier par de ODG, sea significativa. Esta misma observación es válida cuando se analizan los resultados de una prueba de audición subjetiva.

No existe una figura única que describa completamente la exactitud del método de medición objetiva. Por el contrario, hay que considerar varias cifras de mérito diferentes. Una de ellas es la correlación entre los SDG y los ODG. Es importante comprender que no hay ninguna garantía de que el coeficiente de correlación sea superior a un valor predefinido. Lo más probable es que el comportamiento del método de medición varíe con, por ejemplo, el tipo y el nivel de la degradación introducida.

Otra cifra de mérito que ofrece interés es el número de «valores lejanos» («outliers»). Un «valor lejano» es, por definición, un valor medido que no satisface un esquema de tolerancia predefinido. De acuerdo con las exigencias de los usuarios, el método de medición debe proporcionar la mayor exactitud posible para el extremo superior de la escala de las puntuaciones de degradación (es decir, una calidad de audio alta). En consecuencia, se permite que la exactitud obtenida sea menor en el centro y en la parte baja de la gama de la escala de las puntuaciones de degradación.

Aunque la correlación normalmente da una buena estimación de la exactitud del método de medición objetiva, es importante tener presente que incluso un coeficiente de correlación relativamente alto puede ocultar un comportamiento inaceptable (desde la perspectiva de los «valores lejanos») de un método de medición.

Una tercera cifra de mérito que se ha utilizado durante el proceso de validación es la puntuación de error absoluta (AES, *absolute error score*), que refleja el promedio de la relación entre el tamaño del intervalo de confianza del SDG y la distancia entre el SDG y el ODG.

Para más detalles sobre el comportamiento esperado del método de medición, así como sobre el comportamiento durante el proceso de validación, véase el Adjunto 1 al Anexo 2.

### 6 Requisitos y limitaciones

La señal del dispositivo sometido a prueba y la señal de referencia deben estar alineadas en el tiempo con una exactitud de 24 muestras durante el intervalo de medición completo. El mecanismo de sincronización no forma parte de la presente Recomendación y se espera que varíe de una realización a otra.

# Adjunto 1
# al Anexo 1

# Aplicaciones

### 1 Generalidades

Este Adjunto presenta las definiciones y requisitos específicos de las principales aplicaciones para la cuales se recomienda el método de medición objetiva de la calidad de audio percibida.

Algunas de las aplicaciones requieren una realización en tiempo real del método de medición objetiva mientras que para otras aplicaciones no es necesaria la medición en tiempo diferido. Para las aplicaciones en tiempo real se recomienda que el máximo retardo a través del equipo de medición no sea superior a 200 ms, siendo inaceptable todo retardo superior a 1 s.

Además, se ha hecho una distinción entre las mediciones en línea y fuera de línea. En las mediciones fuera de línea, el procedimiento de medición tiene pleno acceso al equipo o la conexión, mientras que la medición en línea implica un programa en ejecución, que no podrá ser interrumpido por la medición.

### 2 Principales aplicaciones

#### 2.1 Evaluación de realizaciones

Los radiodifusores, los operadores de red, y otros, tienen necesidad de evaluar a diferentes realizaciones de equipo, en particular códecs de audio, cuando seleccionan tal equipo con miras a su adquisición, o cuando se realizan pruebas de aceptación.

Para este tipo de aplicaciones se requiere una gran exactitud, especialmente para evaluar pequeñas degradaciones y clasificar correctamente las diferentes realizaciones. En lo que concierne a las variables de salida, una salida simple, como el ODG, es suficiente para los usuarios, pero los desarrolladores de códecs de audio pueden realizar un análisis más completo utilizando un conjunto adecuado de variables de salida del modelo (MOV).

Pueden utilizarse ambas versiones del modelo, pero se recomienda la versión avanzada.

#### 2.2 Ajuste de calidad perceptual

Este es un procedimiento rápido que se sigue antes de poner en servicio una unidad de equipo o un circuito. Tiene por finalidad verificar la funcionalidad y la calidad. El equipo de medición será manipulado por un personal de operaciones. Cualquier clase de distorsión puede estar presente.

Se requiere la medición en tiempo real. Pueden utilizarse señales de prueba o señales de audio predefinidas. Los ODG deben ser adecuadamente visualizados y presentados al menos dos veces por segundo o, si se utiliza una señal de prueba especial, inmediatamente después de finalizada la señal de prueba.

La versión básica es suficiente para este procedimiento.

### 2.3 Supervisión en línea

Se trata de un proceso continuo, que se aplica a una transmisión de audio en curso. El programa no podrá ser interrumpido por el procedimiento de medición. En consecuencia, es necesario utilizar para la medición la propia señal del programa o un fragmento de audio predefinido. Este último puede ser una señal de una estación o una cuña publicitaria (*jingle*). El equipo de medición será manipulado por un personal de operaciones.

Se requiere la medición en tiempo real. Los ODG deberán ser visualizados adecuadamente y presentados al menos dos veces en un segundo o inmediatamente después de finalizada la señal predefinida. No se desea una presentación visual de las variables MOV.

La versión básica es suficiente para la utilización de este procedimiento.

### 2.4 Estado del equipo o de la conexión

Para asegurar la funcionalidad de las conexiones o equipos de audio, cada cierto tiempo hay que realizar una extensa comprobación de la calidad. A diferencia de la supervisión en línea o el ajuste perceptual, esta aplicación requiere la comprobación de varios parámetros técnicos.

El sistema de medición debe dar una información detallada sobre la influencia del estado del equipo o de la conexión en la calidad de audio percibida presentando visualmente el conjunto completo de variables MOV, además de los ODG. No se requiere la medición en tiempo real.

Se recomienda la utilización de la versión avanzada.

### 2.5 Identificación de códec

Para identificar los códecs (de diferentes algoritmos, o diferentes realizaciones del mismo algoritmo), el sistema de medición deberá poder almacenar, extraer y comparar patrones de características. La similitud entre los patrones puede tomarse como una medida de la similitud de las diferentes realizaciones de códecs. Tal procedimiento se utiliza para identificar el tipo y realización de un determinado códec.

El sistema de medición deberá registrar la mayor cantidad posible de información sobre los patrones. Un análisis limitado exclusivamente a los ODG pudiera no proporcionar una información suficiente.

La utilización de la versión básica es suficiente, aun cuando no se requiera una medición en tiempo real.

NOTA – No se dispone de muchos datos experimentales sobre el método recomendado. Además, todavía no se ha definido una medida única de la similitud entre los patrones.

### 2.6 Desarrollo de códec

Para esta aplicación, el método de medición deberá caracterizar el comportamiento del códec sometido a prueba lo más exactamente y con todos los detalles que sea posible, sobre todo cuando se trate de distorsiones pequeñas.

La supervisión continua el procesamiento en tiempo real, que no es necesariamente admitido por la versión avanzada. Sin embargo, las pequeñas degradaciones y una información detallada exigirán la utilización de la versión avanzada. El sistema de medición deberá poder visualizar las salidas a la misma velocidad a la que se calculan. El acceso directo a la historia de las salidas en un periodo de 4 s es deseable.

Se recomienda la utilización de la versión avanzada. Sin embargo, para la medición en tiempo real basta con la versión básica. Se requiere la medición en tiempo real y en tiempo diferido, así como el análisis trama por trama. Es necesario que se indique toda distorsión severa, por ejemplo por una presentación visual de los valores de cresta. El acceso al conjunto completo de las variables MOV es deseable.

### 2.7 Planificación de red

Para la planificación de las redes es necesaria una evaluación de la calidad esperada en diversos puntos durante el proceso de planificación. Se puede utilizar una simulación informatizada de los componentes de la red que permita combinar diferentes etapas del procesamiento de audio, para examinar diferentes configuraciones con el fin de optimizar la calidad de audio. En una etapa ulterior, los componentes reales del procesamiento de audio podrán probarse en la configuración elegida.

La planificación de la red la efectúan ingenieros de sistema, que deberán obtener una información detallada sobre la influencia de la característica de red en la calidad de audio. La determinación de la clasificación que corresponde a las diferentes configuraciones de red posibles deberá basarse en un conjunto adecuado de variables MOV, que dependerá de la aplicación específica de la red. No basta por tanto con una simple presentación visual de los ODG. En esta aplicación no se requiere una medición en tiempo real para la evaluación.

Pueden utilizarse ambas versiones de modelo, pero se recomienda la versión avanzada.

### 2.8 Ayuda para evaluación subjetiva

El método de medición objetiva proporciona una herramienta para el análisis del material de audio crítico que habrá de utilizarse en las pruebas de escucha subjetivas. El conjunto completo de variables MOV puede utilizarse para establecer las distintas categorías del material crítico.

Se requiere la mayor exactitud posible y se recomienda la utilización de la versión avanzada. No obstante, es deseable la medición en tiempo real, para reducir el tiempo requerido para seleccionar el material crítico.

### 2.9 Resumen de aplicaciones

En el Cuadro 2 se recapitulan los requisitos que deberá satisfacer el método de medición para las principales aplicaciones.

CUADRO 2

**Requisitos que debe cumplir el método de medición**

| | Aplicación | Categoría | Tiempo real | Min, ROV[(1)] (Hz) | En/fuera de línea | Versión del modelo |
|---|---|---|---|---|---|---|
| 1 | Evaluación de realizaciones | Diagnóstico | No | – | Desactivada | Ambas |
| 2 | Ajuste de calidad perceptual | Operacional | S/N | 2 | Desactivada | Básica |
| 3 | Supervisión en línea | Operacional | Sí | 2 | Activada | Básica |
| 4 | Estado del equipo o de la conexión | Diagnóstico | S/N | – | Activada/ desactivada | Avanzada |
| 5 | Identificación de códec | Diagnóstico | No | – | Desactivada | Ambas |
| 6 | Desarrollo de códec | Desarrollo | S/N | – | Desactivada | Ambas |
| 7 | Planificación de red | Desarrollo | S/N | – | Desactivada | Ambas |
| 8 | Ayuda para evaluación subjetiva | Desarrollo | S/N | – | Desactivada | Avanzada |

(1) Velocidad de presentación de los valores a la salida (ROV, rate of output values) (por segundo).

## 3 Señales de prueba

Las señales de prueba pueden dividirse en dos grupos: naturales y sintéticas. Las señales de prueba naturales aquí indicadas consisten en secuencias de audio críticas ya utilizadas en pruebas de escucha realizadas tanto por el UIT-R como por otras organizaciones, para la evaluación de la calidad de audio. Estas señales deberán estar disponibles tanto en el lugar transmisor como en el que se efectúa la medición. Por consiguiente, el dispositivo de medición debe estar dotado de una memoria.

Las señales sintéticas están definidas matemáticamente y pueden ser modificadas de una manera controlada. Estas señales pueden ser generadas en el lugar transmisor y en el lugar en que se efectúa la medición. No se requiere una memoria suplementaria en el dispositivo de medición. Debido a la naturaleza de estas señales, es difícil, si no imposible, establecer para ellas una escala de puntuación para una evaluación subjetiva. En consecuencia, este método de medición no ha sido validado cotejando sus resultados con los resultados subjetivos para estas señales.

### 3.1 Selección de señales de prueba naturales

En el Cuadro 3 se presenta un subconjunto de señales de prueba que fueron utilizadas durante el procedimiento de verificación que condujo a la formulación de esta Recomendación. Se indica también el tipo de los artefactos, que estas señales revelan, típicamente, debido a la codificación a baja velocidad binaria.

CUADRO 3

**Un subconjunto de señales de prueba**

| Nº | Elemento (ítem) | Nombre de fichero | Observaciones |
|---|---|---|---|
| 1 | Castañuelas | cas | (1) |
| 2 | Clarinete | cla | (2) |
| 3 | Claves | clv | (1) |
| 4 | Flauta | flu | (2) |
| 5 | Glockenspiel | glo | (1), (2), (5) |
| 6 | Clavicordio | hrp | (1), (2), (4) |
| 7 | Timbal | ket | (1) |
| 8 | Marimba | mar | (1) |
| 9 | Piano Schubert | pia | (2) |
| 10 | Diapasón-pipa | pip | (4) |
| 11 | Ry Cooder | ryc | (2), (4) |
| 12 | Saxofón | sax | (2) |
| 13 | Gaita | sb1 | (2), (4), (5) |
| 14 | Voz fem. inglesa | sfe | (3) |
| 15 | Voz masc. inglesa | sme | (3) |
| 16 | Voz masc. alemana | smg | (3) |
| 17 | Tambores a timbre | sna | (1) |
| 18 | Soprano Mozart | sop | (4) |
| 19 | Tamboril | tam | (1) |
| 20 | Trompeta | tpt | (2) |
| 21 | Triángulo | tri | (1), (2), (5) |
| 22 | Tuba | tub | (2) |
| 23 | Susanne Vega | veg | (3), (4) |
| 24 | Xilófono | xyl | (1), (2) |

(1) Transitorios: sensibles al pre-eco, dispersión de ruido en el dominio del tiempo.

(2) Estructura tonal: sensible al ruido, aspereza.

(3) Lenguaje natural (combinación crítica de partes tonales y ataques): sensible a la distorsión, dispersión de ataques.

(4) Sonido complejo: somete a esfuerzo al dispositivo sometido a prueba.

(5) Gran anchura de banda: somete a esfuerzo al dispositivo sometido a prueba, pérdida de las frecuencias altas, ruido de alta frecuencia modulado por programa.

### 3.2 Duración

La duración de una señal de prueba natural debe ser aproximadamente la misma que si se utilizara en una prueba de audición. La duración típica es del orden de 10 a 20 s. Es muy probable que la parte crítica de la señal de prueba, que revela a la mayor parte de los artefactos, esté limitada a una pequeña parte de la duración.

Las señales de prueba sintéticas deben ser lo suficientemente largas para que sometan a esfuerzo al códec sometido a prueba, el cual podrá estar dotado de una memoria búfer para la señal de audio codificada. Si se tienen en cuenta estos tamaños de memoria búfer y las constantes de tiempo utilizadas en el método de medición, la duración de cada elemento de prueba individual en una secuencia deberá ser superior a 500 ms. La duración puede ser limitada a un valor tan corto porque no es de esperar que estas señales se utilicen en pruebas de escucha subjetivas.

### 4 Sincronización

Para el procedimiento de medición, la señal sometida a prueba y la señal de referencia estarán sincronizadas entre sí. Esto es aplicable tanto a las señales de prueba naturales como a las sintéticas.

### 5 Cuestiones de derecho de autor

Las señales de prueba indicadas en el Cuadro 3 pueden utilizarse sin estar sujetas a derechos de autor solamente para fines de medición junto con el método de mediciones objetivas descrito en el Anexo 2 de esta Recomendación.

NOTA – Para todas las secuencias es necesario observar las disposiciones relativas a los derechos de autor, sobre todo en lo referente a la UER (disco UER SQAM).

## Adjunto 2
## al Anexo 1

## Variables de salida

### 1 Introducción

El método de medición objetiva descrito en esta Recomendación permite medir la calidad de audio y presentar a la salida un valor que se pretende corresponda a la calidad de audio percibida. El método de medición modela las propiedades fundamentales del sistema auditivo del ser humano. Varias etapas intermedias modelan los efectos fisiológicos y sicoacústicos.

Estas variables de salida intermedias pueden utilizarse para caracterizar los denominados artefactos (distorsiones originadas en la señal procesada, ajenas a ésta). Los parámetros se denominan variables de salida del modelo (MOV, *model output variables*). La etapa final del modelo de medición combina los valores MOV para formar un valor de salida único que corresponde directamente a un resultado esperado de una evaluación subjetiva de la calidad.

### 2 Variables de salida del modelo

El Cuadro 4 contiene una descripción de las MOV utilizadas para predecir los grados de diferencia objetiva. Las variables con el subíndice A se derivan de la parte del modelo basada en un banco de filtros, mientras que las variables con el subíndice B se derivan de la parte del modelo basada en una transformada rápida de Fourier (TRF). Los grados de diferencia objetiva pueden predecirse o bien a partir de la parte del modelo basada en una TRF solamente (versión básica) o a partir de una combinación de partes del modelo basadas en TRF y en banco de filtros (versión avanzada). El promediado siempre se realiza en función del tiempo.

## 3 Calidad de audio básica

El parámetro más conocido en las pruebas de escucha subjetivas es la calidad de audio básica (BAQ, *basic audio quality*). La BAQ se mide como un grado de diferencia subjetiva (SDG), que se calcula como el grado atribuido a la referencia, restado del grado atribuido a la señal sometida a prueba en una prueba subjetiva[1]. Normalmente, el valor del SDG es negativo. El correspondiente parámetro de salida del modelo se denomina grado de diferencia objetiva (ODG). La correspondencia de las MOV con un ODG se basa en un gran número de elementos de prueba fiables (véase el Adjunto 2 al Anexo 2).

CUADRO 4

**Descripción de las variables de salida del modelo**

| Variable de salida del modelo | Descripción |
|---|---|
| $WinModDiff_B$ | *Diferencia de (envolventes de) modulación promediada ventanizada entre la señal de referencia y la señal sometida a prueba* |
| $AvgModDiff1_B$ | *Diferencia de modulación promediada* |
| $AvgModDiff2_B$ | *Diferencia de modulación promediada con énfasis en las modulaciones y los cambios de la modulación introducidos cuando la referencia contenga poca modulación, o ninguna* |
| $RmsModDiff_A$ | *Valor medio cuadrático (Rms) de la diferencia de modulación* |
| $RmsMissingComponents_A$ | *Valor medio cuadrático (Rms) de la sonoridad de ruido de los componentes de frecuencia que faltan (utilizados en* $RmsNoiseLoudAsym_A$*)* |
| $RmsNoiseLoud_B$ | *Valor medio cuadrático (Rms) de la sonoridad del ruido promediada con énfasis en los componentes introducidos* |
| $RmsNoiseLoudAsym_A$ | $RmsNoiseLoudA + 0.5RmsMissingComponents_A$ |
| $AvgLinDist_A$ | *Una medida de las distorsiones lineales promedio con un mayor peso en las bandas de frecuencias altas* |
| $BandwidthRef_B$ | *Anchura de banda de la señal de referencia* |
| $BandwidthTest_B$ | *Anchura de banda de la señal de salida del dispositivo sometido a prueba* |
| $TotNMR_B$ | *Logaritmo de la razón de ruido total enmascaramiento* |
| $RelDistFrames_B$ | *Fracción relativa de tramas para las cuales al menos una banda de frecuencias contiene un componente de ruido apreciable* |
| $AvgSegmNMR_B$ | *El logaritmo promediado segmentalmente de la relación ruido-enmascaramiento* |
| $MFPD_B$ | *Máximo de la probabilidad de detección tras un filtrado de paso bajo* |
| $ADB_B$ | *Bloque distorsionado promedio, tomado como el logaritmo de la razón de la distorsión total al número total de tramas severamente distorsionados* |
| $EHS_B$ | *Estructura armónica del error en función del tiempo* |

[1] Véase la Recomendación UIT-R BS.1116.

El ODG es el parámetro medido objetivamente que corresponde a la calidad percibida subjetivamente. Puesto que el cometido del oyente en una prueba de audición es evaluar la BAQ de un elemento de prueba, el ODG es asimismo una medida de la BAQ.

## 4 Margen de codificación

Otro parámetro que en el futuro puede resultar muy útil es el margen de codificación (CM, *coding margin*), que es una forma de describir artefactos inaudibles. El margen de codificación subjetivo (SCM, *subjective coding margin*) puede ser evaluado amplificando los artefactos hasta que se hagan audibles por un participante en la prueba. El SCM describe el margen con respecto al umbral de audibilidad de artefactos.

Para hallar el umbral, los artefactos tienen que ser amplificados o atenuados durante la prueba de audición. Un método adecuado es el método de la diferencia. La señal de diferencia entre la señal original y la señal codificada, sincronizadas en entre sí, se amplifica y se suma a la señal original. La mejor forma de efectuar la detección del umbral de audibilidad es mediante un método de elección forzada. El SCM se obtiene promediando los valores de umbral para amplificación o atenuación obtenidos de los participantes en las pruebas. Los valores CM negativos representan artefactos audibles, mientras que los valores CM positivos representan artefactos inaudibles. A diferencia de la BAQ, el margen de codificación es una medida del instante (y del nivel) en que los artefactos se hacen audibles, y no del grado en que los artefactos son molestos. La definición y validación del método para medir el SCM se presentan en [Feiten, 1997].

El margen de codificación objetivo (OCM, *objective coding margin*) se obtiene también de las MOV. En la actualidad, sólo unos pocos elementos de prueba para el margen de codificación objetivo han sido evaluados. La correspondencia del modelo descrito en esta Recomendación al OCM no ha sido aún investigada.

## 5 Requisitos de usuario

Los requisitos de usuario con respecto a las variables de salida que deben ser satisfechos por el método de medición son diferentes según la aplicación. En algunas aplicaciones, por ejemplo las aplicaciones 2 y 3 (véase el Adjunto 1 al Anexo 1), la medición forma parte de un procedimiento operativo. En estos casos es muy importante que la salida proporcionada por el método sea fácil de leer e interpretar por personas que no tengan un gran conocimiento sobre la técnica de medición. Para ello, lo mejor es que el método presente **un valor único** que corresponda a la calidad de audio percibida.

Lo mismo cabe decir en cuanto a otras aplicaciones, por ejemplo, las aplicaciones 1 y 4. Sin embargo, con relación a éstas, así como con relación a las aplicaciones 5-8, unas variables de salida más complejas pueden ofrecer ventajas a los usuarios que tengan un mayor conocimiento sobre los mecanismos del método de medición.

# Adjunto 3
# al Anexo 1

## Descripción general del modelo

De acuerdo con la Recomendación UIT-R BS.1116, se obtiene un SDG para un elemento de prueba de audio en una prueba de audición, y el SDG medio para un número de oyentes representa la calidad subjetiva del elemento. El elemento puede contener diferentes tipos de distorsiones acústicas, por lo que las variaciones en la calidad están integradas en el tiempo. Por consiguiente, la predicción del SDG basada en mediciones físicas requiere un modelo exacto del sistema auditivo periférico así como aspectos cognoscitivos de los juicios sobre la calidad de audio.

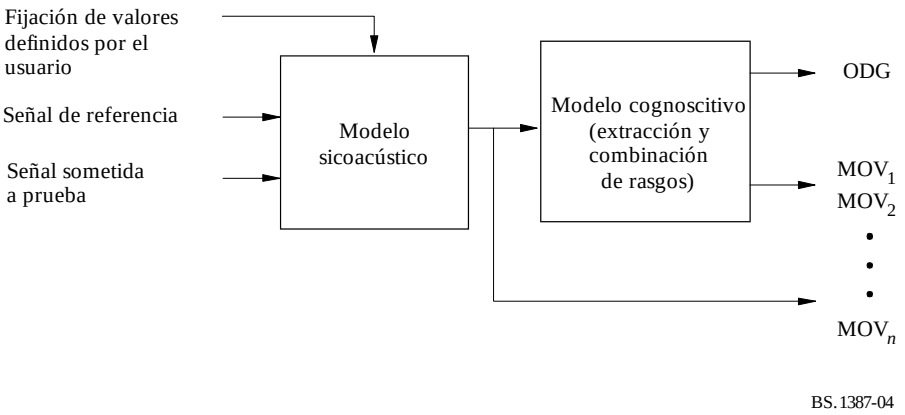
El modelo recomendado para la medición objetiva produce un número de variables de salida del modelo (MOV) basadas en comparaciones entre la señal de referencia y la señal sometida a prueba. Estas MOV se hacen corresponder con un ODG utilizando una técnica de optimización que minimiza la diferencia cuadrática entre la distribución del ODG y la correspondiente distribución de los SDG medios para un conjunto de datos suficientemente grande.

Se describen dos variantes del modelo: una versión basada en la transformada discreta de Fourier (DFT), que pudiera utilizarse para la supervisión en tiempo real, y otra versión, basada en un banco de filtros y en la DFT, que se esperaba que diera resultados más exactos. La versión basada en la DFT se conoce por la versión básica, y la versión combinada se conoce por la versión avanzada.

La estructura de alto nivel de la versión básica y la de la versión avanzada se muestran en la Fig. 4.

FIGURA 4

**Etapas de procesamiento incluidas en el modelo**




### 1 Procesamiento de audio

Al igual que en las pruebas de escucha subjetivas, la calidad de la señal de prueba se juzga con relación a una señal de referencia. Tanto la señal de referencia como la señal sometida a prueba (señales monoaurales o estereofónicas) son transformadas en sus respectivas representaciones sicoacústicas. Estas representaciones son comparadas en orden para obtener un ODG. Estas operaciones se realizan en las etapas de procesamiento que se muestran en la Fig. 4.

### 1.1 Valores definidos por el usuario

El método de medición requiere que el nivel de escucha utilizado se proporcione como un parámetro. En consecuencia, el usuario tiene que suministrar el nivel de presión sonora en dB SPL producido por una onda sinusoidal completa de 1 019,5 Hz. En caso de que no se conozca el nivel de escucha exacto, se recomienda suponer un nivel de escucha de 92 dB SPL.

### 1.2 Modelo sicoacústico

El modelo sicoacústico transforma tramas sucesivas de la señal en el dominio del tiempo en una representación en la membrana basilar. El proceso comienza por la utilización tanto de una DFT como de un banco de filtros. La DFT transforma los datos al dominio de la frecuencia, y el resultado cambia de una escala de frecuencias a una escala de alturas del sonido (*pitch*), que es el equivalente sicoacústico de frecuencia. En la parte del modelo basada en un banco de filtros, la correspondencia de frecuencia a tono (sensación tonal) se tiene directamente en cuenta por las anchuras de banda y el espaciamiento de los filtros de paso de banda.

Se utilizan dos conceptos diferentes para lograr un enmascaramiento simultáneo. Algunas MOV se calculan utilizando el *concepto de umbral enmascarado*, mientras que otras se calculan mediante una *comparación de representaciones internas*. El primer concepto calcula directamente un umbral enmascarado utilizando funciones de enmascaramiento sicofísico. Las variables de salida del modelo se basan en la distancia de la señal de error físico al umbral enmascarado. En la comparación de representaciones internas, las energías de la señal sometida a prueba y de la señal de referencia se ensanchan a regiones de alturas de sonido adyacentes, para obtener patrones de excitación. Las variables de salida del modelos se basan en una comparación entre estos patrones de excitación. El enmascaramiento no simultáneo se realiza «explayando» («smearing») en el tiempo las representaciones de señal.

El umbral absoluto se modela en parte aplicando una función de ponderación dependiente de la frecuencia y en parte añadiendo, a los patrones de excitación, un desplazamiento dependiente de la frecuencia. Este umbral es una aproximación de la mínima presión audible [ISO 389-7, Acoustics – Reference zero for the calibration of audiometric equipment – Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions, 1996].

Las principales salidas del modelo sicoacústico son la excitación y el umbral enmascarado en función del tiempo y de la frecuencia. La salida del modelo en varios niveles está disponible para un ulterior procesamiento.

### 1.3 Modelo cognoscitivo

El modelo cognoscitivo condensa la información de una secuencia de tramas producidas por el modelo sicoacústico. Las fuentes de información más importantes para la realización de mediciones de calidad son las diferencias entre la señal de referencia y la señal sometida a prueba tanto en el dominio de la frecuencia como en el dominio de la altura del sonido. En el dominio de la frecuencia, se miden las anchuras de banda espectrales de ambas señales, así como la estructura de armónicos en el error. En el campo de la altura del sonido, medidas de error se derivan de la modulación de envolvente de la excitación y de la magnitud de la excitación.

Las características calculadas son ponderadas, de modo que su combinación dé por resultado un ODG que esté suficientemente próximo al SDG para la distorsión de audio particular de que se trate. La versión básica utiliza 11 características para producir un ODG, mientras que la versión avanzada utiliza cinco características. La optimización se efectuó utilizando el algoritmo de aprendizaje de red neuronal de retropropagación (véase el Anexo 2, § 6). Los datos de entrenamiento procedían de la totalidad de las bases de datos 1 y 2, y de una parte de la base de datos 3. Se obtuvieron datos de

pruebas de generalización del resto de la base de datos 3 y de la totalidad del conjunto de datos CRC97 (véase al Adjunto 2 al Anexo 2).

# Adjunto 4
# al Anexo 1

## Principios y características de los métodos de medición objetiva de la calidad de audio percibida

### 1 Introducción e historia

La transmisión digital y el almacenamiento de señales de audio se basan cada vez en mayor medida en algoritmos de reducción de datos, que están adaptados a las propiedades del sistema auditivo humano y se basan de manera particular en efectos de enmascaramiento. Tales algoritmos no tienen por objeto principal la minimización de las distorsiones, sino más bien tratan de manipular estas distorsiones de modo que sean percibidas lo menos posible. La calidad de estos codificadores perceptuales ya no puede ser evaluada por métodos de medición convencionales, que normalmente determinan el valor global de la distorsión. Un ejemplo que suele mencionarse para ilustrar estas limitaciones es el denominado milagro de los 13 dB: un ruido superpuesto con una estructura adaptada a la señal de audio es casi inaudible incluso si la relación *S*/*N* no ponderada resultante declina a 13 dB.

Por esta razón, las evaluaciones de los códecs perceptuales requieren una prueba de audición para evaluar la calidad de audio. Para que las pruebas de escucha puedan ser suficientemente fiables y repetibles es necesario realizar operaciones que exigen mucho tiempo y trabajo.

Los esquemas de mediciones objetivas que incorporan propiedades de un sistema auditivo humano pueden ayudar a superar estos problemas. Esta idea fue dada a conocer por primera vez por [Schroeder y otros, 1979]. En su publicación, que trata principalmente sobre la codificación de voz, se describe el esquema de medición «sonoridad del ruido (NL, *noise loudness*)».

En dicha publicación, la sonoridad percibida de la señal de ruido del códec de voz, que es la diferencia entre su señal de entrada y su señal de salida, se estima para cada trama de tiempo de aproximadamente de 20 ms. Si la señal de ruido queda completamente enmascarada, la sonoridad percibida es cero. Un enmascaramiento parcial reduce la sonoridad de la señal de ruido no enmascarada. El umbral enmascarado utilizado es optimizado para ruido enmascarador de tono, y la degradación final de la voz se calcula para cada trama. No se calcula un resumen de la calidad total de la voz.

En 1985, Karjalainen presentó el sistema de medición «Diferencia especial auditiva (ASD, *auditory spectral difference*)» [Karjalainen, 1985]. Este investigador partió de varias ideas de Schroeder, Atal y Hall, pero reemplazó el análisis basado en trama por un banco de filtros en el cual los filtros se superponían, cambió la manera en que se incluye el umbral absoluto y añadió un modelo para enmascaramiento temporal. Las dos señales de entrada al esquema de medición se procesan exactamente de la misma manera y producen una especie de representación interna. Estas representaciones internas se comparan una con otra para explicar diferencias percibidas entre la señal de entrada y la señal de salida de un esquema de codificación de la voz. No se calcula un resumen de la calidad total de una muestra de voz. La resolución temporal de ASD se adapta mejor a las propiedades del sistema auditivo humano, pero aumenta la complejidad del algoritmo.

En 1987, Brandenburg presentó el esquema de medición «Relación ruido-enmascaramiento (NMR, *noise to mask ratio*)» [Brandenburg, 1987], que estaba destinado a ser utilizado como una herramienta para el desarrollo de esquemas de codificación de audio. La complejidad del esquema se redujo en comparación con NL calculando la expansión (*spreading*) en bandas perceptuales utilizando una función de expansión que fue diseñada como una curva de caso más desfavorable. El umbral enmascarado utilizado está optimizado para tono que enmascara ruido. Se añadió un esquema simple del modelado posenmascaramiento y varias formas para evaluar la calidad percibida de extractos largos de audio. Este esquema fue el primero que se realizó en equipo físico en tiempo real.

En 1989, Moore y Glasberg [Moore, 1989] presentaron un modelo perceptual pero no presentaron una forma para juzgar la calidad percibida de señales de audio degradadas.

## 2 Estructura general de los métodos de medición objetiva de la calidad de audio percibida

Todos los esquemas de medición perceptual trabajan con dos señales de entrada: una se denomina la señal de referencia (REF), y la otra la señal sometida a prueba (SUT, *signal under test*). En aquellas situaciones en que la señal de referencia no pueda transmitirse al equipo de medición, pero es bien conocida, se podrá utilizar como señal de referencia una referencia interna almacenada en el propio equipo de medición. Es esencial que las señales de entrada estén alineadas en el equipo.

La incorporación del aspecto sicoacústico en los esquemas de medición puede efectuarse de dos maneras diferentes. La primera es muy similar a la manera en que están estructurados los esquemas de codificación de audio: la señal de referencia se utiliza para calcular una estimación del nivel enmascarado real (véase más adelante). La diferencia entre la señal sometida a prueba y la señal de referencia se compara con este umbral enmascarado. Este método se denomina «concepto de umbral enmascarado» y se utiliza en los procedimientos sonoridad del ruido y MNR. La diferencia entre las señales de entrada puede calcularse en el dominio del tiempo o como la diferencia entre los espectros de energía a corto plazo. Este último procedimiento proporciona una mayor robustez contra errores de alineación en el tiempo, pero disminuye la resolución temporal. La diferencia en el dominio del tiempo por lo general es demasiado sensible a las distorsiones de fases y, por esta razón, ha dejado de utilizarse.

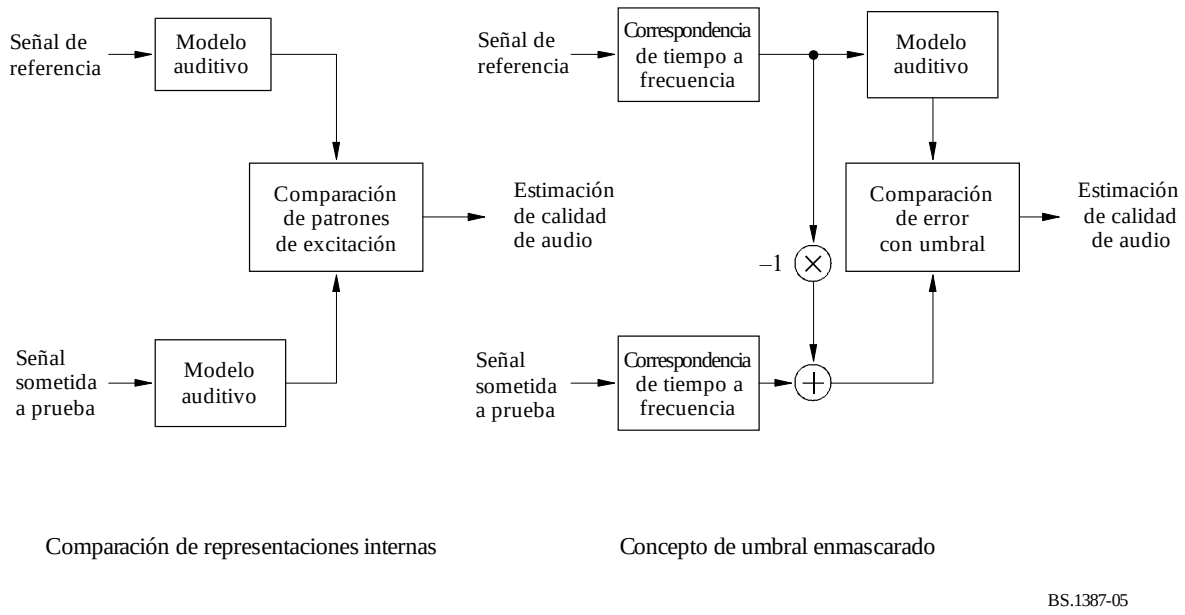
La segunda está más cerca de los procesos fisiológicos del sistema auditivo humano: Se calcula una denominada representación interna de la señal de referencia y de la señal sometida a prueba. Esta representación interna es una estimación de la información disponible por el cerebro humano para la comparación de señales. Este método se denomina «comparación de representaciones internas» y se utiliza en ASD.

## 3 Aspectos sicoacústicos y cognoscitivos

En este punto se analizan las propiedades del sistema auditivo humano que desempeñan un papel más importante en la evaluación de la calidad percibida de las señales de audio. Se trata principalmente de la manera de modelar estas propiedades.

FIGURA 5

**Conceptos sicoacústicos utilizados en diferentes esquemas de medición perceptual**




## 3.1 Características de transferencia del oído externo y del oído medio

En general, las señales sonoras tienen que pasar el oído externo y el oído medio antes de llegar al oído interno, que es donde se realizan los procesos de detección y análisis de los sonidos. El oído externo y el oído medio actúan como un filtro paso banda sobre la señal de entrada. A la señal de entrada viene a sumarse el ruido presente en el nervio auditivo así como también el causado por la circulación de la sangre. La amplitud de este ruido aumenta con las frecuencias bajas. La función de transferencia del oído externo y del oído medio, junto con el ruido interno, limitan la aptitud para detectar señales de audio débiles, y ejercen una gran influencia en el umbral absoluto de audibilidad.

## 3.2 Escalas de frecuencia perceptuales

Los receptores de la presión acústica en el oído humano son los cilios. Los cilios se encuentran en el oído interno, más precisamente en la cóclea (el caracol). En la cóclea, se efectúa una transformación de frecuencia a posición. La posición de la máxima excitación depende de la frecuencia de la señal de entrada. Cada cilio en una posición dada en la cóclea responde a una gama de frecuencias en la escala de frecuencias; estas gamas se superponen unas a otras. La impresión perceptual de la altura del sonido (*pitch*) está correlacionada con una distancia constante de los cilios.

Según el experimento sicoacústico utilizado, se han encontrado diferentes funciones de transformación de la frecuencia a la altura del sonido:

En [Zwicker y Feldtkeller, 1967] se presenta un cuadro que divide la escala de frecuencias en Hz en 24 bandas que no se superponen, y que se han denominado bandas críticas. Las frecuencias de corte superior de estas bandas se indican en el Cuadro 5. El Cuadro representa también una definición de la escala Bark: 1 Bark corresponde a 100 Hz, 24 Bark corresponden a 15 500 Hz.

CUADRO 5

**Escala de bandas críticas definidas por Zwicker**

| **Banda crítica** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frecuencia de corte superior (Hz) | 100 | 200 | 300 | 400 | 510 | 630 | 770 | 920 | 1 080 | 1 270 | 1 480 | 1 720 |

| Banda crítica | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frecuencia de corte superior (Hz) | 2 000 | 2 320 | 2 700 | 3 150 | 3 700 | 4 400 | 5 300 | 6 400 | 7 700 | 9 500 | 12 000 | 15 500 |

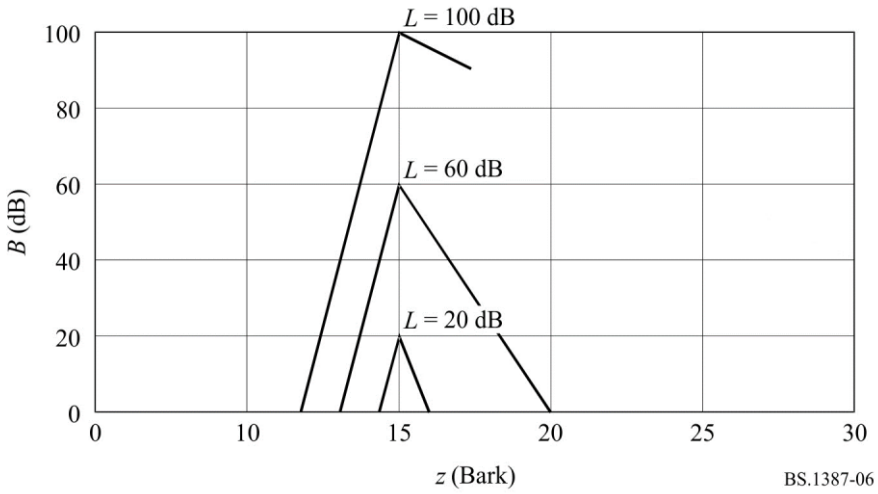
En el pasado se hallaron varias aproximaciones a la escala Bark. En [Cohen y Fielder, 1992] se presenta un análisis detallado de diferentes escalas. En el contexto de la medición objetiva de la calidad de audio percibida, los mejores resultados se obtuvieron utilizando la escala Bark.

## 3.3 Excitación

Cada cilio reacciona a una gama de frecuencias que puede describirse por una característica de filtro. La pendiente de los filtros puede expresarse convenientemente en una escala perceptual, como se ha indicado anteriormente. La forma de los filtros en tal escala es casi independiente de la frecuencia central. La pendiente inferior de la excitación es independiente del nivel L de la señal de entrada (aproximadamente 27 dB/Bark). La pendiente superior es más empinada para los niveles más bajos que para las más altos de la señal de entrada (−5 a −30 dB/Bark). Esta característica empinada es causada por un mecanismo de retroalimentación entre dos clases diferentes de cilios y se necesita cierto tiempo para que actúe. Por tanto, la mejor resolución de frecuencia para la audición se consigue en el caso de señales estacionarias varios milisegundos después de la aparición de la señal. Los patrones de excitación de señales constituidas por varios componentes se suman según una ley no lineal.

FIGURA 6

**Dependencias de la excitación con respecto al nivel, según Terhardt [1979]**



Tras la exposición a una señal, los cilios y el procesamiento neuronal necesitan cierto tiempo para volver a alcanzar la plena sensibilidad. La duración del proceso de recuperación depende del nivel y de la duración de la señal y puede durar hasta varios cientos de milisegundos. Las señales de alto nivel son procesadas más rápidamente que las de bajo nivel en su recorrido entre un cilio y el cerebro. Por consiguiente, la aparición de una señal de elevado nivel sonoro puede enmascarar una señal precedente de un nivel más bajo.

Otro planteamiento para el modelado de la excitación se basa en la escala de anchura de banda rectangular equivalente (ERB, *equivalent rectangular bandwith*) [Moore, 1986]. En este planteamiento se utilizan los denominados filtros exponenciales redondeados (ROEX, *rounded exponential*) [Moore, 1986]. En el contexto de la medición objetiva de la calidad de audio percibida, los mejores resultados se han obtenido con modelos basados [Zwicker y Feldtkeller, 1967] y [Terhardt, 1979].

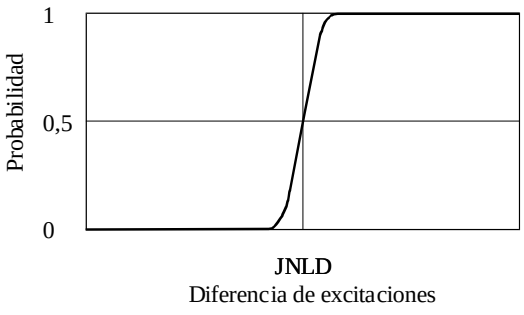
## 3.4 Detección

Las excitaciones de diferentes señales de audio se transfieren al cerebro humano. Hay tres clases de memoria que se diferencian entre sí por el grado de detalle y por el tiempo en que la información está presente: memoria a largo, corto y ultracorto plazo. En el contexto de las pruebas de escucha, las memorias ultracortas desempeñan el papel más importante. La mayor parte de los detalles de una señal se conservan si la duración de un extracto de audio es inferior a unos cinco u ocho segundos, lo que dependerá del oyente y del extracto de audio. Esto se tiene en cuenta en el procedimiento de evaluación definido en la Recomendación UIT-R BS.1116 en el que se permite a los sujetos seleccionar partes muy cortas de un extracto de audio para escucharlo con mayor atención. En el umbral de detección, la probabilidad de detección es 50%. En torno al umbral, la probabilidad de detección de diferencias aumenta uniformemente de 0% a 100%.

La diferencia de nivel justamente perceptible (JNLD, *just-noticeable level difference*) es el umbral de detección de las diferencias de nivel. La JNLD es influida por el nivel de las señales de entrada. En el caso de señales de bajo nivel, se requieren grandes diferencias para la detección (nivel: 20 dB SPL, JNLD: 0,75 dB). En el caso de señales de alto nivel, la sensibilidad a las pequeñas diferencias es mucho mayor (nivel: 80 dB SPL, JNLD: 0,2 dB). Estos números se basan en experimentos de modulación de amplitud.

FIGURA 7

**Principio de la probabilidad de detección**




## 3.5 Enmascaramiento

Una señal que es claramente audible si aparece sola puede ser completamente inaudible en presencia de otra señal, la enmascaradora. Este efecto se denomina enmascaramiento y la señal enmascarada se denomina la enmascarada. Se deben distinguir dos situaciones:

– *Enmascaramiento simultáneo*

En esta situación, la señal enmascaradora y la enmascarada se presentan al mismo tiempo y son casi estacionarias. Si la enmascaradora tiene una anchura de banda discreta, el umbral de audibilidad se eleva incluso para frecuencias por debajo o por encima de la enmascaradora. La cantidad de enmascaramiento depende de las estructuras de la enmascaradora y la enmascarada. En el caso de que una señal similar al ruido esté enmascarando una señal tonal, la cantidad de enmascaramiento es casi independiente de la frecuencia. Si el nivel de presión sonora de la enmascarada está aproximadamente 5 dB por debajo del nivel de la enmascaradora, aquélla es inaudible. En el caso de que una señal tonal esté enmascarando una señal similar al ruido, la cantidad de enmascaramiento depende de la frecuencia de la enmascaradora. Esta cantidad puede estimarse por la fórmula $\left(15{,}5 + \frac{z}{Bark}\right)$ dB, donde $z$ es la tasa de la banda crítica de la enmascaradora. Además, a altos niveles de señal, los efectos no lineales reducen el umbral de enmascaramiento cerca de la enmascaradora. Efectos similares se producen cuando un tono enmascara a otro tono. Los umbrales enmascarados de varias señales se suman según una ley no lineal. En general, el umbral de enmascaramiento resultante está por encima del umbral de enmascaramiento producido por cada señal individual.

– *Enmascaramiento temporal*

En esta situación, la enmascaradora y la enmascarada se presentan en instantes de tiempo diferentes. Poco después de la caída de nivel de una enmascaradora, el umbral de enmascaramiento está más cerca del enmascaramiento simultáneo de esta enmascaradora que del umbral absoluto. Dependiendo de la duración de la enmascaradora, el tiempo de caída del umbral puede estar comprendido entre 5 ms (señal enmascaradora: un impulso gaussiano con una duración de aproximadamente de 0,05 ms) y 150 ms, o más (señal enmascaradora: ruido rosado con una duración de 1 s). Las señales débiles que aparecen justamente antes de señales de mayor nivel, son enmascaradas. La duración de este efecto de enmascaramiento hacia atrás es de aproximadamente de 5 ms. Si la enmascarada está justamente por encima del umbral, no se percibe antes de la enmascaradora, sino como un cambio de la enmascaradora. El enmascaramiento hacia atrás muestra grandes desviaciones de un oyente a otro.

### 3.6 Sonoridad y enmascaramiento parcial

La sonoridad percibida de las señales de audio depende de su frecuencia, su duración y su nivel de presión sonora. Debido al autoenmascaramiento, la sonoridad de una señal compleja es menor que la suma de las sonoridades de todos sus componentes. En el contexto de la medición de la calidad de audio, la sonoridad de la distorsión no deseada sumada a la señal de referencia, lo que se conoce por sonoridad del ruido, se reduce por el enmascaramiento parcial causado por la señal de referencia.

### 3.7 Nitidez

La nitidez, que es uno de los valores básicos de la sensación, está relacionada con el timbre. Un sonido se percibe como nítido si contiene principalmente componentes de alta frecuencia. Por ejemplo, un tono sinusoidal o un ruido limitado en banda a altas frecuencias, o un ruido de paso alto con una frecuencia de corte en la gama de frecuencias por encima de unos 3 kHz se dice que es nítido. La estructura de frecuencia detallada de la señal de audio, sin embargo, no tiene una gran influencia en la nitidez. Von Bismarck [1974] ha realizado una investigación fundamental sobre la nitidez.

Las investigaciones adicionales sobre la nitidez se realizaron por [Auras, 1989]. Como resultado de estas investigaciones se obtuvo una función de ponderación ligeramente modificada, comparada con la función de ponderación definida por Bismarck. La función de ponderación de Aures contribuye menos a la atribución de nitidez en bandas críticas muy bajas y muy altas, y contribuye más a tasas de bandas críticas entre 14 y 20 Bark. Además, estas investigaciones han demostrado que la nitidez de las señales de audio con una alta variación del nivel de presión sonora de la señal de audio y un gran contenido de alta frecuencia no puede basarse solamente en la sonoridad global, sino en una función de ponderación, que depende de la sonoridad global.

### 3.8 Procesamiento cognoscitivo

Es claro que la calidad de audio percibida es muy influida por efectos cognoscitivos. Esto puede demostrarse por un sencillo experimento.

Una señal de referencia con un ruido de fondo claramente audible es procesada por un equipo de audio que no sea capaz de transmitir este ruido de fondo. Como el ruido es una distorsión no deseada, la señal de referencia sería evaluada peor que la señal procesada en una prueba de audición. Por otra parte, la misma señal procesada obtendría una puntuación peor si la parte más importante de la señal de referencia fuera el ruido de fondo ligero.

Aunque una enumeración de todos los efectos cognoscitivos posibles está fuera del ámbito del presente Adjunto, se citan algunos ejemplos:

*Ejemplo 1: Separación de las distorsiones lineales de las no lineales*

Las distorsiones lineales son menos molestas que las distorsiones no lineales. La separación de las distorsiones lineales de las no lineales puede realizarse con bastante facilidad mediante un filtrado inverso adaptativo de la señal de salida. En el método especificado por esta Recomendación se utiliza una separación de las distorsiones lineales de las no lineales.

*Ejemplo 2: Análisis de la escena auditiva*

Análisis del escenario auditivo [Bregman, 1990] es un proceso cognoscitivo que permite a los oyentes separar eventos auditivos diferentes y agruparlos en objetos diferentes. Un planteamiento pragmático como el indicado en [Beerends y Stemerdink, 1994] fue útil para cuantificar un efecto de análisis de escenario auditivo. Si un componente de tiempo-frecuencia no es codificado por un códec, la señal restante continúa formando una escenario auditivo coherente, mientras que la introducción de un nuevo componente de tiempo-frecuencia no relacionado conduce a dos percepciones diferentes. Debido a la división en dos percepciones diferentes, la distorsión será más molesta que lo que cabría esperar atendiendo a la sonoridad del componente de distorsión introducido en último lugar. Esto conduce a una asimetría entre la perturbación percibida de una distorsión que es causada por la ausencia de codificación de un componente de tiempo-frecuencia frente a la perturbación causada por la introducción de un nuevo componente de tiempo-frecuencia.

*Ejemplo 3: Enmascaramiento informativo*

El enmascaramiento informativo puede ser modelado definiendo una medida de complejidad espectral-temporal similar a la entropía. Lo más probable es que el efecto dependa del adiestramiento que los sujetos han recibido antes de efectuar la evaluación subjetiva. Un primer intento de modelar este efecto se presenta en [Beerends y otros, 1996], donde se calcula una estimación de complejidad local en una ventana de tiempo de aproximadamente 100 ms. Si esta complejidad local es alta, las distorsiones dentro de esta ventana de tiempo son más difíciles de oír que cuando la complejidad local es baja. El adiestramiento puede reducir el umbral de enmascaramiento en varias decenas de decibelios [Leek y Watson, 1984].

*Ejemplo 4: Ponderación espectral-temporal*

Algunas regiones espectrales-temporales en la señal de audio transportan más información, y, por lo tanto, pueden ser más importantes que otras. Se ha observado que la ponderación espectral-temporal es importante en los juicios de calidad sobre códecs de voz. En la voz, algunos componentes espectrales-temporales, tales como los formantes, evidentemente transportan más información que otros [Beerends y Stemerdink, 1994]. En la música, sin embargo, todos los componentes espectrales-temporales de la señal, incluso los periodos de silencio, pueden transportar información.

## 4 Modelos incorporados

### 4.1 DIX

El método de medición perceptual DIX (*disturbance index*) [Thiede y Kabot, 1996] se basa en un banco de filtros auditivos que proporcionan una alta resolución temporal y, por tanto, permiten (en comparación con los procedimientos basados en TFR) un modelado más preciso de efectos temporales tales como el preenmascaramiento y el posenmascaramiento. La estructura fina temporal de las envolventes en cada filtro auditivo se preserva y se utiliza para obtener información adicional sobre las señales y las distorsiones introducidas.

Las frecuencias centrales de los filtros individuales están uniformemente distribuidas en una escala perceptual de alturas del sonido. La parte superior de la curva que define la forma del filtro está ligeramente redondeada para asegurar que el número elegido de filtros abarca la totalidad de la gama de frecuencias sin ondulaciones en la respuesta a frecuencia global. Para modelar umbrales enmascarados, las pendientes de los filtros decrecen exponencialmente en la escala Bark. El empinamiento de las pendientes del filtro depende del nivel de las señales de entrada. La gama de frecuencias audibles fue cubierta por 80 filtros en la primera versión de DIX y posteriormente fue cubierta por 40 filtros, es decir, la resolución de frecuencia corresponde aproximadamente a 0,6 Bark. El algoritmo de banco de filtros es bastante rápido en comparación con el de otros bancos de filtrado en los que los filtros son individuales, pero toma todavía mucho más tiempo que los algoritmos de las transformadas basadas en bloques como TFR y los de las transformadas basadas en conjuntos de ondículas (*wavelet-package-transforms*).

DIX adapta dinámicamente los niveles y espectros entre la señal sometida a prueba y la señal de referencia con el fin de separar las distorsiones lineales de las no lineales. Evalúa la estructura de las envolventes temporales en las salidas de filtro para modelar el enmascaramiento incrementado causado por señales enmascaradoras moduladas y por señales enmascaradas similares al ruido, en comparación con tonos puros.

Mediante una comparación de las representaciones internas de la señal sometida a prueba y de la señal de referencia se calcularon numerosas parámetros de salida, entre ellos la sonoridad parcial de distorsiones no lineales, indicadores de la cantidad de distorsión no lineal y medidas de efectos temporales y binaurales. Sin embargo, se puede obtener una buena estimación de la calidad de audio

básica utilizando solamente dos de los parámetros de salida: la sonoridad parcial de distorsiones no lineales junto con uno de los indicadores de la cantidad de distorsión lineal se hacen corresponder con una estimación de la calidad de audio básica esperada de la señal sometida a prueba.

### 4.2 NMR

El esquema de medición NMR (razón de ruido a enmascarado, *noise-to-masked-ratio*) [Brandenburg, 1987] evalúa la diferencia de nivel entre el umbral de enmascaramiento y la señal de ruido. Se utiliza una DFT con una ventana Hann de aproximadamente 20 ms para analizar el contenido de frecuencia de la señal. Los coeficientes de la transformada se combinan en bandas de acuerdo con la escala Bark. El umbral de enmascaramiento se estima para cada banda. La pendiente del umbral de enmascaramiento se obtiene utilizando un procedimiento de caso más desfavorable que tiene en cuenta que las pendientes son más empinadas para señales débiles pero alcanzan el umbral absoluto a niveles más altos. El nivel absoluto se adapta a la resolución de la señal de entrada (usualmente de 16 bits), pero no a las exigencias sicoacústicas. Debido a estas circunstancias, la NMR es reacia a cambios en el nivel de reproducción. La resolución de la escala de alturas del sonido es de aproximadamente 1 Bark. Dado que el poder de cálculo requerido era bajo, fue posible realizar NMR como un sistema en tiempo real en una etapa inicial de su desarrollo.

Este modelo se ha utilizado desde 1987 y ha demostrado su fiabilidad básica.

Los valores de salida más importantes de NMR son la tasa de aviso de enmascaramiento (*masking flag rate*), que da el porcentaje de tramas con distorsiones audibles, así como la NMR total y la NMR media, que son formas diferentes de promediar la distancia entre la energía de error y el umbral de enmascaramiento.

### 4.3 OASE

El esquema de medición de evaluación objetiva de señal audio (OASE, *objective audio signal evaluation*) [Sporer, 1997] utiliza un banco de filtros, formado por 241 filtros, para analizar las señales de entrada. Las frecuencias centrales están uniformemente espaciadas en la escala Bark con una distancia de 0,1 Bark. Los filtros se superponen unos a otros. Cada uno de los filtros está adaptado a la respuesta a frecuencia de un punto en la membrana basilar. La dependencia del nivel de las pendientes se incluye mediante un procedimiento de caso más desfavorable, como se hizo en NMR. Los filtros con frecuencias centrales bajas deben ser calculados a la velocidad de muestreo completa, mientras que los filtros con frecuencias centrales más altas pueden ser calculados a una velocidad de muestreo reducida. Después de los filtros, se calcula un modelo de los efectos temporales del sistema auditivo humano, como se hace en ASD. Después de este paso, es posible una reducción de la velocidad de muestreo en todas las bandas del filtro. Esto conduce a una resolución temporal del banco de filtros de 0,66 ms a una velocidad de muestreo de 48 kHz. Las salidas de los filtros concordantes de la señal de referencia y de la señal sometida a prueba se comparan con una función de probabilidad de detección. Esta función utiliza la sonoridad de las señales de entrada como dato de entrada para calcular la JNLD. La probabilidad total de detección se obtiene a partir de la probabilidad de detección de cada banda. Esta operación se efectúa para ambos canales de entrada y también para el denominado canal central. La probabilidad de detección en el canal central para cada banda es el peor caso de la probabilidad de detección del canal izquierdo y del canal derecho. Para cada trama de 0,66 ms se calcula también la suma de los pasos por encima del umbral.

Se utilizan varias formas de promediado temporal de la probabilidad de detección y de los pasos por encima del umbral:

– el promediado temporal de la probabilidad de detección;
– la frecuencia de tramas con una probabilidad de detección superior a 0,5;
– el máximo de una probabilidad de detección filtrada en paso bajo;

– el máximo de una probabilidad de detección filtrada en paso bajo, sin memoria;
– número medio de pasos por encima del umbral para tramas de una probabilidad de detección superior a 0,5;
– el número promedio de pasos por encima del umbral;
– el número máximo de pasos por encima del umbral;
– el promedio del número de pasos por encima del umbral del 10% de las tramas peores.

### 4.4 PAQM (*perceptual audio quality measure*)

El principio básico de PAQM [Beerends y Stemerdink, 1992] consiste en restar las representaciones internas (representaciones dentro de la cabeza del sujeto) de la señal de referencia y de la señal degradada, y hacer corresponder su diferencia con una correspondencia cognoscitiva de la calidad de audio percibida subjetivamente. La transformación del dominio físico, externo, al dominio sicofísico, interno, se efectúa mediante cuatro operaciones:

– una correspondencia de tiempo-frecuencia que se efectúa mediante una DFT con una ventana Hann de aproximadamente de 40 ms de duración;
– deformación de frecuencia (*frequency warping*) mediante la escala Bark;
– expansión en tiempo-frecuencia (convolución no lineal);
– deformación de intensidad (*intensity warping*) (compresión).

La combinación de dispersión (*smearing*) y compresión permite modelar el comportamiento enmascarador del sistema auditivo humano en el umbral de enmascaramiento y por encima de este umbral. La optimización de la compresión se obtiene utilizando resultados subjetivos de la primera evaluación con códec de audio MPEG (ISO/IEC JTC 1/SC 2/WG 11 MPEG/Audio test report, Document MPEG90/N0030, octubre 1990) (ISO/IEC JTC 1/SC 2/WG 11 MPEG/Audio test report, Document MPEG91/N0010, junio 1991). La diferencia en representación interna se expresa en forma de perturbación debida al ruido. En las últimas versiones de PAQM, presentadas al UIT-R GTE 10/4, se incluyeron dos efectos cognoscitivos en la correspondencia de la perturbación por ruido a la calidad subjetiva: el «streaming» perceptual [Beerends y Stemerdink, 1994] y el enmascaramiento informativo [Beerends y otros, 1996].

Una versión simplificada de la PAQM, la medida de la calidad vocal perceptual (PSQM, *perceptual speech quality measure*) [Beerends y Stemerdink, 1994] se desarrolló utilizando un modelo cognoscitivo como el presentado en [Beerends y Stemerdink, 1994], pero ampliado con una ponderación de los intervalos de silencio. Durante el desarrollo de PSQM pudo observarse que al juzgar la calidad vocal en un contexto de telefonía, el ruido que se produce durante los intervalos de silencio tiene menos importancia que el ruido que se produce durante intervalos de actividad de la voz. En pruebas realizadas por el UIT-T, la propuesta PSQM mostró la correlación más alta entre la calidad objetiva y subjetiva (Comisión de Estudio 12 de Normalización de las Telecomunicaciones, contribución COM 12-74 – Review of validation tests for objective speech quality measures). En la Recomendación UIT-T P862 se especifica la Evaluación perceptual de la calidad vocal, que contiene un algoritmo mejorado para la evaluación de la calidad vocal.

### 4.5 PERCEVAL

PERCEVAL (evaluación perceptual, *PERCeptual EVALuation*) [Paillard y otros, 1992] modela las características de transferencia del oído medio y del oído interno para formar una representación interna de la señal. La señal de entrada se descompone en una representación de tiempo-frecuencia utilizando una DFT. Por lo general, se aplica una ventana Hann de aproximadamente 40 ms a los datos de entrada, con una superposición del 50% entre ventanas sucesivas. El espectro de energía se multiplica por una función dependiente de la frecuencia que modela el efecto del canal auditivo y del

oído interno. Los valores de energía espectral atenuados se convierten de la escala de frecuencia a una escala de alturas del sonido que es más lineal con respecto a las propiedades físicas del oído interno y también a los efectos sicofísicos observados. Los componentes de energía transformados son convolucionados con una función de expansión para simular la dispersión de energía a lo largo de la membrana basilar. Por último, se añade, a cada componente de altura del sonido, una energía intrínseca dependiente de la frecuencia, para tener en cuenta el umbral absoluto de audibilidad. Una conversión de la energía a decibelios proporciona una representación de señal en la membrana basilar.

En simulaciones de experimentos de enmascaramiento auditivo, se forma una representación en membrana basilar para cada estímulo, y la diferencia entre las representaciones es la información disponible para la realización de la tarea. Una representación es la de la señal enmascaradora sola, y la otra es la de la señal enmascaradora y la señal de prueba combinadas. Su diferencia representa el componente de la señal que no está enmascarado. PERCEVAL calcula la probabilidad de detectar esta diferencia. La probabilidad de no detección de esta diferencia, para cada detector en la totalidad de la membrana basilar simulada se estima utilizando una función de probabilidad sigmoidal. Suponiendo que los detectores son estadísticamente independientes, la probabilidad de detección global, para el conjunto completo de detectores, se calcula como el complemento del producto de las probabilidades de no detección individuales. Se realizaron con éxito varios experimentos de enmascaramiento, que se simularon mediante este procedimiento, y el modelo se utilizó para evaluar las posibilidades de modelar oyentes individuales [Treurniet, 1996].

Como una herramienta para estimar la calidad de audio, PERCEVAL calcula la diferencia entre las representaciones de la señal de referencia y de la señal sometida a prueba. Utilizando hipótesis razonables sobre los procesos perceptuales y cognoscitivos de alto nivel, cierto número de variables de interés en materia de percepción se calcularon y se hicieron corresponder a una medida de la calidad objetiva de la señal sometida a prueba. La correspondencia se optimizó minimizando la diferencia entre la distribución de la calidad objetiva y la correspondiente distribución de las evaluaciones subjetivas medias de la calidad, para el conjunto de datos disponible.

### 4.6 POM

La medición objetiva perceptual (POM, *perceptual objective measurement*) [Colomes y otros, 1995] tiene por finalidad cuantificar una determinada degradación que puede producirse entre una señal de referencia y su versión «degradada». Esto se efectúa comparando la representación interna, en membrana basilar, de ambas señales, cualquiera que sea el agente que produjo la degradación. La representación basilar modela los diferentes procesos sufridos por una señal de audio cuando atraviesa el oído humano. Por consiguiente, la primera etapa de POM es el cálculo de la representación interna de una señal de audio. El patrón de excitación (dado en dB), ensanchado sobre la membrana basilar, ha sido elegido de modo que modele la tasa de disparo de las neuronas a lo largo de la membrana basilar.

El proceso de calcular el patrón de excitación se denomina el oído artificial. Seguidamente, una vez que se ha conseguido comparar, una con otra, las dos representaciones internas de las señales, la POM tiene que señalar si la diferencia entre sus representaciones internas es audible o no, y, si lo es, de qué manera. Este proceso se denomina el proceso de detección.

La POM utiliza una DFT con una ventana Hann con una duración de aproximadamente 40 ms (con una superposición del 50% entre las dos ventanas Hann). El número de canales basilares de análisis es 620. Las partes restantes del modelo auditivo son casi idénticas a las utilizadas en PAQM o en PERCEVAL.

La función de expansión se describe con una buena exactitud por una aproximación más precisa que tiene en cuenta tanto la dependencia con respecto al nivel, de acuerdo con [Terhardt, 1979] como la forma redondeada, de acuerdo con [Schroeder y otros, 1979].

Este modelo presenta la probabilidad de detectar una distorsión entre las dos señales comparadas, así como una denominada distancia basilar que representa la diferencia perceptual entre las dos excitaciones comparadas.

### 4.7 The Toolbox approach (caja de herramientas)

Toolbox utiliza un procedimiento de tres pasos para medir la distancia percibida en la calidad de audio de una señal de prueba en relación con la de una señal de referencia, dando así una indicación del nivel de calidad de audio subjetiva de la señal de prueba. El método se basa en modelos perceptuales bien conocidos que se utilizan para describir la representación perceptual de las diferencias entre las dos señales de audio. Además, incluye un procedimiento de ponderación para la calidad de audio percibida de una señal de prueba estereofónica, teniendo en cuenta los resultados del canal izquierdo y del canal derecho. No se requiere una correlación rígida, muestra por muestra, de la señal de referencia y de la señal de audio sometida a prueba.

La principal funcionalidad de Toolbox, paso 1, se basa en el cálculo de la sonoridad específica, que se efectúa de acuerdo con [Zwicker y Feldtkeller, 1967], utilizando una FFT de 2 048 puntos, tratada por una ventana Hann, que corresponde aproximadamente a una duración de 40 ms. La totalidad de la ventana se desplaza en incrementos de 10 ms. Además, se aplican efectos de enmascaramiento temporal, tales como el posenmascaramiento y el preenmascaramiento, de acuerdo con Zwicker. A partir de estos valores básicos de sensación se calculan otros parámetros perceptuales, tales como la sonoridad integrada, sonoridad parcialmente enmascarada, nitidez, de acuerdo con [von Bismarck, 1974] y [Aures, 1984], y la cantidad de preecos se calcula como un resultado de una etapa de preprocesamiento para los pasos siguientes.

El segundo paso de Toolbox incluye procedimientos de ponderación que dependen principalmente de la magnitud de la diferencia percibida en sonoridad y de la variación de la sonoridad en función del tiempo.

El tercer paso de Toolbox incluye la generación de un conjunto de valores de salida Toolbox intermedios que se basan en un análisis estadístico de los valores obtenidos en los pasos 1 y 2. El resultado de este análisis estadístico incluye los valores medio y máximo cuadrático, así como la desviación típica de los valores medios. Se utiliza una suma ponderada de estos valores de salida Toolbox intermedios para el ajuste final de la distancia perceptual entre la señal sometida a prueba y la señal de referencia. Si es necesario, este valor de salida único puede cotejarse con un grado de diferencia subjetiva (SDG), obtenido usualmente en pruebas de escucha subjetivas, efectuando un ajuste secuencial de los datos de salida para cada incremento de tiempo, mediante el empleo de una función lineal o de una función polinómica de orden más alto.

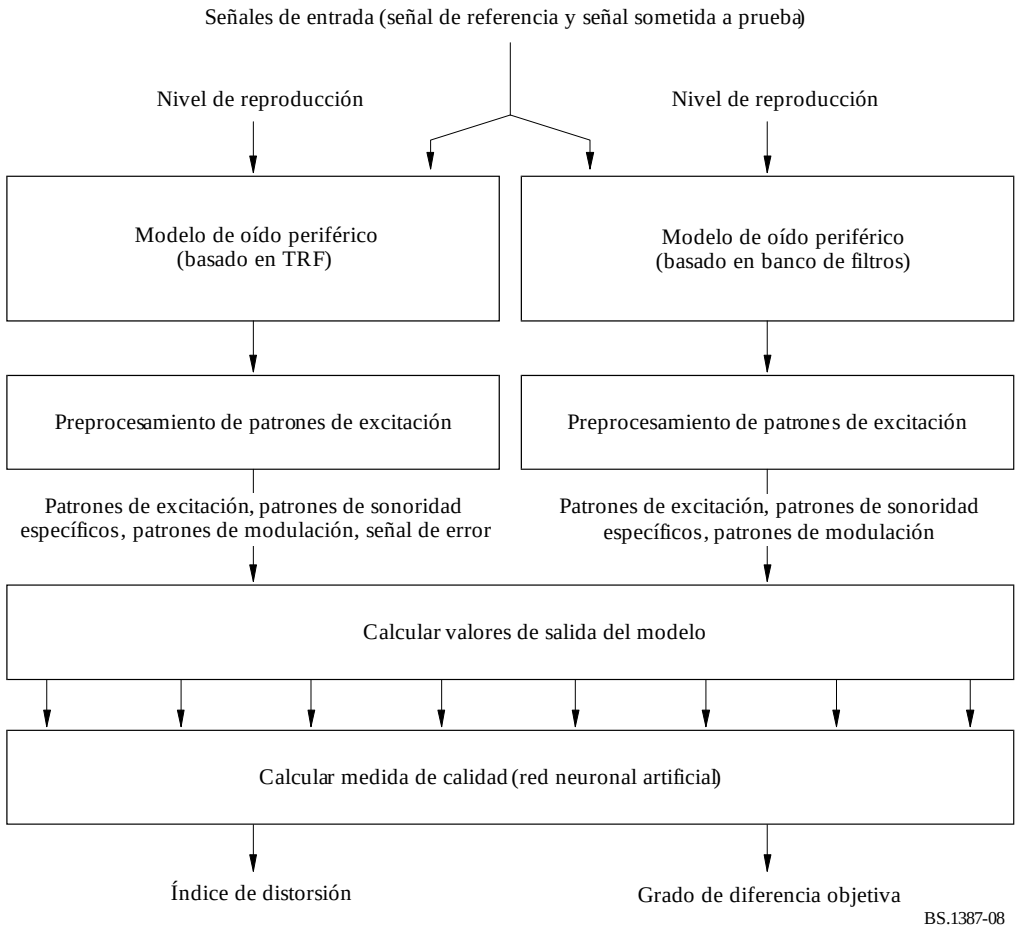
# Anexo 2

# Descripción del modelo[2]

## 1 Descripción general

FIGURA 8

**Diagrama de principio del esquema de medición**




El *método para la medición objetiva de la calidad de audio percibida* propuesto consiste en un modelo de oído periférico, varios pasos intermedios (a los que se hace referencia en este Anexo como «*preprocesamiento de patrones de excitación*», el cálculo de las *variables de salida del modelo* («MOV», *model output variables*) basadas (en su mayor parte) en nociones sicoacústicas, y una correspondencia de un conjunto de variables de salida del modelo a un valor único que representa la *calidad de audio básica* de la señal sometida a prueba. Incluye dos modelos de oído periféricos, uno

[2] Los proponentes de la tecnología descrita en esta Recomendación han presentado declaraciones de patentes conformes al Anexo 1 de la Resolución UIT-R 1. La tecnología descrita en esta Recomendación está protegida por patentes internacionales y, como todas las Recomendaciones de la UIT, están sujetas a los derechos de autor. Para explotar esta tecnología es indispensable el previo consentimiento de los propietarios, en forma de una licencia. Para obtener más información en materia de licencias sobre esta tecnología, véase la base de datos del UIT-R sobre patentes, o diríjase a la Secretaría de BR.

basado en la TRF y el otro basado en un banco de filtros. Salvo el cálculo de la señal de error (que sólo se utiliza con la parte del modelo basada en una TRF), la estructura general es la misma para ambos modelos de oído periféricos.

Las entradas para el cálculo de las MOV son:

– Los patrones de excitación para la señal de prueba y la señal de referencia.
– Los patrones de excitación adaptados espectralmente para la señal de prueba y la señal de referencia.
– Los patrones de sonoridad específicos para la señal de prueba y la señal de referencia.
– Los patrones de modulación para la señal de prueba y la señal de referencia.
– La señal de error calculada como la diferencia espectral entre la señal de prueba y la señal de referencia (sólo para el modelo de oído basado en TRF).

Si no se indica otra cosa, en el caso de señales estereofónicas todas los cálculos se realizan independientemente y de la misma manera para el canal izquierdo y el derecho.

En la descripción se definen dos configuraciones: una denominada la «*versión básica*» y otra denominada la «*versión avanzada*».

En todas las ecuaciones presentadas, el índice «Ref.» representa todos los patrones calculados a partir de la señal de referencia, el índice «Test» representa todos los patrones calculados a partir de la señal sometida a prueba. El Índice «*k*» representa la variable de frecuencia discreta (por ejemplo, la banda de frecuencias) y «*n*» representa la variable de tiempo discreta (por ejemplo, el contador de tramas o el contador de muestras). Si los valores de *k* o *n* no están explícitamente definidos, los cálculos deberán efectuarse para todos los valores posibles de *k* y *n*. Todas las demás abreviaturas se explican en el lugar en que aparecen.

En los nombres de variables de salida del modelo, el índice «*A*» representa todas las variables calculadas utilizando la parte del modelo de oído basada en un banco de filtros y el índice «*B*» representa las variables calculadas usando la parte del modelo de oído basada en una TRF.

### 1.1 Versión básica

La *versión básica* incluye solamente MOV que son calculadas a partir del modelo de oído basado en TRF. La parte del modelo basada en un banco de filtros no se utiliza. La *versión básica* emplea un total de 11 MOV para la predicción de la *calidad audio básica* percibida.

### 1.2 Versión avanzada

La *versión avanzada* incluye MOV que son calculadas a partir del modelo de oído basado en banco de filtros, así como MOV que son calculadas a partir del modelo de oído basado en TRF. Los patrones de excitación adaptados espectralmente y los patrones de modulación se calculan a partir de la parte del modelo basada en un banco de filtros, solamente. La *versión avanzada* emplea un total de 5 MOV para la predicción de la *calidad de audio básica* percibida.

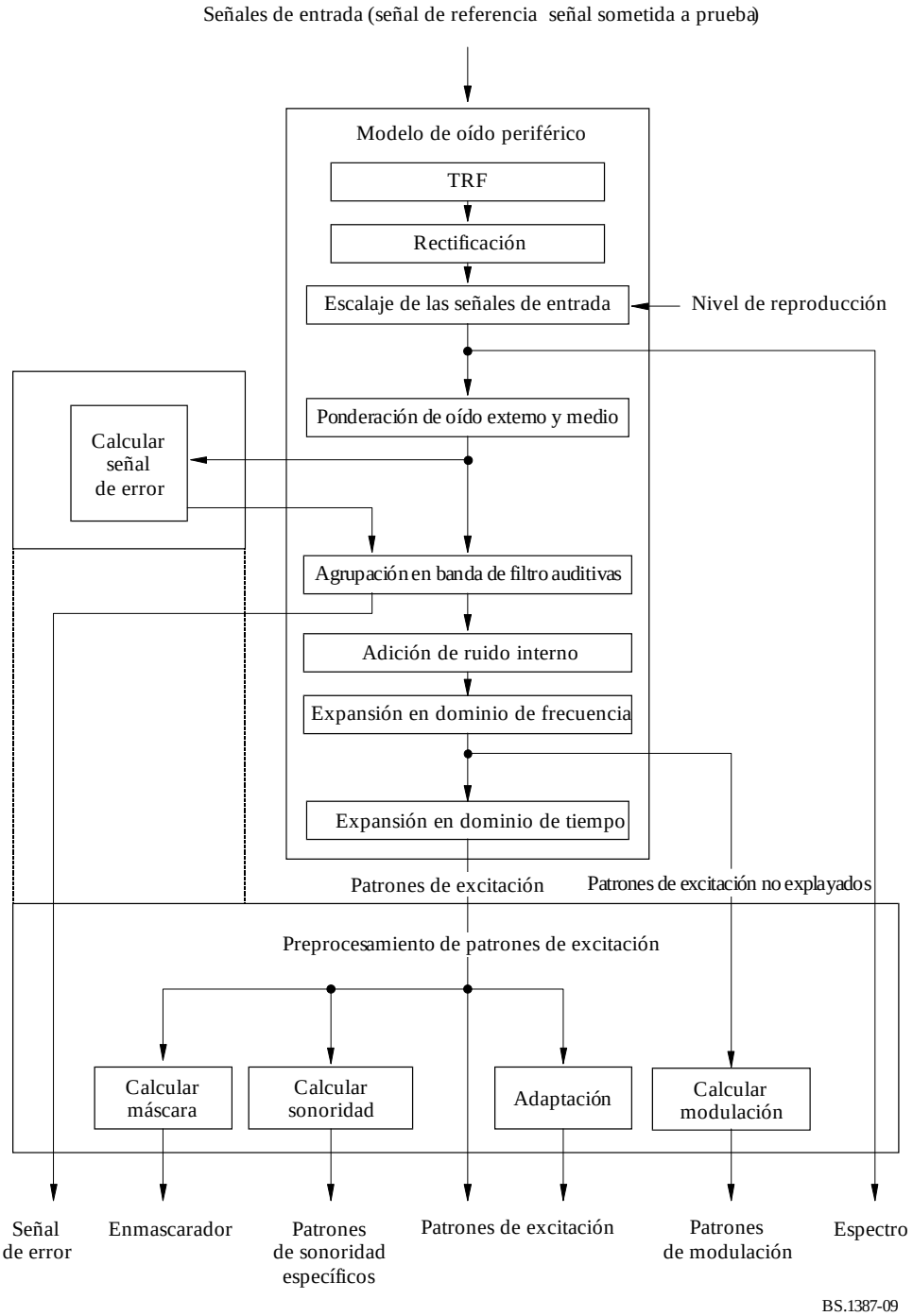
# 2 Modelo de oído periférico

## 2.1 Modelo de oído basado en TRF

### 2.1.1 Visión de conjunto

FIGURA 9

**Modelo de oído periférico y preprocesamiento de los patrones de excitación para la parte del modelo basada en una TRF**




La entrada del modelo de oído basado en TRF, constituida por la señal de referencia y la señal de prueba, alineadas en el tiempo y muestreadas a 48 kHz, se divide en tramas de aproximadamente

0,042 s con una superposición de 50%. Cada trama es transformada al dominio de la frecuencia utilizando una ventana Hann y una TRF a corto plazo, y escalada al nivel de reproducción. A los coeficientes espectrales se les aplica una función de ponderación que modela la respuesta a frecuencia del oído externo y del oído medio. La transformación a la representación en altura del sonidos (*pitch*) se efectúa agrupando los coeficientes espectrales ponderados en bandas críticas. Se añade un desplazamiento dependiente de la frecuencia para simular el ruido interno en el sistema auditivo. Se utiliza una función de expansión dependiente del nivel para modelar los filtros auditivos espectrales en el dominio de la frecuencia. A esto sigue una expansión en el dominio del tiempo que tiene en cuenta los efectos de enmascaramiento hacia adelante.

Los *patrones de excitación* así obtenidos se utilizan para calcular *patrones de sonoridad específicos* y los *patrones de enmascaramiento*. Los patrones obtenidos antes de la última expansión en el dominio del tiempo («*patrones de excitación no explayados*») se utilizan para calcular *patrones de modulación*.

Para modelar la señal de error, los patrones de la señal de referencia y de la señal de prueba de la salida del filtro de oído externo y de oído medio son combinados y hechos corresponder con la escala de alturas del sonido agrupándolos en bandas críticas.

Estas salidas se utilizan junto con los patrones de excitación para calcular los valores de las variables de salida del modelo.

### 2.1.2 Procesamiento en el tiempo

La entrada del modelo de oído basado en TRF, señal de prueba y señal de referencia, se divide en tramas de 2 048 muestras con una superposición de 1 024 muestras:

$$t_n[k_t, n] = t[1\,024 \cdot n + k_t] \quad n = 0, 1, 2 \ldots k_t = 0..2\,047 \tag{1}$$

donde $n$ es el número de la trama de tiempo y $k_t$ es un contador de tiempo dentro de una trama.

### 2.1.3 TRF

La conversión del dominio del tiempo al dominio de la frecuencia se efectúa por medio de una ventana Hann:

$$h_w[k] = \frac{1}{2}\sqrt{\frac{8}{3}}\left[1 - \cos\left(2\pi\frac{k}{N-1}\right)\right] \quad \Big| \quad N = 2\,048 \tag{2}$$

$$t_w[k_t, n] = h_w[k_t] \cdot t_n[k_t, n] \tag{3}$$

seguida de una transformada de Fourier a corto plazo:

$$F_f[k_f, n] = \frac{1}{2\,048}\sum_{k_t=0}^{2\,047} t_w[k_t, n]\, \mathrm{e}^{-j\frac{2\pi}{2\,048}k_f k_t} \tag{4}$$

El factor de escalaje para la TRF se calcula a partir del nivel de presión sonora *Lp* supuesto de una onda sinusoidal a plena escala, por las fórmulas:

$$fac = \frac{10^{\frac{Lp}{20}}}{Norm} \tag{5}$$

$$F\,[k_f, n] = fac \cdot F_f\,[k_f, n] \tag{6}$$

donde el factor de normalización *Norm* se calcula tomando una onda sinusoidal de 1 019,5 Hz y 0 dB a plena escala como la señal de entrada y calculando el valor absoluto máximo de los coeficientes espectrales en 10 tramas.

Si el nivel de presión sonora es desconocido, se recomienda dar a *Lp* un valor de 92 $dB_{SPL}$.

### 2.1.4 Oído externo y oído medio

La respuesta a frecuencia del oído externo y del oído medio se modela por una función de ponderación dependiente de la frecuencia:

$$W[k]/\text{dB} = -0{,}6 \cdot 3{,}64 \cdot \left(\frac{f[k]}{\text{kHz}}\right)^{-0{,}8} + 6{,}5 \cdot \text{e}^{-0{,}6 \cdot \left(\frac{f[k]}{\text{kHz}} - 3{,}3\right)^2} - 10^{-3} \cdot \left(\frac{f[k]}{\text{kHz}}\right)^{3{,}6} \tag{7}$$

donde:

$$f[k]/\text{Hz} = k \cdot 23{,}4375 \tag{8}$$

es la representación en frecuencia en la línea *k* que se aplica a las salidas TRF (ecuación (9)).

$$F_e[k_f, n] = \left| F[k_f, n] \right| \cdot 10^{\frac{W[k_f]}{20}} \tag{9}$$

$F_e[k_f]$ se designan como las «*salidas TRF ponderadas por el oído externo*».

### 2.1.5 Agrupación en bandas críticas

La escala auditiva de alturas del sonido se calcula utilizando una formula aproximada presentada por [Schroeder y otros, 1979].

$$z/Bark = 7 \cdot \text{arsenh}\left(\frac{f/\text{Hz}}{650}\right) \tag{10}$$

Las unidades de altura del sonido (*pitch*) se denominan *Bark* (debe señalarse, no obstante, que esta escala no representa exactamente la escala Bark definida por [Zwicker y Feldtkeller, 1967]).

Las frecuencias en los bordes de las bandas de frecuencias de los filtros están comprendidas entre 80 Hz y 18 000 Hz. Las anchuras y el espaciamiento de las bandas de filtro corresponden a una resolución de *res = 0,25* Bark para la versión básica y *res = 0,5* Bark para la versión avanzada.

Esto conduce a un número de bandas de frecuencia *Z = 109* para la versión básica y *Z = 55* para la versión avanzada.

CUADRO 6

**Bandas de frecuencias del modelo de oído basado en TRF utilizado en la versión básica**

| **Grupo** | **Frecuencia inferior/Hz** | **Frecuencia central/Hz** | **Frecuencia superior/Hz** | **Anchura en frecuencia/Hz** |
|---|---|---|---|---|
| *k* | $f_l[k]$ | $f_c[k]$ | $f_u[k]$ | $f_w[k]$ |
| 0 | 80 | 91,708 | 103,445 | 23,445 |
| 1 | 103,445 | 115,216 | 127,023 | 23,577 |
| 2 | 127,023 | 138,87 | 150,762 | 23,739 |
| 3 | 150,762 | 162,702 | 174,694 | 23,932 |
| 4 | 174,694 | 186,742 | 198,849 | 24,155 |
| 5 | 198,849 | 211,019 | 223,257 | 24,408 |
| 6 | 223,257 | 235,566 | 247,95 | 24,693 |
| 7 | 247,95 | 260,413 | 272,959 | 25,009 |
| 8 | 272,959 | 285,593 | 298,317 | 25,358 |
| 9 | 298,317 | 311,136 | 324,055 | 25,738 |
| 10 | 324,055 | 337,077 | 350,207 | 26,151 |
| 11 | 350,207 | 363,448 | 376,805 | 26,598 |
| 12 | 376,805 | 390,282 | 403,884 | 27,079 |
| 13 | 403,884 | 417,614 | 431,478 | 27,594 |
| 14 | 431,478 | 445,479 | 459,622 | 28,145 |
| 15 | 459,622 | 473,912 | 488,353 | 28,731 |
| 16 | 488,353 | 502,95 | 517,707 | 29,354 |
| 17 | 517,707 | 532,629 | 547,721 | 30,014 |
| 18 | 547,721 | 562,988 | 578,434 | 30,713 |
| 19 | 578,434 | 594,065 | 609,885 | 31,451 |
| 20 | 609,885 | 625,899 | 642,114 | 32,229 |
| 21 | 642,114 | 658,533 | 675,161 | 33,048 |
| 22 | 675,161 | 692,006 | 709,071 | 33,909 |
| 23 | 709,071 | 726,362 | 743,884 | 34,814 |
| 24 | 743,884 | 761,644 | 779,647 | 35,763 |
| 25 | 779,647 | 797,898 | 816,404 | 36,757 |
| 26 | 816,404 | 835,17 | 854,203 | 37,799 |
| 27 | 854,203 | 873,508 | 893,091 | 38,888 |
| 28 | 893,091 | 912,959 | 933,119 | 40,028 |
| 29 | 933,119 | 953,576 | 974,336 | 41,218 |
| 30 | 974,336 | 995,408 | 1 016,797 | 42,461 |
| 31 | 1 016,797 | 1 038,511 | 1 060,555 | 43,758 |
| 32 | 1 060,555 | 1 082,938 | 1 105,666 | 45,111 |

CUADRO 6 (*continuación*)

| **Grupo** | **Frecuencia inferior/Hz** | **Frecuencia central/Hz** | **Frecuencia superior/Hz** | **Anchura en frecuencia/Hz** |
|---|---|---|---|---|
| *k* | $f_l[k]$ | $f_c[k]$ | $f_u[k]$ | $f_w[k]$ |
| 33 | 1 105,666 | 1 128,746 | 1 152,187 | 46,521 |
| 34 | 1 152,187 | 1 175,995 | 1 200,178 | 47,991 |
| 35 | 1 200,178 | 1 224,744 | 1 249,7 | 49,522 |
| 36 | 1 249,7 | 1 275,055 | 1 300,816 | 51,116 |
| 37 | 1 300,816 | 1 326,992 | 1 353,592 | 52,776 |
| 38 | 1 353,592 | 1 380,623 | 1 408,094 | 54,502 |
| 39 | 1 408,094 | 1 436,014 | 1 464,392 | 56,298 |
| 40 | 1 464,392 | 1 493,237 | 1 522,559 | 58,167 |
| 41 | 1 522,559 | 1 552,366 | 1 582,668 | 60,109 |
| 42 | 1 582,668 | 1 613,474 | 1 644,795 | 62,128 |
| 43 | 1 644,795 | 1 676,641 | 1 709,021 | 64,226 |
| 44 | 1 709,021 | 1 741,946 | 1 775,427 | 66,406 |
| 45 | 1 775,427 | 1 809,474 | 1 844,098 | 68,671 |
| 46 | 1 844,098 | 1 879,31 | 1 915,121 | 71,023 |
| 47 | 1 915,121 | 1 951,543 | 1 988,587 | 73,466 |
| 48 | 1 988,587 | 2 026,266 | 2 064,59 | 76,003 |
| 49 | 2 064,59 | 2 103,573 | 2 143,227 | 78,637 |
| 50 | 2 143,227 | 2 183,564 | 2 224,597 | 81,371 |
| 51 | 2 224,597 | 2 266,34 | 2 308,806 | 84,208 |
| 52 | 2 308,806 | 2 352,008 | 2 395,959 | 87,154 |
| 53 | 2 395,959 | 2 440,675 | 2 486,169 | 90,21 |
| 54 | 2 486,169 | 2 532,456 | 2 579,551 | 93,382 |
| 55 | 2 579,551 | 2 627,468 | 2 676,223 | 96,672 |
| 56 | 2 676,223 | 2 725,832 | 2 776,309 | 100,086 |
| 57 | 2 776,309 | 2 827,672 | 2 879,937 | 103,627 |
| 58 | 2 879,937 | 2 933,12 | 2 987,238 | 107,302 |
| 59 | 2 987,238 | 3 042,309 | 3 098,35 | 111,112 |
| 60 | 3 098,35 | 3 155,379 | 3 213,415 | 115,065 |
| 61 | 3 213,415 | 3 272,475 | 3 332,579 | 119,164 |
| 62 | 3 332,579 | 3 393,745 | 3 455,993 | 123,415 |
| 63 | 3 455,993 | 3 519,344 | 3 583,817 | 127,823 |
| 64 | 3 583,817 | 3 649,432 | 3 716,212 | 132,395 |
| 65 | 3 716,212 | 3 784,176 | 3 853,348 | 137,136 |
| 66 | 3 853,348 | 3 923,748 | 3 995,399 | 142,051 |
| 67 | 3 995,399 | 4 068,324 | 4 142,547 | 147,148 |
| 68 | 4 142,547 | 4 218,09 | 4 294,979 | 152,432 |

CUADRO 6 (*continuación*)

| **Grupo** | **Frecuencia inferior/Hz** | **Frecuencia central/Hz** | **Frecuencia superior/Hz** | **Anchura en frecuencia/Hz** |
|---|---|---|---|---|
| *k* | $f_l[k]$ | $f_c[k]$ | $f_u[k]$ | $f_w[k]$ |
| 69 | 4 294,979 | 4 373,237 | 4 452,89 | 157,911 |
| 70 | 4 452,89 | 4 533,963 | 4 616,482 | 163,592 |
| 71 | 4 616,482 | 4 700,473 | 4 785,962 | 169,48 |
| 72 | 4 785,962 | 4 872,978 | 4 961,548 | 175,585 |
| 73 | 4 961,548 | 5 051,7 | 5 143,463 | 181,915 |
| 74 | 5 143,463 | 5 236,866 | 5 331,939 | 188,476 |
| 75 | 5 331,939 | 5 428,712 | 5 527,217 | 195,278 |
| 76 | 5 527,217 | 5 627,484 | 5 729,545 | 202,329 |
| 77 | 5 729,545 | 5 833,434 | 5 939,183 | 209,637 |
| 78 | 5 939,183 | 6 046,825 | 6 156,396 | 217,214 |
| 79 | 6 156,396 | 6 267,931 | 6 381,463 | 225,067 |
| 80 | 6 381,463 | 6 497,031 | 6 614,671 | 233,208 |
| 81 | 6 614,671 | 6 734,42 | 6 856,316 | 241,646 |
| 82 | 6 856,316 | 6 980,399 | 7 106,708 | 250,392 |
| 83 | 7 106,708 | 7 235,284 | 7 366,166 | 259,458 |
| 84 | 7 366,166 | 7 499,397 | 7 635,02 | 268,854 |
| 85 | 7 635,02 | 7 773,077 | 7 913,614 | 278,594 |
| 86 | 7 913,614 | 8 056,673 | 8 202,302 | 288,688 |
| 87 | 8 202,302 | 8 350,547 | 8 501,454 | 299,152 |
| 88 | 8 501,454 | 8 655,072 | 8 811,45 | 309,996 |
| 89 | 8 811,45 | 8 970,639 | 9 132,688 | 321,237 |
| 90 | 9 132,688 | 9 297,648 | 9 465,574 | 332,887 |
| 91 | 9 465,574 | 9 636,52 | 9 810,536 | 344,962 |
| 92 | 9 810,536 | 9 987,683 | 10 168,013 | 357,477 |
| 93 | 10 168,013 | 10 351,586 | 10 538,46 | 370,447 |
| 94 | 10 538,46 | 10 728,695 | 10 922,351 | 383,891 |
| 95 | 10 922,351 | 11 119,49 | 11 320,175 | 397,824 |
| 96 | 11 320,175 | 11 524,47 | 11 732,438 | 412,264 |
| 97 | 11 732,438 | 11 944,149 | 12 159,67 | 427,231 |
| 98 | 12 159,67 | 12 379,066 | 12 602,412 | 442,742 |
| 99 | 12 602,412 | 12 829,775 | 13 061,229 | 458,817 |
| 100 | 13 061,229 | 13 296,85 | 13 536,71 | 475,48 |
| 101 | 13 536,71 | 13 780,887 | 14 029,458 | 492,748 |
| 102 | 14 029,458 | 14 282,503 | 14 540,103 | 510,645 |
| 103 | 14 540,103 | 14 802,338 | 15 069,295 | 529,192 |
| 104 | 15 069,295 | 15 341,057 | 15 617,71 | 548,415 |

CUADRO 6 (*fin*)

| Grupo | Frecuencia inferior/Hz | Frecuencia central/Hz | Frecuencia superior/Hz | Anchura en frecuencia/Hz |
|---|---|---|---|---|
| *k* | $f_l[k]$ | $f_c[k]$ | $f_u[k]$ | $f_w[k]$ |
| 105 | 15 617,71 | 15 899,345 | 16 186,049 | 568,339 |
| 106 | 16 186,049 | 16 477,914 | 16 775,035 | 588,986 |
| 107 | 16 775,035 | 17 077,504 | 17 385,42 | 610,385 |
| 108 | 17 385,42 | 17 690,045 | 18 000 | 614,58 |

CUADRO 7

**Bandas de frecuencias del modelo de oído basado en TRF utilizado en la versión avanzada**

| Grupo | Frecuencia inferior/Hz | Frecuencia central/Hz | Frecuencia superior/Hz | Anchura en frecuencia/Hz |
|---|---|---|---|---|
| *k* | $f_l[k]$ | $f_c[k]$ | $f_u[k]$ | $f_w[k]$ |
| 0 | 80 | 103,445 | 127,023 | 47,023 |
| 1 | 127,023 | 150,762 | 174,694 | 47,671 |
| 2 | 174,694 | 198,849 | 223,257 | 48,563 |
| 3 | 223,257 | 247,95 | 272,959 | 49,702 |
| 4 | 272,959 | 298,317 | 324,055 | 51,096 |
| 5 | 324,055 | 350,207 | 376,805 | 52,75 |
| 6 | 376,805 | 403,884 | 431,478 | 54,673 |
| 7 | 431,478 | 459,622 | 488,353 | 56,875 |
| 8 | 488,353 | 517,707 | 547,721 | 59,368 |
| 9 | 547,721 | 578,434 | 609,885 | 62,164 |
| 10 | 609,885 | 642,114 | 675,161 | 65,277 |
| 11 | 675,161 | 709,071 | 743,884 | 68,723 |
| 12 | 743,884 | 779,647 | 816,404 | 72,52 |
| 13 | 816,404 | 854,203 | 893,091 | 76,687 |
| 14 | 893,091 | 933,119 | 974,336 | 81,245 |
| 15 | 974,336 | 1 016,797 | 1 060,555 | 86,219 |
| 16 | 1 060,555 | 1 105,666 | 1 152,187 | 91,632 |
| 17 | 1 152,187 | 1 200,178 | 1 249,7 | 97,513 |
| 18 | 1 249,7 | 1 300,816 | 1 353,592 | 103,892 |
| 19 | 1 353,592 | 1 408,094 | 1 464,392 | 110,801 |
| 20 | 1 464,392 | 1 522,559 | 1 582,668 | 118,275 |
| 21 | 1 582,668 | 1 644,795 | 1 709,021 | 126,354 |
| 22 | 1 709,021 | 1 775,427 | 1 844,098 | 135,077 |
| 23 | 1 844,098 | 1 915,121 | 1 988,587 | 144,489 |

CUADRO 7 (*fin*)

| **Grupo** | **Frecuencia inferior/Hz** | **Frecuencia central/Hz** | **Frecuencia superior/Hz** | **Anchura en frecuencia/Hz** |
|---|---|---|---|---|
| *k* | $f_l[k]$ | $f_c[k]$ | $f_u[k]$ | $f_w[k]$ |
| 24 | 1 988,587 | 2 064,59 | 2 143,227 | 154,64 |
| 25 | 2 143,227 | 2 224,597 | 2 308,806 | 165,579 |
| 26 | 2 308,806 | 2 395,959 | 2 486,169 | 177,364 |
| 27 | 2 486,169 | 2 579,551 | 2 676,223 | 190,054 |
| 28 | 2 676,223 | 2 776,309 | 2 879,937 | 203,713 |
| 29 | 2 879,937 | 2 987,238 | 3 098,35 | 218,414 |
| 30 | 3 098,35 | 3 213,415 | 3 332,579 | 234,229 |
| 31 | 3 332,579 | 3 455,993 | 3 583,817 | 251,238 |
| 32 | 3 583,817 | 3 716,212 | 3 853,348 | 269,531 |
| 33 | 3 853,348 | 3 995,399 | 4 142,547 | 289,199 |
| 34 | 4 142,547 | 4 294,979 | 4 452,89 | 310,343 |
| 35 | 4 452,89 | 4 616,482 | 4 785,962 | 333,072 |
| 36 | 4 785,962 | 4 961,548 | 5 143,463 | 357,5 |
| 37 | 5 143,463 | 5 331,939 | 5 527,217 | 383,754 |
| 38 | 5 527,217 | 5 729,545 | 5 939,183 | 411,966 |
| 39 | 5 939,183 | 6 156,396 | 6 381,463 | 442,281 |
| 40 | 6 381,463 | 6 614,671 | 6 856,316 | 474,853 |
| 41 | 6 856,316 | 7 106,708 | 7 366,166 | 509,85 |
| 42 | 7 366,166 | 7 635,02 | 7 913,614 | 547,448 |
| 43 | 7 913,614 | 8 202,302 | 8 501,454 | 587,84 |
| 44 | 8 501,454 | 8 811,45 | 9 132,688 | 631,233 |
| 45 | 9 132,688 | 9 465,574 | 9 810,536 | 677,849 |
| 46 | 9 810,536 | 10 168,013 | 10 538,46 | 727,924 |
| 47 | 10 538,46 | 10 922,351 | 11 320,175 | 781,715 |
| 48 | 11 320,175 | 11 732,438 | 12 159,67 | 839,495 |
| 49 | 12 159,67 | 12 602,412 | 13 061,229 | 901,56 |
| 50 | 13 061,229 | 13 536,71 | 14 029,458 | 968,229 |
| 51 | 14 029,458 | 14 540,103 | 15 069,295 | 1 039,837 |
| 52 | 15 069,295 | 15 617,71 | 16 186,049 | 1 116,754 |
| 53 | 16 186,049 | 16 775,035 | 17 385,42 | 1 199,371 |
| 54 | 17 385,42 | 17 690,045 | 18 000 | 614,58 |

La conversión de frecuencia a altura del sonido (*pitch*) se efectúa por el algoritmo descrito en el punto siguiente, donde $Fsp[k_f]$ es la representación en energía de las «*salidas TRF ponderadas por el oído externo*»:

$$Fsp[k_f, n] = \left| F_e[k_f, n] \right|^2 \tag{11}$$

o la representación en energía de la señal de error:

$$Fsp[k_f, n] = \left| F_{noise}[k_f, n] \right|^2 \quad (12)$$

respectivamente. Véase el § 3.4 para el cálculo de la señal de error.

La salida de esta etapa de procesamiento son las energías de los grupos de frecuencias, $P_e[k,n]$.

#### 2.1.5.1 Seudocódigo

| | | |
|---|---|---|
| /* inputs */ | | |
| Fsp[ ] | : | energías de entrada |
| /* outputs */ | | |
| Pe[ ] | : | energías correspondientes a altura del sonido (*pitch*) |
| /* intermediate values */ | | |
| i | : | índice de grupos de frecuencias |
| k | : | \ índice de línea fft |
| Z | | número de grupos de frecuencias:<br>109 para la versión básica<br>55 para la versión avanzada |
| fl[] | : | frecuencia inferior de grupo de frecuencias |
| fu[] | : | frecuencia superior de grupo de frecuencias |
| Fres | : | constante para resolución de frecuencia |

```
Fres = 48000/2048;
for(i=0; i<Z; i++ )
{
 Pe[i]=0;
 for(k=0;k<1024;k++)
 {
  /* line inside frequency group */
  if( (( k-0.5)*Fres >= fl[i]) && ((k+0.5)*Fres <= fu[i]))
  {
   Pe[i] += Fsp[k];
  }
  /* frequency group inside*/
  else if( (( k-0.5)*Fres < fl[i]) && ((k+0.5)*Fres > fu[i]))
  {
   Pe[i] += Fsp[k]*(fu[i]-fl[i])/Fres;
  }
  /* left border */
  else if( ((k-0.5)*Fres < fl[i]) && ((k+0.5)*Fres > fl[i]))
  {
  Pe[i] += Fsp[k]*( (k+0.5)*Fres - fl[i])/Fres;
  }
  /* right border
  else if( ((k-0.5)*Fres < fu[i]) && ((k+0.5)*Fres > fu[i]);
```

```
  {
  Pe[i] += Fsp[k]*(fu[i]- (k-0.5)*Fres)/Fres;
  }
  /* line outside frequency group */
  else
  {
   Pe[i] += 0;
  }
 }

 /* limit result */
 Pe[i]=max(Pe[i],0.000000000001);
}
```

### 2.1.6 Adición de ruido interno

Se añade un desplazamiento $P_{Thres}$ a las energías en cada grupo de frecuencias:

$$P_{Thres}[k] = 10^{0,4 \cdot 0,364 \cdot \left( \frac{f_c[k]}{\text{kHz}} \right)^{-0,8}} \tag{13}$$

$$P_p[k,n] = P_e[k,n] + P_{Thres}[k] \tag{14}$$

La salida de esta etapa de procesamiento, $P_p[k,n]$ se designa como «*patrones de altura del sonido*» (*pitch patterns*).

### 2.1.7 Expansión (*spreading*)

Los patrones de altura del sonido (*pitch patterns*), $P_p[k,n]$, son explayados en frecuencia utilizando una función de expansión dependiente del nivel. La función de expansión es una exponencial a la derecha y a la izquierda. La pendiente inferior es siempre 27 dB/Bark y la pendiente superior depende de la frecuencia y de la energía.

Las pendientes se calculan de acuerdo con:

$$\frac{S_u[k, L[k,n]]}{\text{dB / } Bark} = -24 - \frac{230 \text{ Hz}}{f_c[k]} + 0,2 \cdot L[k,n] / \text{dB} \tag{15}$$

$$S_l[k, L[k,n]] = 27 \frac{\text{dB}}{Bark} \tag{16}$$

con:

$$L[k,n] = 10 \cdot \log_{10}\left(P_p[k,n]\right)$$

La expansión se efectúa independientemente para cada grupo de frecuencias *k*:

$$E_2[k,n] = \frac{1}{Norm_{SP}[k]} \left( \sum_{j=0}^{Z-1} E_{line}[j,k,n]^{0,4} \right)^{\frac{1}{0,4}} \tag{17}$$

donde $E_{line}$ viene dado por:

$$E_{line}[j,k,n] = \begin{cases} \dfrac{10^{\frac{L[j,n]}{10}} \cdot 10^{\frac{-res \cdot (j-k) \cdot s_l[j, L[j,n]]}{10}}}{\sum_{\mu=0}^{j-1} 10^{\frac{-res \cdot (j-\mu) \cdot s_l[j, L[j,n]]}{10}} + \sum_{\mu=j}^{Z-1} 10^{\frac{res \cdot (\mu-j) \cdot s_u[j, L[j,n]]}{10}}} & \text{si } k < j \\ \dfrac{10^{\frac{L[j,n]}{10}} \cdot 10^{\frac{res \cdot (k-j) \cdot s_u[j, L[j,n]]}{10}}}{\sum_{\mu=0}^{j-1} 10^{\frac{-res \cdot (j-\mu) \cdot s_l[j, L[j,n]]}{10}} + \sum_{\mu=j}^{Z-1} 10^{\frac{res \cdot (\mu-j) \cdot s_u[j, L[j,n]]}{10}}} & \text{si } k \geq j \end{cases} \tag{18}$$

$Norm_{SP}[k]$ se calcula de acuerdo con:

$$Norm_{SP}[k] = \left( \sum_{j=0}^{Z-1} \tilde{E}_{line}[j,k]^{0,4} \right)^{\frac{1}{0,4}} \tag{19}$$

con:

$$\tilde{E}_{line}[j,k] = \begin{cases} \dfrac{10^{\frac{-res\cdot(j-k)\cdot s_l[j,0]}{10}}}{\sum_{\mu=0}^{j-1} 10^{\frac{-res\cdot(j-\mu)\cdot s_l[j,0]}{10}} + \sum_{\mu=j}^{Z-1} 10^{\frac{res\cdot(\mu-j)\cdot s_u[j,0]}{10}}} & \text{si } k < j \\ \dfrac{10^{\frac{res\cdot(k-j)\cdot s_u[j,0]}{10}}}{\sum_{\mu=0}^{j-1} 10^{\frac{-res\cdot(j-\mu)\cdot s_l[j,0]}{10}} + \sum_{\mu=j}^{Z-1} 10^{\frac{res\cdot(\mu-j)\cdot s_u[j,0]}{10}}} & \text{si } k \geq j \end{cases} \tag{20}$$

y *res* es la resolución de la escala de altura del sonido en Bark (0,25 para la versión básica y 0,5 para la versión avanzada).

Los patrones en esta etapa de procesamiento, $E_2[k,n]$, se utilizan más adelante para el cálculo de patrones de modulación y se denominan «patrones de excitación no explayados» («*unsmeared excitation patterns*»).

### 2.1.8 Expansión en el dominio del tiempo

Con el fin de modelar el enmascaramiento hacia adelante, las energías en cada grupo de frecuencias son explayadas en el tiempo mediante filtros de paso bajo de primer orden. Las constantes de tiempo dependen de la frecuencia central de cada grupo (tal como se indica en la ecuación (10) y en el Cuadro 6) y se calculan de acuerdo con:

$$\tau = \tau_{mín} + \frac{100 \text{ Hz}}{f_c[k]} \cdot (\tau_{100} - \tau_{mín}) \quad \left| \begin{array}{l} \tau_{100} = 0{,}030 \text{ s} \\ \tau_{mín} = 0{,}008 \text{ s} \end{array} \right. \tag{21}$$

Los filtros de paso bajo de primer orden se calculan de acuerdo con:

$$E_f[k,n] = a \cdot E_f[k,n-1] + (1-a) \cdot E_2[k,n] \tag{22}$$

$$E[k,n] = \text{máx}\left(E_f(k,n), E_2(k,n)\right) \tag{23}$$

donde *a* se calcula a partir de las mencionadas constantes de tiempo por la fórmula:

$$a = \text{e}^{-\frac{4}{187,5} \cdot \frac{1}{\tau}} \tag{24}$$

*n* es el número de la trama en curso, *k* es el índice del grupo y $E_f[k,0] = 0$.

Los patrones en esta etapa de procesamiento, $E[k,n]$, se denominan «*patrones de excitación (excitation patterns)*».

### 2.1.9 Umbral de enmascaramiento

El enmascaramiento describe el efecto en virtud del cual una señal claramente audible, aunque débil, se hace inaudible cuando aparece una señal correspondientemente más alta (de mayor sonoridad). Este umbral se calcula ponderando los patrones de excitación con la función de ponderación $m[k]$.

$$m[k] = \begin{cases} 3{,}0 & \text{para} \quad k \cdot \text{res} \le 12 \\ 0{,}25 \cdot k \cdot \text{res} & \text{para} \quad k \cdot \text{res} > 12 \end{cases} \tag{25}$$

$$M[k,n] = \frac{E[k,n]}{10^{\frac{m[k]}{10}}} \tag{26}$$

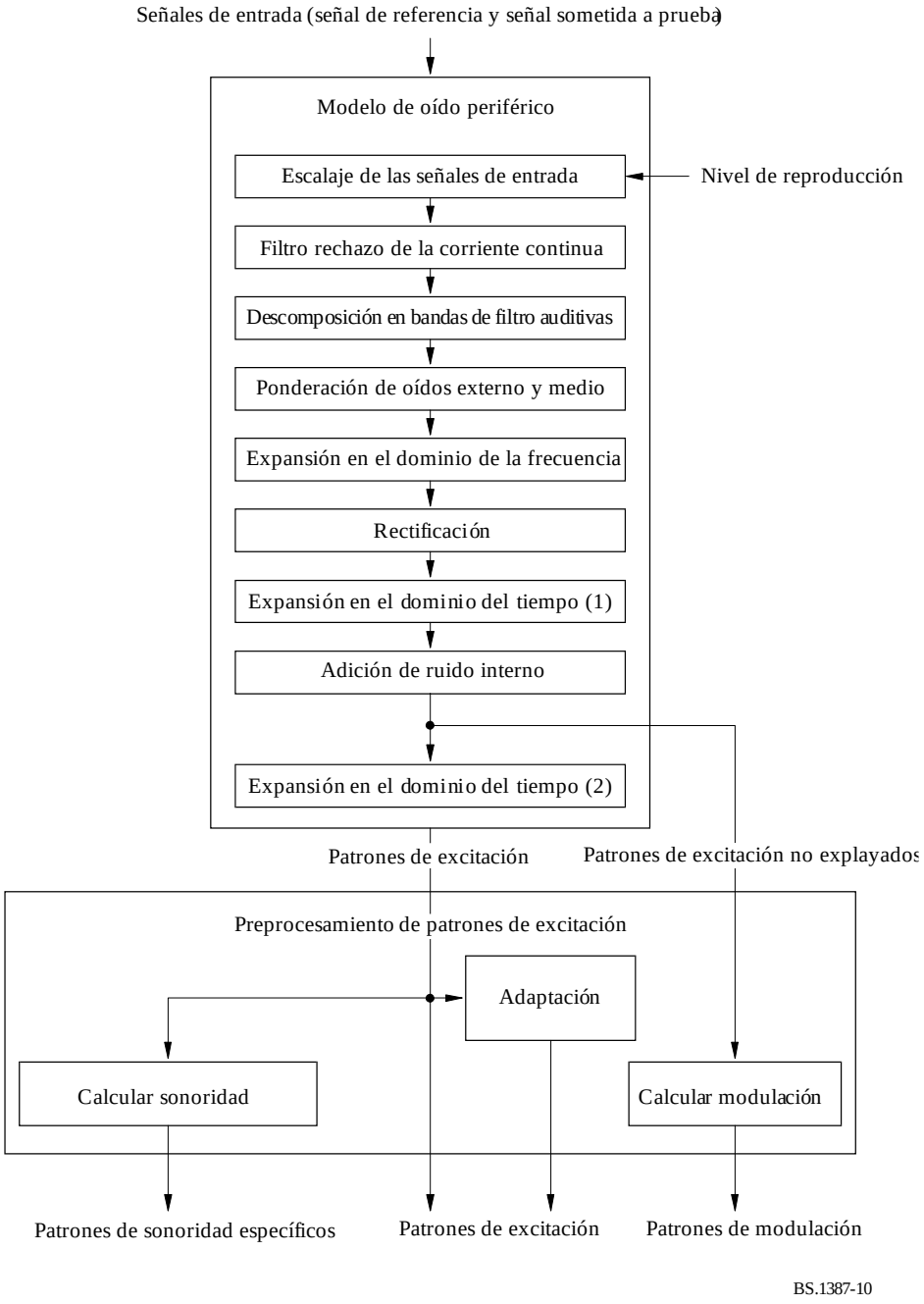
Los patrones en esta etapa de procesamiento, $M[k,n]$, se denominan «*patrones de enmascaramiento (mask patterns)*».

## 2.2 Modelo de oído basado en banco de filtros

### 2.2.1 Visión de conjunto

FIGURA 10

**Modelo de oído periférico y preprocesamiento de patrones de excitación para la parte del modelo basada en un banco de filtros**




A la entrada del modelo de oído basado en banco de filtros, la señal sometida a prueba y la señal de referencia son ajustadas al nivel de reproducción supuesto y pasadas a través de un filtro de paso alto para eliminar los componentes de corriente continua y los de frecuencias inferiores a la audibles de las señales. Seguidamente, las señales son descompuestas en señales de paso de banda por filtros de fases lineales que están distribuidos uniformemente a lo largo de la escala de alturas del sonido perceptuales. A las señales de paso de banda se les aplica una ponderación dependiente de la

frecuencia para modelar las características espectrales del oído externo y del oído medio. La resolución espectral dependiente del nivel, de los filtros auditivos, se modela en el dominio de la frecuencia por una convolución de las salidas con una función de expansión dependiente del nivel.

Las envolventes de las señales se calculan utilizando la transformada de Hilbert de las señales de paso de banda («rectificación») y una convolución en el dominio del tiempo con una función de ventana aplicada para modelar el enmascaramiento hacia atrás. Después, se añade un desplazamiento dependiente de la frecuencia que tiene en cuenta el ruido interno del sistema auditivo y modela el umbral en condiciones de silencio. Por última, se efectúa una segunda convolución en el dominio del tiempo, utilizando una función de expansión exponencial que modela el enmascaramiento hacia adelante.

Los *patrones de excitación* así obtenidos se utilizan para calcular *patrones de sonoridad específicos*, y los patrones obtenidos antes de la última expansión en el dominio del tiempo («patrones de excitación no explayados») («*unsmeared excitation patterns*») se utilizan para calcular «*patrones de modulación*». Estos patrones, junto con los patrones de excitación propiamente dichos, sirven de base para calcular los valores del modelo. Con el fin de separar la influencia de la respuesta a frecuencia en estado estacionario del dispositivo sometido a prueba, de otras distorsiones, los patrones de excitación de la señal de prueba y de la señal de referencia son también adaptados espectralmente uno al otro («adaptación»). Los patrones de modulación y los patrones de sonoridad específicos se calculan a partir de los patrones de excitación adaptados y de los patrones de excitación no adaptados.

### 2.2.2 Submuestreo

A la salida del banco de filtros, estas señales son submuestreadas por un factor de 32 y, después del primera expansión en el dominio del tiempo, las señales son submuestreadas por un factor de 6 (véase la Fig. 11).

### 2.2.3 Fijación del nivel de reproducción

El factor de escalaje (*scaling factor*) para la entrada se calcula a partir del nivel de reproducción supuesto de una señal de entrada a plena escala por la fórmula:

$$fac = \frac{10^{L_{máx}/20}}{32\,767} \tag{27}$$

Cuando no se conozca el nivel exacto de reproducción, se recomienda dar a $L_{máx}$ el valor 92 $dB_{SPL}$.

### 2.2.4 Filtro de rechazo de la corriente continua

Como el banco de filtros es sensible a los componentes de frecuencias subsónicas en las señales de entrada, se aplica a éstas un filtro de rechazo de la corriente continua. Se utiliza un filtro de paso alto Butterworth de cuarto orden con una frecuencia de corte de 20 Hz. El filtro se realiza como dos filtros de respuesta a impulso infinito (filtros IIR) de segundo orden conectados en cascada:

$$y_n = x_n - 2x_{n-1} + x_{n-2} + b_1 y_{n-1} + b_2 y_{n-2} \tag{28}$$

donde los coeficientes del primer bloque son:

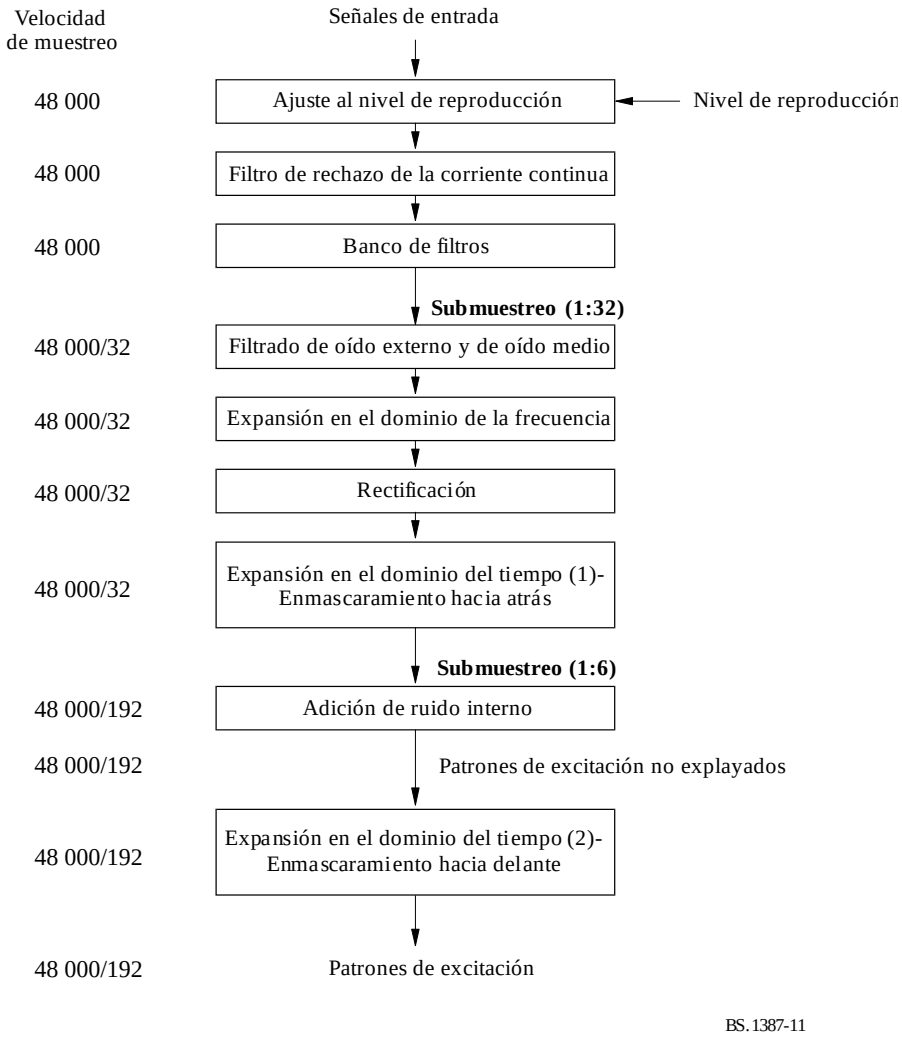
$$b_{1,2} = 1{,}99517,\ -0{,}995174$$

y los coeficientes del segundo bloque son:

$$b_{1,2} = 1{,}99799,\ -0{,}997998$$

FIGURA 11

**Submuestreo en el modelo de oído periférico basado en banco de filtros**




### 2.2.5 Banco de filtros

El banco de filtros está formado por 40 pares de filtros para cada canal de señal de prueba y de señal de referencia. Los filtros están uniformemente espaciados y tienen una anchura de banda absoluta constante cuando se relacionan con una escala auditiva de alturas del sonido. Los dos filtros que constituyen el par tienen una respuesta a frecuencia idéntica en magnitud pero con una diferencia de fase de 90°. Por tanto, la salida del segundo filtro representa la transformada de Hilbert de la salida del primero (o la parte imaginaria, si se supone que el primer filtro representa la parte real de una señal compleja). Las envolventes de sus respuestas a impulso tienen una forma de coseno cuadrado ($\cos^2$). Los filtros están definidos por el Cuadro 8 y la ecuación (29) (donde $k$ es el índice del filtro, $n$ es el índice de la muestra temporal y $T$ es el tiempo entre dos muestras: $T = 1/48\,000$). Pueden ser realizados como filtros de respuesta a impulso finito (filtros FIR) utilizando los valores $h_{re}(k,n)$ y $h_{im}(k,n)$ como coeficientes. Cuando las señales de entrada están limitadas en el tiempo, las salidas de los filtros pueden también ser calculadas por algoritmos recurrentes muy rápidos.

$$\left.\begin{aligned} h_{re}(k,n) &= \frac{4}{N[k]} \cdot \text{sen}^2\left(\pi \cdot \frac{n}{N[k]}\right) \cdot \cos\left(2\pi \cdot f_c[k] \cdot \left(n - \frac{N[k]}{2}\right) \cdot T\right) \\ h_{im}(k,n) &= \frac{4}{N[k]} \cdot \text{sen}^2\left(\pi \cdot \frac{n}{N[k]}\right) \cdot \text{sen}\left(2\pi \cdot f_c[k] \cdot \left(n - \frac{N[k]}{2}\right) \cdot T\right) \end{aligned}\right| \; 0 \le n < N[k]$$

$$h_{re}(k,n) = h_{im}(k,n) = 0 \quad \left| \begin{array}{l} n < 0 \\ n \ge N[k] \end{array} \right. \qquad (29)$$

CUADRO 8

**Frecuencia central, longitud de respuesta a impulso y retardo adicional para cada filtro**

| **Índice de filtro** | **Frecuencia central/Hz** | **Longitud de respuesta a impulso/muestras** | **Retardo adicional/muestras** |
|---|---|---|---|
| *k* | $f_C[k]$ | *N[k]* | *D[k]* |
| 0 | 50,00 | 1 456 | 1 |
| 1 | 116,19 | 1 438 | 10 |
| 2 | 183,57 | 1 406 | 26 |
| 3 | 252,82 | 1 362 | 48 |
| 4 | 324,64 | 1 308 | 75 |
| 5 | 399,79 | 1 244 | 107 |
| 6 | 479,01 | 1 176 | 141 |
| 7 | 563,11 | 1 104 | 177 |
| 8 | 652,97 | 1 030 | 214 |
| 9 | 749,48 | 956 | 251 |
| 10 | 853,65 | 884 | 287 |
| 11 | 966,52 | 814 | 322 |
| 12 | 1 089,25 | 748 | 355 |
| 13 | 1 223,10 | 686 | 386 |
| 14 | 1 369,43 | 626 | 416 |
| 15 | 1 529,73 | 570 | 444 |
| 16 | 1 705,64 | 520 | 469 |
| 17 | 1 898,95 | 472 | 493 |
| 18 | 2 111,64 | 430 | 514 |
| 19 | 2 345,88 | 390 | 534 |
| 20 | 2 604,05 | 354 | 552 |
| 21 | 2 888,79 | 320 | 569 |
| 22 | 3 203,01 | 290 | 584 |
| 23 | 3 549,90 | 262 | 598 |
| 24 | 3 933,02 | 238 | 610 |
| 25 | 4 356,27 | 214 | 622 |
| 26 | 4 823,97 | 194 | 632 |

CUADRO 8 (*fin*)

| **Índice de filtro** | **Frecuencia central/Hz** | **Longitud de respuesta a impulso/muestras** | **Retardo adicional/muestras** |
|---|---|---|---|
| *k* | $f_C[k]$ | *N[k]* | *D[k]* |
| 27 | 5 340,88 | 176 | 641 |
| 28 | 5 912,30 | 158 | 650 |
| 29 | 6 544,03 | 144 | 657 |
| 30 | 7 242,54 | 130 | 664 |
| 31 | 8 014,95 | 118 | 670 |
| 32 | 8 869,13 | 106 | 676 |
| 33 | 9 813,82 | 96 | 681 |
| 34 | 10 858,63 | 86 | 686 |
| 35 | 12 014,24 | 78 | 690 |
| 36 | 13 292,44 | 70 | 694 |
| 37 | 14 706,26 | 64 | 697 |
| 38 | 16 270,13 | 58 | 700 |
| 39 | 18 000,02 | 52 | 703 |

Las frecuencias centrales están comprendidas entre 50 Hz y 18 000 Hz. La escala auditiva de alturas del sonido se calcula mediante una formula aproximada presentada por [Schroeder y otros, 1979]:

$$z / Bark = 7 \cdot \text{arsenh}\left(\frac{f / \text{Hz}}{650}\right) \tag{30}$$

Las unidades de altura del sonido se denominan *Bark* (debe señalarse, no obstante, que esta escala no representa exactamente la escala Bark definida por [Zwicker y Feldtkeller, 1967]).

Con el fin de tener iguales retardos para todos los filtros, la entrada de cada filtro se retarda en *D* muestras, donde *D* es la mitad de la diferencia entre la longitud de su respuesta a impulso y la longitud de la respuesta a impulso del filtro que tenga la respuesta a impulso más larga[3]:

$$D[k] = 1 + \frac{1}{2}\left(N[0] - N[k]\right) \tag{31}$$

Las salidas de los filtros son submuestreadas por un factor de 32, es decir, se calculan valores de salida cada 32ª muestra de entrada para todos los filtros[4].

[3] El retardo adicional de una muestra no es necesario para la realización. Obsérvese que la realización de referencia utilizada para pruebas de observancia (*compliance*) incluye este retardo adicional.

[4] En realidad, las envolventes de los filtros en las bandas de frecuencias superiores no cumplen necesariamente el teorema del muestreo. Aun cuando la distorsión por solapamiento (*aliasing*) sólo ocurra en condiciones muy particulares (por ejemplo, componentes de alta frecuencia modulados con frecuencias mayores que 1,5 kHz) y nunca se hayan presentado problemas relacionados con esos efectos en las bases de datos conocidas, debe señalarse que pueden presentarse problemas de distorsión por solapamiento, especialmente cuando se utilizan señales de prueba artificiales.

### 2.2.6 Filtrado de oído externo y de oído medio

La respuesta a frecuencia del oído externo y del oído medio se modela por una función de ponderación dependiente de la frecuencia que se aplica a las salidas de los filtros (ecuación (32)).

$$W[k]/\text{dB} = -0{,}6 \cdot 3{,}64 \cdot \left(\frac{f_c[k]}{\text{kHz}}\right)^{-0{,}8} + 6{,}5 \cdot \text{e}^{-0{,}6\cdot\left(\frac{f_c[k]}{\text{kHz}}-3{,}3\right)^2} - 10^{-3} \cdot \left(\frac{f_c[k]}{\text{kHz}}\right)^{3{,}6} \quad (32)$$

#### 2.2.6.1 Seudocódigo

| | | |
|---|---|---|
| /* inputs */ | | |
| out_re,out_im | : | salidas de bancos de filtros (partes real e imaginaria) |
| W | : | función de ponderación (véase la ecuación (32)) |
| /* outputs */ | | |
| out_re,out_im | : | salidas de bancos de filtros |
| /* intermediate values */ | | |
| k | : | índice del filtro |
| Wt | : | factor de ponderación |

```
/* outer and middle ear filtering */
for(k=0..39)
{
        Wt    = pow(10,W[k]/20)
        out_re[k] *= Wt;
        out_im[k]  *= Wt;
}
```

### 2.2.7 Expansión (*spreading*) en el dominio de la frecuencia

Los valores de salida del banco de filtros son explayados (*smeared*) en el campo de la frecuencia utilizando una función de expansión dependiente del nivel. La función de expansión es una exponencial a la derecha y a la izquierda. La pendiente inferior es siempre de 31 dB/Bark y la pendiente superior varía entre −24 y −4 dB/Bark.

La pendiente superior, $s[k]$, se calcula de acuerdo con:

$$\frac{s[k]}{\text{dB}/Bark} = \text{mín}\left(-4{,}-24 - \frac{230\ \text{Hz}}{f_c[k]} + 0{,}2 \cdot L[k]/\text{dB}\right) \quad (33)$$

El nivel $L[k]$ se calcula independientemente para cada canal de filtro tomando el valor absoluto cuadrático de la salida del filtro y convirtiéndolo a la escala de dB. Las frecuencias centrales, $f_c[k]$, se toman del Cuadro 8, las representaciones lineales de las pendientes son alisadas en el tiempo por un filtro de paso bajo de primer orden con una constante de tiempo de 100 ms.

La expansión se lleva a cabo independientemente para los filtros que representan la parte real de las señales y los filtros que representan las partes imaginarias de las señales (ecuación (29)). La expansión en frecuencia se efectúa primero para la pendiente superior (dependiente del nivel), y después para la pendiente inferior, utilizando un algoritmo de filtro IIR de primer orden.

#### 2.2.7.1 Seudocódigo

| | | |
|---|---|---|
| /* inputs */ | | |
| out_re,out_im | : | salidas de banco de filtros (partes real e imaginaria) |
| z[ ] | : | tasas de bandas críticas para las frecuencias centrales de bandas de filtro en Bark (de acuerdo con el Cuadro 8 y la ecuación (30)) |
| /* outputs */ | | |
| A_re,A_im | : | características de salida |
| /* intermediate values */ | | |
| j,k | : | patrones de salida |
| a,b | : | coeficientes para alisamiento temporal |
| dist | : | constante para calcular la diafonía |
| L[ ] | : | nivel en cada salida de filtro |
| s[ ] | : | pendiente local para expansión ascendente |
| d1,d2 | : | memorias tampón |
| /* static */ | | (los valores de la trama precedente son preservados; los valores son inicializados con ceros al comienzo de la medición) |
| cl, cu[ ] | : | fracción ensanchada de la señal |

```
/* level dependent upward spreading */
dist      = pow(0.1,(z[39]-z[0])/(39*20));
/* (z[39]-z[0])/39 is the distance in Bark between two adjacent filter bands */
a         = exp(-32/(48000*0.1));
b         = 1 - a;
for(k=0..39)
{
          A_re[k]    = out_re[k];
          A_im[k]   = out_im[k];
}
for(k=0..39)
{
          /* calculate level dependent slope */
          L[k]   = 10*log10(out_re[k]* out_re[k] + out_im[k]* out_im[k];
          s[k]    = max(4,(24 + 230/fcentre[k] - 0.2*L[k]));
          /* calculate spread fraction and smooth it over time*/
          cu[k] = a*pow(dist,s[k])+b*cu[k];
          /* spreading of band k */
          d1      = out_re[k]
          d2      = out_im[k]
          for(j=k+1..39)
          {
           d1   *= cu[k];
           d2   *= cu[k];
           A_re[j]   += d1;
           A_im[j]  += d2;
          }
}
/* downward spreading */
cl         = pow(dist,31);
d1         = 0;
d2         = 0;
for(k=39..0)
{
          /* spreading of band k */
          d1     = d1 * cl + A_re[k];
          d2     = d2 * cl + A_im[k];
          A_re[k]    = d1;
          A_im[k]   = d2;
}
```

### 2.2.8 Rectificación

Las energías a la salida de los filtros se calculan sumando los valores cuadráticos del filtro de representa la parte real de la señal y del que representa la parte imaginaria de la señal.

$$E_0[k,n] = A_{re}[k,n]^2 + A_{im}[k,n]^2 \tag{34}$$

Sobre estas energías se efectúan las operaciones siguientes.

### 2.2.9 Dispersión (smearing) en el dominio del tiempo (1) – Enmascaramiento hacia atrás

Para modelar el enmascaramiento hacia atrás, las energías en las salidas de filtro son explayadas (*smeared*) en el dominio del tiempo por medio de un filtro con una respuesta a impulso en forma de coseno cuadrado ($\cos^2$) con 12 puntos de derivación (que corresponde a una respuesta de filtro de 384 muestras a la velocidad de muestreo de entrada del banco de filtros). Después de la dispersión en el dominio del tiempo, las salidas son submuestreadas por un factor de 6. Los valores resultantes se multiplican por un factor de calibración $cal_1 = 0{,}9761$ para obtener los niveles de salida adecuados para el nivel de reproducción dado.

$$E_1[k,n] = \frac{0{,}9761}{6} \cdot \sum_{i=0}^{11} E_0[k, 6n-i] \cdot \cos^2\left(\pi \cdot \frac{(i-5)}{12}\right) \tag{35}$$

### 2.2.10 Adición de ruido interno

Tras la primera expansión en el dominio del tiempo se añade un desplazamiento dependiente de la frecuencia, $E_{Thres}$, a las energías en cada canal de filtro.

$$E_{Thres}[k] = 10^{0{,}4 \cdot 0{,}364 \cdot \left(\frac{f_c[k]}{\text{kHz}}\right)^{-0{,}8}} \tag{36}$$

$$E_2[k,n] = E_1[k,n] + E_{Thres}[k,n] \tag{37}$$

Los patrones en esta etapa de procesamiento, $E_2[k,n]$, se utilizarán posteriormente para el cálculo de los patrones de modulación y se denominan «*patrones de excitación no explayados*».

### 2.2.11 Dispersión en el dominio del tiempo (2) – Enmascaramiento hacia adelante

Para modelar el enmascaramiento hacia adelante, las energías de cada canal de filtro son explayadas en el tiempo por filtros de paso bajo de primer orden. Las constantes de tiempo dependen de la frecuencia central de cada filtro (indicada en el Cuadro 5) y se calculan de acuerdo con:

$$\tau = \tau_{mín} + \frac{100\ \text{Hz}}{f_c[k]} \cdot (\tau_{100} - \tau_{mín}) \quad \left| \begin{array}{l} \tau_{100} = 0{,}020\ \text{s} \\ \tau_{mín} = 0{,}004\ \text{s} \end{array} \right. \tag{38}$$

Los filtros de paso bajo de primer orden se calculan por la siguiente fórmula:

$$E[k,n] = a \cdot E[k,n-1] + (1-a) \cdot E_2[k,n] \tag{39}$$

donde $a$ se calcula a partir de las mencionadas constantes de tiempo por:

$$a = \text{e}^{-\frac{192}{48\,000 \cdot \tau}} \tag{40}$$

Los patrones en esta etapa de procesamiento, $E[k,n]$, se denominan «*patrones de excitación*».

## 3 Preprocesamiento de patrones de excitación

La mayor parte de las cálculos descritos en este punto se utilizan con el modelo de oído basado en un banco de filtros así como en el modelo de oído basado en TRF. Como el factor de submuestreo y el número de bandas de frecuencias son diferentes en los dos modelos de oído, las constantes que dependen de este factor se describen utilizando las variables dependientes del modelo de oído *StepSize* y *Z*. Para el modelo de oído basado en TRF, el valor de *StepSize* es *1 024* y *Z* es, o bien *55* (versión avanzada) o *109* (versión básica). Para el modelo de oído basado en un banco de filtros, el valor de *StepSize* es *192* y *Z* es *40*. Cuando no se indica otra cosa, todas las variables y los filtros recursivos se inicializan a cero.

### 3.1 Adaptación de nivel y de patrón

Para compensar las diferencias de nivel y distorsiones lineales entre la señal de prueba y la señal de referencia, los niveles promedio de la señal de prueba y de la señal de referencia se adaptan uno a otro.

En un primer paso, las energías en cada canal de filtro son alisadas por filtros de paso bajo de primer orden. Las constantes de tiempo dependen de las frecuencias centrales de los filtros y se eligen como:

$$\tau = \tau_{mín} + \frac{100\ \text{Hz}}{f_c[k]} \cdot \left(\tau_{100} - \tau_{mín}\right) \quad \left| \begin{array}{l} \tau_{100} = 0{,}050\ \text{s} \\ \tau_{mín} = 0{,}008\ \text{s} \end{array} \right. \tag{41}$$

Los filtros de paso bajo de primer orden se calculan de acuerdo con:

$$P_{Ref}[k,n] = a \cdot P_{Ref}[k,n-1] + (1-a) \cdot E_{Ref}[k,n] \tag{42}$$

$$P_{Test}[k,n] = a \cdot P_{Test}[k,n-1] + (1-a) \cdot E_{Test}[k,n] \tag{43}$$

donde $E_{Test}$ y $E_{Ref}$ son los patrones de excitación que habrán de ser adaptados uno a otro y $a$ se calcula a partir de las constantes de tiempo por:

$$a = \mathrm{e}^{-\frac{StepSize}{48\,000 \cdot \tau}} \tag{44}$$

#### 3.1.1 Adaptación de nivel

A partir de los patrones de entrada pasados por filtro de paso bajo $P_{Test}$ y $P_{Ref}$, se calcula un factor de corrección instantánea *LevCorr* por:

$$LevCorr[n] = \left( \frac{\sum_{k=0}^{Z-1} \sqrt{P_{Test}[k,n] \cdot P_{Ref}[k,n]}}{\sum_{k=0}^{Z-1} P_{Test}[k,n]} \right)^2 \tag{45}$$

Si el factor de corrección es mayor que uno, la señal de referencia se divide por el factor de corrección, y en otro caso se multiplica por el factor de corrección.

$$E_{L,Ref}[k,n] = E_{Ref}[k,n] / LevCorr[n] \quad \left| \quad LevCorr[n] > 1 \right. \tag{46}$$

$$E_{L,Test}[k,n] = E_{Test}[k,n] \cdot LevCorr[n] \quad \left| \quad LevCorr[n] \leq 1 \right. \tag{47}$$

### 3.1.2 Adaptación de patrón

Los factores de corrección para cada canal se calculan comparando las envolventes temporales de las salidas de filtro de la señal de prueba y de la señal de referencia:

$$R[k,n] = \frac{\sum_{i=0}^{n} a[k]^i \cdot E_{L,Test}[k,n-i] \cdot E_{L,Ref}[k,n-i]}{\sum_{i=0}^{n} a[k]^i \cdot E_{L,Ref}[k,n-i] \cdot E_{L,Ref}[k,n-i]} \tag{48}$$

Los valores para *a* se calculan como anteriormente (ecuación (44)) a partir de las constantes de tiempo dadas por la ecuación (41). Si $R[k,n]$ es mayor que uno, el factor de corrección para la señal de prueba se fija a $R[k,n]^{-1}$ y el factor de corrección para la señal de referencia se fija a uno. En el caso contrario, el factor de corrección para la señal de referencia se fija a $R[k,n]$ y el factor de corrección para la señal de prueba se fija a uno.

$$\begin{array}{lll} R_{Test}[k,n] = \dfrac{1}{R[k,n]}, & R_{Ref}[k,n] = 1 & \left| R[k,n] \geq 1 \right. \\ R_{Test}[k,n] = 1, & R_{Ref}[k,n] = R[k,n] & \left| R[k,n] < 1 \right. \end{array} \tag{49}$$

Si el denominador de (48) es cero (en cuyo caso $R[k,n]$ no estaría definido) y el numerador es mayor que cero, $R_{Test}[k,n]$ se fija a cero y $R_{Ref}[k,n]$ se fija a uno. Si el numerador de (48) es también cero, las razones $R_{Test}[k,n]$ y $R_{Ref}[k,n]$ se copian de la banda de frecuencias situada por debajo. Si no hay banda de frecuencias por debajo (es decir, $k = 0$) las razones $R_{Test}[k,n]$ y $R_{Ref}[k,n]$ se fijan a uno.

Los factores de corrección son promediados sobre *M* canales de filtro y alisados en el tiempo (ecuación (50)) utilizando las mismas constantes de tiempo dadas más arriba (ecuaciones (41) a (44)). La anchura de la ventana de frecuencias *M* es 3 para el modelo de oído basado en banco de filtros. Para el modelo de oído basado en TRF es 4 (versión avanzada) u 8 (versión básica) respectivamente.

$$\begin{aligned} PattCorr_{Test}[k,n] &= a \cdot PattCorr_{Test}[k,n-1] + (1-a) \cdot \frac{1}{M} \cdot \sum_{i=-M_1}^{M_2} R_{Test}[k+i,n] \\ PattCorr_{Ref}[k,n] &= a \cdot PattCorr_{Ref}[k,n-1] + (1-a) \cdot \frac{1}{M} \cdot \sum_{i=-M_1}^{M_2} R_{Ref}[k+i,n] \end{aligned} \tag{50}$$

$$\begin{array}{ll} M_1 = M_2 = \dfrac{M-1}{2} & \left| M \text{ impar} \right. \\ M_1 = \dfrac{M}{2} - 1, \quad M_2 = \dfrac{M}{2} & \left| M \text{ par} \right. \end{array}$$

En los bordes de la escala de frecuencias donde la ventana de frecuencias excedería de la gama de bandas de filtro, la anchura de la ventana de frecuencias se reduce en consecuencia:

$$M_1 = \text{mín}(M_1, k), \quad M_2 = \text{mín}(M_2, z-k-1), \quad M = M_1 + M_2 + 1 \tag{51}$$

Los patrones de entrada adaptados en nivel son ponderados con los correspondientes factores de corrección $PattCorr_{Test/Ref}[k,n]$ para obtener los patrones adaptados espectralmente.

$$E_{P,Ref}[k,n] = E_{L,Ref}[k,n] \cdot PattCorr_{Ref}[k,n] \tag{52}$$

$$E_{P,Test}[k,n] = E_{L,Test}[k,n] \cdot PattCorr_{Test}[k,n] \tag{53}$$

### 3.2 Modulación

A partir de los *patrones de excitación explayados*, $E_2[k,n]$, se calcula una sonoridad simplificada elevando la excitación a una potencia de 0.3. Este valor y el valor absoluto de su derivación temporal son explayados en el tiempo.

$$\overline{E}_{der}[k,n] = a \cdot \overline{E}_{der}[k,n-1] + (1-a) \cdot \frac{48\,000}{StepSize} \cdot \left| E_2[k,n]^{0,3} - E_2[k,n-1]^{0,3} \right| \tag{54}$$

$$\overline{E}[k,n] = a \cdot \overline{E}[k,n-1] + (1-a) \cdot E_2[k,n]^{0,3} \tag{55}$$

Los valores de $a$ se calculan como en la ecuación (44) a partir de las constantes de tiempo dadas por:

$$\tau = \tau_0 + \frac{100\ \text{Hz}}{f_c} \cdot (\tau_{100} - \tau_0) \left| \begin{array}{l} \tau_{100} = 0{,}050\ \text{s} \\ \tau_0 = 0{,}008\ \text{s} \end{array} \right. \tag{56}$$

A partir de los valores resultantes, $\overline{E}_{der}$ y $\overline{E}$, se calcula una medida para la modulación de la envolvente en cada salida de filtro:

$$Mod[k,n] = \frac{\overline{E}_{der}[k,n]}{1 + \overline{E}[k,n]/0{,}3} \tag{57}$$

Los valores $\overline{E}$ se utilizan también, más adelante, en el cálculo de la diferencia de modulación.

### 3.3 Sonoridad

Los patrones de sonoridad específica de la señal sometida a prueba y de la señal de referencia se calculan por la fórmula:

$$N[k,n] = const \cdot \left( \frac{1}{s[k]} \cdot \frac{E_{Thres}[k]}{10^4} \right)^{0,23} \cdot \left[ \left( 1 - s[k] + \frac{s[k] \cdot E[k,n]}{E_{Thres}[k]} \right)^{0,23} - 1 \right] \tag{58}$$

como se indica en [Zwicker y Feldtkeller, 1967]. La sonoridad global de la señal sometida a prueba y la señal de referencia se calcula como la suma, para todos los canales de filtro, de todos los valores de sonoridad específica superiores a cero.

$$N_{total}[n] = \frac{24}{Z} \cdot \sum_{k=0}^{Z-1} \text{máx}\left(N[k,n], 0\right) \tag{59}$$

La constante de escalaje se elige como *const* = 1,07664 para el modelo de oído periférico basado en TRF y *const* = 1,26539 para el modelo de oído periférico basado en banco de filtros, con el fin de obtener un valor de sonoridad global de un sonido para un tono de 1 kHz con un nivel de 40 $dB_{SPL}$. El índice de umbral $s$ y la excitación en el umbral $E_{Thres}$ se calculan de acuerdo con:

$$E_{Thres}[k] = 10^{0,364 \cdot \left( \frac{f}{1\,\text{kHz}} \right)^{-0,8}} \tag{60}$$

y

$$s[k] = 10^{\frac{1}{10}\left( -2 - 2{,}05 \cdot atn\left( \frac{f}{4\,\text{kHz}} \right) - 0{,}75 \cdot atn\left( \left( \frac{f}{1\,600\ \text{Hz}} \right)^2 \right) \right)} \tag{61}$$

respectivamente.

NOTA – Debido a los diferentes modelos de oído periféricos, la sonoridad calculada aquí no es idéntica a la sonoridad definida en ISO 532 (Acoustics – Method for calculating loudness level, 1975).

### 3.4 Cálculo de la señal de error

La señal de error sólo se calcula en el modelo basado en TRF. Para el cálculo en el dominio de la frecuencia se toma la diferencia entre los espectros de magnitud, filtrados en el oído externo y en el oído medio, de la señal de referencia y de la señal de prueba (véase el § 2.1.4).

$$F_{noise}[k_f, n] = \left|\left|F_{eref}[k_f, n]\right| - \left|F_{etest}[k_f, n]\right|\right| \quad (62)$$

$F_{noise}$ se convierte al dominio de la altura del sonido (*pitch*) utilizando el algoritmo descrito en el § 2.1.5.

Los valores producidos por este algoritmo, $P_{noise}[n,k]$, se designan por «*patrones de ruido*».

## 4 Cálculo de las variables de salida del modelo

### 4.1 Visión de conjunto

CUADRO 9

**Visión de conjunto de las variables de salida del modelo utilizadas para la predicción de la calidad de audio básica**

| Variable de salida del modelo (MOV) | Calculada en el modelo de oído … | | Utilizada en la versión … | |
|---|---|---|---|---|
| | **TRF** | **banco filtros** | **básica** | **avanzada** |
| $WinModDiff1_B$ | sí | no | sí | no |
| $AvgModDiff1_B$ | sí | no | sí | no |
| $AvgModDiff2_B$ | sí | no | sí | no |
| $RmsModDiff_A$ | no | sí | no | sí |
| $RmsNoiseLoud_B$ | sí | no | sí | no |
| $RmsNoiseLoudAsym_A$ | no | sí | no | sí |
| $AvgLinDist_A$ | no | sí | no | sí |
| $BandwidthRef_B$ | sí | no | sí | no |
| $BandwidthTest_B$ | sí | no | sí | no |
| Total $NMR_B$ | sí | no | sí | no |
| $RelDistFrames_B$ | sí | no | sí | no |
| Segmental $NMR_B$ | sí | no | no | sí |
| $MFPD_B$ | sí | no | sí | no |
| $ADB_B$ | sí | no | sí | no |
| $EHS_B$ | sí | no | sí | sí |

### 4.2 Diferencia de modulación

Las diferencias en la modulación de las envolventes temporales de la señal sometida a prueba y de la señal de referencia se miden calculando una diferencia de modulación local para cada canal de filtro (ecuación (63)) en la que $Mod_{test}$ y $Mod_{Ref}$ se obtienen aplicando la ecuación (57) a la referencia, señal $R_{test}$.

$$ModDiff[k,n] = w \cdot \frac{\left|Mod_{test}[k,n] - Mod_{Ref}[k,n]\right|}{offset + Mod_{Ref}[k,n]}$$
$$\begin{cases} w = 1{,}0 & \left|Mod_{test}[k,n] > Mod_{Ref}[k,n]\right. \\ w = negWt & \left|Mod_{test}[k,n] < Mod_{Ref}[k,n]\right. \end{cases} \quad (63)$$

Una diferencia de modulación instantánea se calcula como la suma de las diferencias de modulación locales para todos los canales de filtro (ecuación (64)).

$$ModDiff[n] = \frac{100}{Z} \sum_{k=0}^{Z-1} ModDiff[k,n] \quad (64)$$

El umbral en condiciones de silencio se tiene en cuenta por un factor de ponderación dependiente del nivel (ecuación (65)) calculado a partir de los patrones de excitación modificados para la señal de referencia indicados en la ecuación (55) y la función de ruido interno definida en la ecuación (36) para el modelo de oído basado en banco de filtros, y la ecuación (13) para el modelo de oído basado en TRF.

$$TempWt[n] = \sum_{k=0}^{Z-1} \frac{\overline{E}_{ref}[k,n]}{\overline{E}_{ref}[k,n] + levWt \cdot E_{Thres}[k]}^{\wedge\, 0{,}3} \quad (65)$$

El promediado temporal de las diferencias de modulación instantáneas *ModDiff[n]* mediante los factores de ponderación *TempWt[n]* se describe en el § 5.2. Los valores de las constantes *negWt*, *offset* y *levWt* se indican en el Cuadro 10.

CUADRO 10

**Variables de salida del modelo con estimación de la diferencia de modulación global**

| MOV (Xxx=Win/Avg/Rms) | negWt | offset | levWt |
|---|---|---|---|
| $XxxModDiff1_B$ | 1 | 1 | 100 |
| $XxxModDiff2_B$ | 0,1 | 0,01 | 100 |
| $XxxModDiff_A$ | 1 | 1 | 1 |

#### 4.2.1 RmsModDiff$_A$

La variable de salida del modelo $RmsModDiff_A$ es la media cuadrática de la diferencia de modulación calculada a partir del modelo de oído basado en banco de filtros. Véase el § 5.2.2 para el promediado temporal y el Cuadro 10 para las constantes.

### 4.2.2 WinModDiff1$_B$

La variable de salida del modelo *WinModDiff1*$_B$ es el promedio enventanado de la diferencia de modulación calculada a partir del modelo de oído basado en TRF. Véase el § 5.2.3 para el promediado temporal y el Cuadro 10 para las constantes. El factor de ponderación temporal que aparece en la ecuación (65) no se aplica a esta variable de salida del modelo.

### 4.2.3 AvgModDiff1$_B$ y AvgModDiff2$_B$

Las variables de salida del modelo *AvgModDiff1*$_B$ y *AvgModDiff2*$_B$ son los promedios lineales de la diferencia de modulación calculada a partir del modelo de oído basado en TRF. *AvgModDiff2*$_B$ y *AvgModDiff1*$_B$ se diferencian en que las constantes se eligen de una manera diferente. Véase el § 5.2.1 para el promediado temporal y el Cuadro 10 para las constantes.

## 4.3 Sonoridad de ruido

Estas variables de salida del modelo estiman la sonoridad parcial de distorsiones aditivas en presencia de la señal de referencia enmascaradora. La fórmula para la sonoridad parcial (ecuación (66)) está concebida para que dé la sonoridad específica del ruido de acuerdo con [Zwicker y Feldtkeller, 1967] si no hay una señal enmascaradora presente y que dé algo así como la relación ruido-enmascaramiento si el ruido es muy pequeño en relación de la señal enmascaradora.

La sonoridad parcial del ruido se calcula de acuerdo con:

$$NL[k,n] = \left( \frac{1}{s_{test}} \cdot \frac{E_{Thres}}{E_0} \right)^{0,23} \cdot \left[ \left( 1 + \frac{\text{máx}(s_{test} \cdot E_{test} - s_{ref} \cdot E_{ref}, 0)}{E_{Thres} + s_{ref} \cdot E_{ref} \cdot \beta} \right)^{0,23} - 1 \right] \tag{66}$$

donde $E_0$ es siempre 1, $E_{Thres}$ es la función de ruido interno $E_{Thres}[k]$ definida en la ecuación (36) y $s$ se calcula de acuerdo con:

$$s = ThresFac_0 \cdot Mod[k,n] + S_0 \tag{67}$$

A menos que se describa de una manera diferente, los patrones de excitación adaptados espectralmente (véase el § 3.1) se utilizan como entradas: $E_{Test} = E_{P,Test}[k,n]$ y $E_{Ref} = E_{P,Ref}[k,n]$. El coeficiente β, que determina la cantidad de enmascaramiento, se calcula por:

$$\beta = \exp\left( -\alpha \cdot \frac{E_{test} - E_{ref}}{E_{ref}} \right) \tag{68}$$

Los valores de sonoridad de ruido instantáneo no se tienen en cuenta hasta que hayan transcurrido 50 ms desde el instante en que la sonoridad global del canal audio izquierdo, o del derecho, haya excedido un valor de $N_{Thres} = 0{,}1$ *sonios* para la señal de prueba y para la señal de referencia (véase el § 5.2.4.2).

En el promediado espectral, los valores instantáneos están normalizados al número de bandas de filtro por banda crítica en vez del número total de bandas de filtro; es decir, el resultado del valor promediado espectral se multiplica por un factor de 24.

Si la sonoridad de ruido instantánea es inferior a un valor umbral, $NL_{mín}$, se fija a cero.

CUADRO 11

**Variables de salida del modelo con estimación de la sonoridad de ruido global**

| MOV (Xxx=Win/Avg/Rms) | $\alpha$ | $ThresFac_0$ | $S_0$ | $NL_{mín}$ |
|---|---|---|---|---|
| $XxxMissingComponents_B$ | 1,5 | 0,15 | 1 | 0 |
| $XxxNoiseLoud_B$ | 1,5 | 0,15 | 0,5 | 0 |
| $XxxMissingComponents_A$ | 1,5 | 0,15 | 1 | 0 |
| $XxxNoiseLoud_A$ | 2,5 | 0,3 | 1 | 0,1 |
| $XxxAdaptSharp_A$ | 1,5 | 0,15 | 1 | 0 |

### 4.3.1 $RmsNoiseLoud_A$

La variable de salida del modelo $RmsNoiseLoud_A$ es la media cuadrática de la sonoridad de ruido calculada a partir del modelo de oído basado en banco de filtros. Véase el § 5.2.2 para el promediado temporal y el Cuadro 11 para las constantes.

### 4.3.2 $RmsMissingComponents_A$

La variable de salida del modelo $RmsMissingComponents_A$ es la media cuadrática de la sonoridad de ruido calculada a partir del modelo de oído basado en banco de filtros. Se calcula con los patrones de excitación adaptados según el espectro de la señal de prueba y de la señal de referencia intercambiados, con el fin de obtener la sonoridad de componentes en la señal de referencia que se pierden en la señal de prueba. Véase el § 5.2.2 para el promediado temporal y el Cuadro 11 para las constantes.

### 4.3.3 $RmsNoiseLoudAsym_A$

La variable de salida del modelo $RmsNoiseLoudAsym_A$ es la suma ponderada de las medias cuadráticas de la sonoridad de ruido (véase el § 4.3.1) y la sonoridad de los componentes de señal perdidos (véase el § 4.3.2), calculadas ambas a partir del modelo de oído basado en banco de filtros.

$$RmsNoiseLoudAsym = RmsNoiseLoud + 0{,}5 \cdot RmsMissingComponents \quad (69)$$

### 4.3.4 $AvgLinDist_A$

La variable de salida del modelo $AvgLinDist_A$ mide la sonoridad de los componentes de señal perdidos durante la adaptación espectral de la señal sometida a prueba y la señal de referencia. Utiliza la excitación adaptada espectralmente de la señal de referencia como referencia y la excitación no adaptada de la referencia como la señal de prueba. La MOV se calcula a partir del modelo de oído basado en banco de filtros. Véase el § 5.2.1 para el promediado temporal y el Cuadro 11 para las constantes.

### 4.3.5 $RmsNoiseLoud_B$

La variable de salida del modelo $RmsNoiseLoud_B$ es la media cuadrática de la sonoridad de ruido calculada a partir del modelo de oído basado en TRF. Véase el § 5.2.2 para el promediado temporal y el Cuadro 11 para las constantes.

## 4.4 Anchura de banda

Estos valores de salida del modelo estiman la anchura de banda media de la señal sometida a prueba y de la señal de referencia en líneas TRF.

Para cada trama, la anchura de banda local $Bw_{Ref}[n]$ y $Bw_{Test}[n]$ se calcula de acuerdo con el seudocódigo que se indica a continuación.

### 4.4.1 Seudocódigo

```
/* inputs */
FLevRef[], FlevelTest[]     :  nivel de salidas TFR en dB
/* outputs */
BwRef, BwTest               :  patrones de salida
/* intermediate values */
k                           :  índice de líneas TFR
ZeroThreshold               :  umbral de anchura de banda
```

```
ZeroThreshold = FlevelTst(921);
BwRef = BwTst = 0.0;
for(k=921;k<1024;k++)
{
 ZeroThreshold=max(ZeroThreshold,FLevelTst(k));
}

        for (k = 920; k>=0; k--)
        {
         if (FLevelRef[k] >= 10.0+ZeroThreshold)
          {
              BwRef = k+1;
              break;
          }
        }
        for (k = BwRef-1; k>=0; k--)
        {
         if(FLeveltest[k] >= 5.0+ZeroThreshold)
          {
              BwTest=k+1;
              break;
          }
        }
```

### 4.4.2 $BandwidthRef_B$ y $BandwidthTest_B$

$BandwidthRef_B$ es el promedio lineal de BwRef y $BandwidthTest_B$ es el promedio lineal de BwTest. Para el promediado sólo se tienen en cuenta las tramas en que BwRef > 346. Las tramas con bajo nivel de energía en los extremos de comienzo y de fin tampoco se tienen en cuenta (véase el § 5.2.4.4). Véase el § 5.2.1 para el promediado temporal.

### 4.5 Relación ruido-enmascaramiento

Los siguientes valores del modelo se calculan a partir de los valores de ruido y de enmascaramiento.

La *NMR* local de la trama *n* en curso se define como:

$$NMR_{local}[n] = 10 * \log_{10} \frac{1}{Z} \sum_{k=0}^{Z-1} \frac{P_{noise}[k,n]}{M[k,n]} \tag{70}$$

#### 4.5.1 $NMR_B$ total

La variable de salida del modelo *$NMR_B$ Total* es el promedio lineal de la relación ruido-enmascaramiento calculada mediante:

$$NMR_{tot} = 10 * \log_{10} \frac{1}{N} \sum_{n} \left( \frac{1}{Z} \sum_{k=0}^{Z-1} \frac{P_{noise}[k,n]}{M[k,n]} \right) \tag{71}$$

Las tramas con bajo nivel de energía en los extremos de comienzo y fin de los elementos no se tienen en cuenta (véase el § 5.2.4.4).

#### 4.5.2 $NMR_B$ segmental

La variable de salida del modelo *$NMR_B$ Segmental* es el promedio lineal de la NMR local. Véase el § 5.2.1 para el promediado temporal.

Las tramas con bajo nivel de energía en los extremos de comienzo y fin de los elementos no se tienen en cuenta (véase el § 5.2.4.4).

### 4.6 Relative Disturbed $Frames_B$

La variable de salida del modelo *Relative Disturbed $Frames_B$* (abreviatura: $RelDistFrames_B$) representa el número de tramas con:

$$\max_{\forall k} \left( 10 \cdot \log \left( \frac{P_{noise}[k,n]}{M[k,n]} \right) \right) \geq 1{,}5 \text{ dB} \qquad k \in [0, Z-1]$$

en relación con el número total de tramas del elemento.

Las tramas con bajo nivel de energía en los extremos de comienzo y fin de los elementos no se tienen en cuenta (véase el § 5.2.4.4).

### 4.7 Probabilidad de detección

Las MOV definidas en este punto se basan en $\tilde{E}[k,n]$ (banda *k*, trama *n*), que son los *patrones de excitación* $E[k,n]$ expresados en dB:

$$\tilde{E}[k,n] = 10 \cdot \log_{10}\left(E[k,n]\right) \tag{72}$$

Para cada trama *n*:

Los siguientes pasos se efectúan independientemente para cada canal *c* (los valores de *c* son izquierdo y derecho). Los *patrones de excitación logarítmica* son $\tilde{E}_{ref}[k,n]$ para la señal de referencia y $\tilde{E}_{test}[k,n]$ para la señal sometida a prueba, respectivamente.

Para cada banda *k*:

– Calcular la excitación promedio asimétrica.

$$L[k,n] = 0{,}3 \cdot \text{máx}(\tilde{E}_{ref}[k,n], \tilde{E}_{test}[k,n]) + 0{,}7 \cdot \tilde{E}_{test}[k,n] \quad (73)$$

– Calcular el tamaño del paso de detección efectivo. La fórmula siguiente es una aproximación de la diferencia de nivel justamente perceptible medida por [Zwicker y Fastl, 1990].

Si $L[k,n] > 0$:

$$s[k,n] = 5{,}95072 \cdot ((6{,}39468)/L[k,n])^{1{,}71332} + 9{,}01033 \times 10^{-11} \cdot L[k,n]^4 + 5{,}05622 \times 10^{-6} \times L[k,n]^3 - 0{,}00102438 \cdot L[k,n]^2 + 0{,}0550197 \cdot L[k,n] - 0{,}198719$$

en otro caso:

$$s[k,n] = 1{,}0 \times 10^{30} \quad (74)$$

– Calcular el error con signo e:

$$e[k,n] = \tilde{E}_{ref}[k,n] - \tilde{E}_{test}[k,n] \quad (75)$$

– Si $\tilde{E}_{ref}[k,n] > \tilde{E}_{test}[k,n]$, entonces la pendiente $b$ se fija a 4,0 y en otro caso se fija a 6,0. De esta manera se modela el efecto que consiste en que un aumento en la energía de la señal sometida a prueba, en comparación con la señal de referencia, produce un efecto mayor que una disminución en dicha energía.

– Calcular el factor de escala $a$:

$$a[k,n] = \frac{10^{\frac{\log_{10}(\log_{10}(2{,}0))}{b}}}{s[k,n]} \quad (76)$$

– Calcular la probabilidad de detección. La ecuación (76) fija el factor de escala $a$ de tal modo que si $e[k,n]$ es igual a $s[k,n]$, $p_c[k,n]$ se convierte en *0,5*:

$$p_c[k,n] = 1 - 10^{(-a[k,n] \cdot e[k,n])^b)} \quad (77)$$

– Calcular el número total de pasos por encima del umbral:

$$q_c[k,n] = \frac{|\text{INT}(e[k,n])|}{s[k,n]} \quad (78)$$

– La probabilidad de detección binaural es:

$$p_{bin}[k,n] = \text{máx}(p_{left}[k,n], p_{right}[k,n]) \quad (79)$$

– El número de pasos por encima del umbral para el canal binaural es:

$$q_{bin}[k,n] = \text{máx}(q_{left}[k,n], q_{right}[k,n]) \quad (80)$$

La probabilidad total de detección del canal $c$ de la trama $n$ es:

$$P_c[n] = 1 - \prod_{\forall k} (1 - p_c[k,n]) \quad (81)$$

donde $c$ puede ser *left*, *right* o *bin*. El número total de pasos por encima del umbral para el canal $c$ de la trama $n$ es:

$$Q_c[n] = \sum_{\forall k} q_c[k,n] \quad (82)$$

### 4.7.1 Probabilidad máxima filtrada de detección ($MFPD_B$)

Una versión alisada de la probabilidad de detección para cada canal *c* se calcula por:

$$\tilde{P}_c[n] = (1 - c_0) \cdot P_c[n] + c_0 \cdot \tilde{P}_c[n-1] \tag{83}$$

donde $P_c[-1] = 0$. La constante $c_0$ depende de *StepSize*:

$$c_0 = 0{,}9^{StepSize/1\,024} \tag{84}$$

$c_0$ reduce la sensibilidad a las distorsiones muy cortas.

La probabilidad máxima filtrada de detección (abreviatura: MFPD) se calcula por:

$$PM_c[n] = \max\left(PM_c[n-1] \cdot c_1, \tilde{P}_c[n]\right) \tag{85}$$

donde $PM_c[-1]$ es cero. La constante $c_1$ depende de *StepSize*:

$$c_1 = 0{,}99^{StepSize/1\,024} \tag{86}$$

$c_1$ modela el efecto que consiste en que las distorsiones que se producen al principio de un extracto son menos severas que las que se producen al final del extracto, debido al olvido. Obsérvese que esta constante es útil para el modelado de pruebas de escucha en las que los sujetos no están autorizados a seleccionar partes más cortas del extracto. Para el presente modelo, que se ha calibrado con datos de pruebas de escucha conformes a la Recomendación UIT-R BS.1116, $c_1$ debe ser 1,0.

La MOV MFPD es el valor de $PM_{bin}[n]$ para la última trama.

### 4.7.2 Bloque distorsionado promedio[5] ($ADB_B$)

El número de tramas válidas con una probabilidad de detección del canal central $P_{bin}[n]$ superior a 0,5 se cuenta como ($n_{distorted}$).

Para todas las tramas válidas, el número total de pasos por encima del umbral del canal central $Q_{bin}[n]$ se calcula:

$$Q_{sum} = \sum_{\forall n} Q_{bin}[n]$$

La distorsión del bloque distorsionado medio *ADB* se calcula:

– si $n_{distorted}$ es cero entonces $ADB = 0$ (sin distorsión audible);

– si $n_{distorted} > 0$ y $Q_{sum} > 0$ entonces $ADB = \log_{10}((Q_{sum}) / n_{distorted}))$;

– si $n_{distorted} > 0$ y $Q_{sum}$ es cero entonces $ADB = -0{,}5$.

### 4.8 Estructura armónica del error

Una señal de referencia que contiene fuertes armónicos (por ejemplo, clarinete bajo, clavicordio) tiene un espectro caracterizado por un número de puntas regularmente espaciadas, separadas por profundas depresiones. En algunas condiciones, la señal de error puede heredar esa estructura. Por ejemplo, el ruido mezclado con tal señal es más probable que permanezca no enmascarado donde la señal sea baja en las depresiones espectrales. El espectro de error resultante contendría entonces una estructura similar a la del espectro original, pero desplazado en frecuencia de modo que corresponda

---

5 El término «bloque» es equivalente en este texto al término «trama».

con las ubicaciones de las depresiones. Esta estructura puede tener por resultado una distorsión con cualidades tonales que podrían aumentar las posibilidades de que el error sobresalga.

El error se define como la diferencia en espectro de las señales de referencia y procesada, cada una ponderada por la respuesta en frecuencia del oído externo y del oído medio (véase el § 2.1.4, ecuación (7)). En este caso no se utiliza el diagrama de excitación del modelo sicoacústico porque la frecuencia no lineal de la transformación de Bark degradaría la estructura armónica.

### 4.8.1 $EHS_B$

La magnitud de la estructura armónica se mide identificando y midiendo el pico más grande en el espectro de la función de autocorrelación. La correlación se calcula como el coseno del ángulo entre dos vectores de acuerdo con la siguiente fórmula, donde $\vec{F}_0$ es el vector de error y $\vec{F}_t$ es el mismo vector retrasado en cierta cantidad. La longitud de la correlación corresponde al retardo máximo (es decir, 256 en el siguiente ejemplo).

$$C = \frac{\vec{F}_0 \cdot \vec{F}_t}{\left|\vec{F}_0\right| \cdot \left|\vec{F}_t\right|} \tag{87}$$

El máximo retraso para obtener la función de autorrelación de la mayor potencia de dos que es menor que la mitad del número de componente de frecuencia TRF que corresponde a 18 kHz.

Por ejemplo, a una velocidad de muestreo de 48 kHz y un tamaño de ventana TRF de 2 048 muestras, la componente TRF que corresponde a 18 kHz es $(18/24) \times 1\,024 = 768$. Por lo tanto, el retraso máximo es 384. El número real de retrasos es 256, que es la mayor potencia de dos que es inferior a 384. El primer valor de la función de correlación se obtendría alineando $F_t[0]$ con $F_0[0]$ y el último, alineando $F_t[0]$ con $F_0[255]$.

Al vector resultante de correlaciones se le aplica una ventana Hann normalizada y, tras excluir la componente de continua sustrayendo el valor medio, se calcula el espectro de potencia con una TRF. El valor de cresta máximo del espectro después del primer valle identifica la frecuencia dominante en la función de autocorrelación. El valor medio de este máximo a lo largo de las tramas multiplicado por 1 000,0 es la estructura armónica del error (EHS, *error harmonic structure*).

## 5 Promediado

### 5.1 Promediado espectral

Si no se indica otra cosa en las descripciones de las variables de salida del modelo (véase el § 4), se utiliza el siguiente algoritmo cuando se promedian los valores locales en las bandas de frecuencias.

### 5.1.1 Promedio lineal

El valor del promedio lineal se calcula por:

$$AvgS = \frac{1}{Z} \cdot \sum_{k=0}^{Z-1} S[k] \tag{88}$$

donde $S$ representa el nombre de la variable de salida del modelo y $Z$ es el número de grupos de frecuencias.

### 5.2 Promediado temporal

Si no se indica otra cosa en las descripciones de las variables de salida del modelo (véase el § 4), se utiliza uno o varios de los siguientes algoritmos cuando los valores momentáneos son promediados en el tiempo. El factor de ponderación temporal (caso de aplicarse) se indica mediante el símbolo *W* y *Z* es el número de bandas de frecuencias.

#### 5.2.1 Promedio lineal

El valor del promedio lineal (prefijo «*Avg*») se calcula por:

$$AvgX = \frac{1}{N} \cdot \sum_{n=0}^{N-1} X[n] \tag{89}$$

donde *X* representa el nombre de la variable de salida del modelo y *N* es el número de muestras temporales para las cuales se han calculado valores momentáneos de *X*.

Cuando se aplica ponderación temporal (véase el § 4.2), para el cálculo del promedio lineal no se utiliza la anterior fórmula, sino la siguiente:

$$AvgX = \frac{\sum_{n=0}^{N-1} W[n] \cdot X[n]}{\sum_{n=0}^{N-1} W[n]} \tag{90}$$

#### 5.2.2 Media cuadrática

El valor medio cuadrático (prefijo «*Rms*») se calcula por:

$$RmsX = \sqrt{\frac{1}{N} \cdot \sum_{n=0}^{N-1} X[n]^2} \tag{91}$$

donde *X* representa el nombre de la variable de salida del modelo y *N* es el número de muestras temporales para las cuales se han calculado valores momentáneos de *X*.

Cuando se aplica ponderación temporal (véase el § 4.2), para el cálculo de la media cuadrática no se utiliza la anterior fórmula, sino la siguiente:

$$RmsX = \sqrt{Z} \cdot \sqrt{\frac{\sum_{n=0}^{N-1} W[n]^2 \cdot X[n]^2}{\sum_{n=0}^{N-1} W[n]^2}} \tag{92}$$

#### 5.2.3 Promedio enventanado

El valor promedio enventanado (prefijo «*Win*») se calcula por:

$$WinX = \sqrt{\frac{1}{N-L+1} \cdot \sum_{n=L-1}^{N-1} \left( \frac{1}{L} \cdot \sum_{i=0}^{L-1} \sqrt{X[n-i]} \right)^4} \tag{93}$$

donde *X* representa el nombre de la variable de salida del modelo y *N* es el número de muestras temporales para las cuales se han calculado valores momentáneos de *X*, y *L* es la longitud de la ventana temporal deslizante en las muestras temporales. La longitud de la ventana es de aproximadamente 100 ms, es decir, *L* es *4* para el modelo de oído basado en TRF y *25* para el modelo de oído basado en banco de filtros.

### 5.2.4 Selección de trama

#### 5.2.4.1 Promediado diferido

Para las variables de salida del modelo que utilizan este criterio, los valores calculados durante los primeros 0,5 s de la medición no se tienen en cuenta en el promediado temporal. El *promediado diferido* se utiliza para todas las variables siguientes:

WinModDiff1, AvgModDiff1, AvgModDiff2, RmsNoiseLoudness, RmsNoiseLoudAsym, RmsModDiff, AvgLinDist.

#### 5.2.4.2 Umbral de sonoridad

Para las variables de salida del modelo que utilizan este criterio, todos los valores momentáneos calculados hasta 50 ms después de que la sonoridad global de uno de los canales de audio correspondientes haya alcanzado un valor de $N_{Thres}$ sonios para la señal de prueba y la señal de referencia no se tienen en cuenta en el promediado temporal. *El umbral de sonoridad* se utiliza solamente para las variables de salida del modelo descritas en el § 4.3.

#### 5.2.4.3 Umbral de energía

Cuando la energía de la última mitad de una trama de 2 048 muestras es de menos de 8 000*, en el canal mono o tanto en el canal izquierdo como en el canal derecho de la referencia y de los datos de prueba, dicha trama no se toma en consideración. Las tramas tienen una superposición del 50% y sólo se evalúa la mitad de la trama que contiene nuevos datos. La aplicación de este criterio evita que se procesen tramas que contengan muy poca energía.

Este criterio sólo se utiliza para la variable de salida del modelo descrita en el § 4.8.

#### 5.2.4.4 Datos límites

Si el fichero procesado contiene ruido antes o después de datos ficheros de referencia legítimos, el error relativo puede ser muy grande ya que el nivel de referencia tiende a $-\infty$. Cuando este error se considera un artefacto, puede ser ignorado aplicando el criterio de rechazo de datos límite.

Cuando los ficheros son abiertos por primera vez, se identifican las ubicaciones de comienzo y de fin de los datos reales. El comienzo o el fin de datos se define como la primera ubicación, explorando desde el principio o desde el final del fichero, donde la suma de los valores absolutos en cinco muestras sucesivas es superior a 200 en uno de los canales de audio correspondientes. Las tramas que caigan fuera de esta gama serán subsiguientemente ignoradas.

Este criterio se utiliza para todas las variables de salida del modelo.

### 5.3 Promediado en canales de audio

Cuando no se indique otra cosa, en el caso de señales estereofónicas, los valores MOV para los canales izquierdo y derecho son promediados linealmente después del promediado temporal.

---

* Este número se refiere a los datos de entrada con un formato de número entero de 16 bits que se encuentra en la gama comprendida entre −32 768 y 32 767, como se utiliza en los discos compactos.

## 6 Estimación de la calidad de audio básica percibida

La *calidad de audio básica percibida* se estima haciendo corresponder varias variables de salida del modelo a un número único, utilizando una estructura de red neuronal artificial con una capa oculta.

### 6.1 Red neuronal artificial

La función de activación de la red neuronal es un sigmoide asimétrico:

$$\mathrm{sig}(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{94}$$

La red utiliza $I$ entradas y $J$ nodos en la capa oculta. La correspondencia se define por un conjunto de factores de escalaje de entrada $a_{mín}[i]$, $a_{máx}[i]$, un conjunto de pesos de entrada $w_x[i]$, un conjunto de pesos de salida $w_y[j]$ y un par de factores de escalaje de salida $b_{mín}$ y $b_{máx}$. Las entradas se han hecho corresponder a un índice de distorsión

$$DI = w_y[J] + \sum_{j=0}^{J-1} \left( w_y[j] \cdot \mathrm{sig} \left( w_x[I, j] + \sum_{i=0}^{I-1} w_x[i, j] \cdot \frac{x[i] - a_{mín}[i]}{a_{máx}[i] - a_{mín}[i]} \right) \right) \tag{95}$$

que está directamente relacionado con la *calidad de audio básica percibida* en términos de un grado de diferencia objetiva (ODG). La relación entre el *índice de distorsión* y el *grado de diferencia objetiva* viene dada por

$$ODG = b_{mín} + (b_{máx} - b_{mín}) \cdot \mathrm{sig}(DI) \tag{96}$$

### 6.2 Versión básica

La versión básica emplea el modelo de oído basado en TRF. Utiliza las variables de salida del modelo: *BandwidthRef*$_B$, *BandwidthTest*$_B$, *Total NMR*$_B$, *WinModDiff1*$_B$, *ADB*$_B$, *EHS*$_B$, *AvgModDiff1*$_B$, *AvgModDiff2*$_B$, *RmsNoiseLoud*$_B$, *MFPD*$_B$ *y RelDistFrames*$_B$. Estas 11 variables de salida del modelo se han hecho corresponder a un índice de calidad único utilizando una red neuronal descrita en el § 6.1 con tres nodos en la capa oculta. Los parámetros de la correspondencia se indican en los Cuadros 12 a 16.

CUADRO 12

**Variables de salida del modelo utilizadas en la versión básica**

| Variable de salida del modelo (MOV) | Finalidad |
|---|---|
| $WinModDiff1_B$ | Cambios en modulación (relacionados con la aspereza (*roughness*)) |
| $AvgModDiff1_B$ | |
| $AvgModDiff2_B$ | |
| $RmsNoiseLoud_B$ | Sonoridad de la distorsión |
| $BandwidthRef_B$ | Distorsiones lineales (respuesta a frecuencia, etc.) |
| $BandwidthTest_B$ | |
| $RelDistFrames_B$ | Frecuencia de distorsiones audibles |
| Total $NMR_B$ | Relación ruido-enmascaramiento |
| $MFPD_B$ | Probabilidad de detección |
| $ADB_B$ | |
| $EHS_B$ | Estructura armónica del error |

CUADRO 13

**Factores de escalaje para las entradas de la versión básica**

| índice (*i*) | MOV (x[i]) | $a_{mín}[i]$ | $a_{máx}[i]$ |
|---|---|---|---|
| 0 | $BandwidthRef_B$ | 393,916656 | 921 |
| 1 | $BandwidthTest_B$ | 361,965332 | 881,131226 |
| 2 | Total $NMR_B$ | −24,045116 | 16,212030 |
| 3 | $WinModDiff1_B$ | 1,110661 | 107,137772 |
| 4 | $ADB_B$ | −0,206623 | 2,886017 |
| 5 | $EHS_B$ | 0,074318 | 13,933351 |
| 6 | $AvgModDiff1_B$ | 1,113683 | 63,257874 |
| 7 | $AvgModDiff2_B$ | 0,950345 | 1 145,018555 |
| 8 | $RmsNoiseLoud_B$ | 0,029985 | 14,819740 |
| 9 | $MFPD_B$ | 0,000101 | 1 |
| 10 | $RelDistFrames_B$ | 0 | 1 |

CUADRO 14

**Pesos para los nodos de entrada de la versión básica**

| índice (*i*) | MOV (x[i]) | nodo 1 ($w_x[i,0]$) | nodo 2 ($w_x[i,1]$) | nodo 3 ($w_x[i,2]$) |
|---|---|---|---|---|
| 0 | $BandwidthRef_B$ | −0,502657 | 0,436333 | 1,219602 |
| 1 | $BandwidthTest_B$ | 4,307481 | 3,246017 | 1,123743 |
| 2 | Total $NMR_B$ | 4,984241 | −2,211189 | −0,192096 |
| 3 | $WinModDiff1_B$ | 0,051056 | −1,762424 | 4,331315 |
| 4 | $ADB_B$ | 2,321580 | 1,789971 | −0,754560 |
| 5 | $EHS_B$ | −5,303901 | −3,452257 | −10,814982 |
| 6 | $AvgModDiff1_B$ | 2,730991 | −6,111805 | 1,519223 |
| 7 | $AvgModDiff2_B$ | 0,624950 | −1,331523 | −5,955151 |
| 8 | $RmsNoiseLoud_B$ | 3,102889 | 0,871260 | −5,922878 |
| 9 | $MFPD_B$ | −1,051468 | −0,939882 | −0,142913 |
| 10 | $RelDistFrames_B$ | −1,804679 | −0,503610 | −0,620456 |
| 11 | bias | −2,518254 | 0,654841 | −2,207228 |

CUADRO 15

**Pesos para el nodo de salida de la versión básica**

| nodo 1 ($w_y[0]$) | nodo 2 ($w_y[1]$) | nodo 3 ($w_y[2]$) | bias ($w_y[3]$) |
|---|---|---|---|
| −3,817048 | 4,107138 | 4,629582 | −0,307594 |

CUADRO 16

**Factores de escalaje para la salida de la versión básica**

| | $b_{mín}$ | $b_{máx}$ |
|---|---|---|
| ODG | −3,98 | 0,22 |

## 6.3 Versión avanzada

La versión avanzada emplea tanto el modelo de oído basado en TRF como el modelo de oído basado en banco de filtros. Utiliza las variables de salida del modelo $RmsModDiff_A$, $RmsNoiseLoudAsym_A$, $AvgLinDist_A$, Segmental NMRB y EHSB. Estas 5 variables de salida del modelo se han hecho corresponder a un índice de calidad único, utilizando una red neuronal descrita en el § 6.1 con cinco nodos en la capa oculta. Los parámetros de la correspondencia se indican en los Cuadros 17 a 21.

CUADRO 17

**Variables de salida del modelo utilizadas en la versión avanzada**

| Variable de salida del modelo (MOV) | Finalidad |
|---|---|
| $RmsNoiseLoudAsym_A$ | Sonoridad de la distorsión |
| $RmsModDiff_A$ | Cambios en modulación (relacionados con la aspereza (*roughness*)) |
| $AvgLinDist_A$ | Distorsiones lineales (respuesta a frecuencia, etc.) |
| Segmental $NMR_B$ | Relación ruido-enmascaramiento |
| $EHS_B$ | Estructura armónica del error |

CUADRO 18

**Factores de escalaje para los nodos de entrada de la versión avanzada**

| índice (*i*) | MOV (x[i]) | $a_{mín}[i]$ | $a_{máx}[i]$ |
|---|---|---|---|
| 0 | $RmsModDiff_A$ | 13,298751 | 2166,5 |
| 1 | $RmsNoiseLoudAsym_A$ | 0,041073 | 13,24326 |
| 2 | Segmental $NMR_B$ | −25,018791 | 13,46708 |
| 3 | $EHS_B$ | 0,061560 | 10,226771 |
| 4 | $AvgLinDist_A$ | 0,024523 | 14,224874 |

CUADRO 19

**Pesos para las entradas de la versión avanzada**

| índice (*i*) | MOV (x[i]) | nodo 1 ($w_x[i,0]$) | nodo 2 ($w_x[i,1]$) | nodo 3 ($w_x[i,2]$) | nodo 4 ($w_x[i,3]$) | nodo 5 ($w_4[i,4]$) |
|---|---|---|---|---|---|---|
| 0 | $RmsModDiff_A$ | 21,211773 | −39,913052 | −1,382553 | −14,545348 | −0,320899 |
| 1 | $RmsNoiseLoudAsym_A$ | −8,981803 | 19,956049 | 0,935389 | −1,686586 | −3,238586 |
| 2 | Segmental $NMR_B$ | 1,633830 | −2,877505 | −7,442935 | 5,606502 | −1,783120 |
| 3 | $EHS_B$ | 6,103821 | 19,587435 | −0,240284 | 1,088213 | −0,511314 |
| 4 | $AvgLinDist_A$ | 11,556344 | 3,892028 | 9,720441 | −3,287205 | −11,031250 |
| 5 | bias | 1,330890 | 2,686103 | 2,096598 | −1,327851 | 3,087055 |

CUADRO 20

**Pesos para el nodo de salida de la versión avanzada**

| nodo 1 (wx[i,0]) | nodo 2 ($w_x[i,1]$) | nodo 3 ($w_x[i,2]$) | nodo 4 ($w_x[i,3]$) | nodo 5 ($w_4[i,4]$) | bias ($w_y[5]$) |
|---|---|---|---|---|---|
| −4,696996 | −3,289959 | 7,004782 | 6,651897 | 4,009144 | −1,360308 |

CUADRO 21

**Factores de escalaje para la salida de la versión avanzada**

| | $b_{mín}$ | $b_{máx}$ |
|---|---|---|
| ODG | −3,98 | 0,22 |

## 7 Conformidad de las realizaciones

### 7.1 Generalidades

En este punto se presenta un conjunto de elementos de prueba para verificar la correcta realización del método.

### 7.2 Selección

Los elementos de prueba se seleccionaron tomándolos de la base de datos 3 (DB3), que se utilizaron para la validación de los modelos. Para simplificar las pruebas se seleccionó un subconjunto de los 84 elementos de la DB3. Este subconjunto está formado por 16 elementos. El criterio de mayor peso para la selección fue que las variables de salida del modelo (MOV) y los valores de índice de distorsión (DI, *distorsion index*) resultantes abarcaran una amplia gama.

### 7.3 Fijación de valores para la prueba de conformidad

Los elementos de prueba se pueden descargar en ficheros WAV (formato RIFF de Microsoft). Todos los elementos fueron muestreados a 48 kHz, MIC 16 bits. Las señales de referencia y de prueba tal como son proporcionadas por la UIT están adaptadas entre sí en tiempo y nivel, por lo que no se requiere ninguna compensación adicional de ganancia o de tiempo de propagación. El algoritmo de medición debe ajustarse a un nivel de escucha de 92 dB SPL.

### 7.4 Intervalo de tolerancia aceptable

Para ser conformes con la presente Recomendación, los valores DI calculados deben reproducir los valores indicados en los Cuadros 22 y 23, con una tolerancia comprendida en ±0,02 [6] para todos los elementos de prueba. Una realización cuyos resultados no estén dentro de esta tolerancia no es conforme con esta Recomendación.

### 7.5 Elementos de prueba

Los siguientes Cuadros indican el nombre de los elementos de referencia y de prueba[7], los números de pista y los valores DI resultantes. El Cuadro 22 indica los valores relacionados con la versión básica, y el Cuadro 23 los relacionados con la versión avanzada.

---

[6] Para conseguir esta exactitud es necesario utilizar la aritmética de coma flotante del IEEE.

[7] Los nombres de los correspondientes elementos de referencia se obtienen sustituyendo la subcadena «cod» en los nombres de los elementos de prueba por «ref», por ejemplo, el elemento de referencia para «bcodtri.wav» es «breftri.wav».

CUADRO 22

**Elementos de prueba y valores DI resultantes para la versión básica**

| Elemento | DI | ODG |
|---|---|---|
| acodsna.wav | 1,304 | −0,676 |
| bcodtri.wav | 1,949 | −0,304 |
| ccodsax.wav | 0,048 | −1,829 |
| ecodsmg.wav | 1,731 | −0,412 |
| fcodsb1.wav | 0,677 | −1,195 |
| fcodtr1.wav | 1,419 | −0,598 |
| fcodtr2.wav | −0,045 | −1,927 |
| fcodtr3.wav | −0,715 | −2,601 |
| gcodcla.wav | 1,781 | −0,386 |
| icodsna.wav | −3,029 | −3,786 |
| kcodsme.wav | 3,093 | 0,038 |
| 1codhrp.wav | 1,041 | −0,876 |
| 1codpip.wav | 1,973 | −0,293 |
| mcodcla.wav | −0,436 | −2,331 |
| ncodsfe.wav | 3,135 | 0,045 |
| scodclv.wav | 1,689 | −0,435 |

CUADRO 23

**Elementos de prueba y valores DI resultantes para la versión avanzada**

| Elemento | DI | ODG |
|---|---|---|
| acodsna.wav | 1,632 | −0,467 |
| bcodtri.wav | 2,000 | −0,281 |
| ccodsax.wav | 0,567 | −1,300 |
| ecodsmg.wav | 1,594 | −0,489 |
| fcodsb1.wav | 1,039 | −0,877 |
| fcodtr1.wav | 1,555 | −0,512 |
| fcodtr2.wav | 0,162 | −1,711 |
| fcodtr3.wav | −0,783 | −2,662 |
| gcodcla.wav | 1,457 | −0,573 |
| icodsna.wav | −2,510 | −3,664 |
| kcodsme.wav | 2,765 | −0,029 |
| lcodhrp.wav | 1,538 | −0,523 |
| lcodpip.wav | 2,149 | −0,219 |
| mcodcla.wav | 0,430 | −1,435 |
| ncodsfe.wav | 3,163 | 0,050 |
| scodclv.wav | 1,972 | −0,293 |

# Adjunto 1
# al Anexo 2

# Proceso de validación

## 1 Generalidades

En 1994, el UIT-R adoptó la Cuestión UIT-R 210/10 – Métodos objetivos para la evaluación de la calidad percibida, y estableció un Grupo Especial para dicha Cuestión. Uno de los primeros pasos fue convocar a la presentación de propuestas de modelos, y se recibieron respuestas de seis proponentes de modelos.

Se trabajó intensamente en la definición de los procedimientos para la evaluación. Se consideró conveniente compilar una primera base de datos, la base de datos 1 (DB1), formada por material tomado de pruebas de escucha ya realizadas. El interés principal estaba concentrado en la calidad de audio media y alta, por lo que sólo se consideraron los resultados de pruebas de escucha de conformidad con la Recomendación UIT-R BS.1116. El material de estas pruebas representaba material de radiodifusión crítico para códecs de baja velocidad binaria como MPEG1 Capa II, MPEG1 Capa III, Dolby AC2, Mini Disc, NICAM. La base de datos 1 se creó para proporcionar a los proponentes de modelos una plataforma común constituida por material que abarcara una amplia gama de factores de degradación, una diversidad de códecs y la degradación introducida por varios códecs dispuestos en cascada. En el Adjunto 2 al Anexo 2 se presenta una descripción detallada de las pruebas compiladas en la base de datos 1.

Evidentemente, un método de medición objetiva de la calidad de audio percibida que imite el comportamiento humano sólo puede ser validado sobre una base de datos que contenga resultados de pruebas subjetivas. Una validación adecuada requiere una base de datos que se funde sobre un material desconocido. Por esta razón fue necesario llevar a cabo nuevas pruebas de escucha. Puesto que, en un orden ideal, el método de medición debe comprender cualquier tipo de artefacto que pudiera aparecer en aplicaciones de radiodifusión, deben incluirse no solamente artefactos de codificación, sino también artefactos tradicionales como la distorsión y el ruido. Las bases de datos 2 (DB2) y 3 (DB3) se crearon en 1996 y 1997 respectivamente para responder a estas exigencias. Además de los códecs ya incluidos en DB1, también fueron incluidos Dolby-AC3 y AAC. Para más detalles, véase el Adjunto 2 al Anexo 2.

La validación debe tener en cuenta estas incertidumbres, a menudo presentadas como intervalos de confianza, inherentes a las pruebas de escucha subjetivas. El tamaño del intervalo de confianza depende de varios factores. Entre éstos, los más importantes son la experiencia de los sujetos, los procedimientos de adiestramiento, el contexto en que se presentan los elementos (o ítems) de prueba, y el número de sujetos.

La adaptación y la validación del método objetivo presentadas en esta Recomendación se basan en un «oyente experto promedio». Los valores medios obtenidos de las evaluaciones subjetivas de la calidad junto con los intervalos de confianza del 95% se utilizan para caracterizar al «oyente experto promedio».

Las pruebas de escucha subjetivas son muy sensibles a diversos factores que influyen en los resultados. Los SDG para las bases de datos 2 y 3 se produjeron en tres lugares de prueba diferentes, y se realizaron varios estudios para determinar si los datos podían realmente combinarse. Aunque no en todos los estudios se llegó a las mismas conclusiones, se consideró razonable fusionar los datos, y esta base de datos refundida se utilizó como base para la validación.

El proceso de validación se dividió en tres fases:

– Fase 1: Fase competitiva

– Fase 2: Fase colaborativa

– Fase 3: Fase de selección final

Estas fases de describirán detalladamente en las secciones siguientes.

## 2 Fase competitiva

Para la medición objetiva de la calidad de audio percibida se propusieron seis métodos (DIX, NMR, PAQM, PERCEVAL, POM, TTA) y se decidió comparar sus comportamientos utilizando la base de datos 2 y un subconjunto de la base de datos 1. La base de datos 2 se creó a principios de 1996. La selección del material de prueba final se efectuó gracias a trabajos realizados conjuntamente por SR (Suecia) y BBC (Reino Unido). Las pruebas de escucha se efectuaron en NRK en Noruega, DR en Dinamarca y NHK en Japón. Deutsche Telekom (Alemania) y Teracom (Suecia) prepararon un análisis estadístico de los datos obtenidos en las pruebas. En la fase 1, los datos objetivos fueron generados en un lugar neutral (Swisscom, Suiza). Los proponentes de modelos recibieron entonces la primera mitad de la base de datos 2 con miras a una adaptación final de los métodos (fase 2). Por último, se generaron nuevos grados de diferencia objetiva (ODG) en Swisscom.

Los análisis de los comportamientos de los métodos fueron realizados por Teracom (Suecia), así como por los propios proponentes. Aunque los resultados de algunos de los métodos propuestos mostraban una alta correlación con los SDG, hubo consenso en cuanto a que ninguno de los métodos propuestos era apreciablemente mejor que los demás ni cumplía íntegramente los requisitos deseados. En consecuencia, se decidió desarrollar un método de medición mejorado mediante el trabajo en común de todos los actuales proponentes de modelos. El comportamiento del nuevo método deberá compararse con el de uno de los métodos ya establecidos, que se ha llamado modelo B3.

## 3 Fase colaborativa

La fase colaborativa tenía por objeto combinar los mejores elementos de los diferentes métodos para formar un nuevo método. Para responder mejor a las necesidades de los usuarios se decidió crear dos versiones del método. Una, apropiada para realizaciones en tiempo real, y otra que puede requerir un mayor poder de cálculo para alcanzar una mayor exactitud.

El procedimiento de validación para las dos versiones del nuevo método se diseñó de manera similar al de la fase competitiva. Hubo que crear una nueva base de datos (la base de datos 3, o DB3) Los elementos (o ítems) y condiciones fueron finalmente definidos en la primavera de 1997, y compilados en SR, Swisscom y BBC. Una descripción completa de la base de datos puede encontrarse en el Adjunto 2 al Anexo 2. Las pruebas de escucha subjetivas se efectuaron en tres lugares de prueba: Deutsche Telekom, NHK y SR. Todos los lugares de prueba aplicaron el «método ciego doble, referencia oculta, triple estímulo», descrito en la Recomendación UIT-R BS.1116. Los resultados de las pruebas de escucha se recopilaron en Suecia. En Teracom, así como en otras partes, se realizó un extenso análisis estadístico de los resultados de las pruebas de escucha. Como consecuencia de este análisis, algunos oyentes fueron excluidos de las ulteriores evaluaciones. Los resultados obtenidos en los lugares de prueba se combinaron para formar la base de datos 3.

En otoño de 1997, 52 elementos de los contenidos en la base de datos fueron liberados y proporcionados a los proponentes. Los nuevos métodos fueron adaptados a los nuevos datos. Como había varios juegos de valores de parámetros que daban resultados similares, la decisión sobre la selección final se tomó lo más tarde posible. Finalmente, en Suiza, los 32 elementos restantes se utilizaron para validar los nuevos métodos sobre un conjunto de datos «desconocido».

Además, los resultados de una nueva prueba de audición, realizada por CRC (Canadá) se utilizaron para validar los nuevos métodos sobre material «desconocido». Los procesos de selección y verificación se describen en los puntos siguientes.

## 4 Verificación

Se realizaron extensas pruebas de las 18 versiones especificadas del método de medición objetiva. En este punto se describen los criterios de selección, así como los resultados de la comparación de los SDG con los resultados obtenidos de las 18 versiones del método de medición. El objetivo era seleccionar y verificar las versiones óptimas que se recomendarán a la UIT.

### Criterios de selección

La correlación entre resultados subjetivos y objetivos es el criterio más natural para validar un método objetivo. Además, para la evaluación se han otros dos criterios que consideran la fiabilidad de los valores medios: la puntuación de error absoluta (AES, *absolute error score*) y el esquema de tolerancia.

La AES, introducida para relacionar la exactitud del modelo con la exactitud de la prueba de audición, se define por la siguiente expresión:

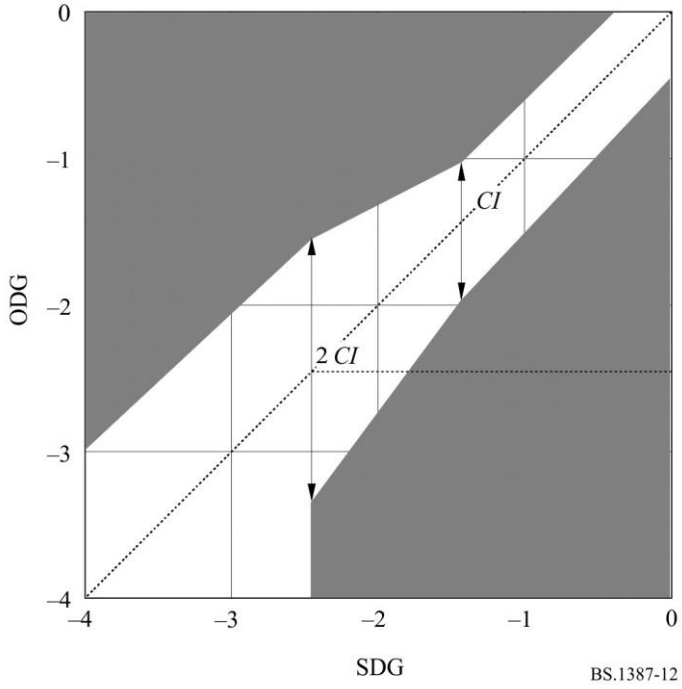
$$AES = 2 * \sqrt{\frac{\sum((ODG - SDG)/CI)^2}{N}} \qquad \text{si } CI < 0{,}25 \text{ entonces } CI = 0{,}25$$

donde *CI* es el intervalo de confianza.

El esquema de tolerancia fue concebido para permitir diferentes desviaciones de los ODG con respecto a los SDG en los extremos superior e inferior de la escala de degradación. La gama tolerada está relacionada con los intervalos de confianza de las pruebas de escucha. Esta gama está limitada a un valor mínimo de 0,25 grados. Las distancias de los ODG fuera del esquema de tolerancia, al esquema de tolerancia, se utilizaron para evaluar la calidad del método de medición.

FIGURA 12

**Esquema de tolerancia, intervalo de confianza $CI \geq 0{,}25$**




## 4.1 Comparación de valores SDG y ODG

Las mediciones objetivas se dividieron en tres fases. En la fase 1, los 84 elementos de prueba eran desconocidos para todos, salvo para el panel de selección. En la fase 2, se dio información sobre 52 elementos. La información comprendía los valores SDG y los extractos de audio reales. En la fase 3 se utilizó este conocimiento para optimizar el comportamiento de las versiones del método. Debe señalarse que en la fase 3 se probaron cuatro versiones más que en la fase 1. Los valores SDG presentados se calcularon a partir de los datos generados por 75 sujetos calificados.

Hay muchas formas diferentes de evaluar hasta qué punto los ODG reflejan adecuadamente los SDG. Desafortunadamente, no existe un valor único que refleje el comportamiento completo. Por consiguiente, hay que examinar un número de perspectivas. Las correlaciones se presentan en el § 4.2, y las AES se presentan en el § 4.3. El modelo B3 es uno de los probados por el UIT-R en 1996 y se había acordado que las diversas nuevas versiones se compararan con esta otra, más antigua.

## 4.2 Correlación

Los valores de correlación de las fases 1 y 3 se han representado en la Fig. 13 (84 elementos) y en la Fig. 14 (32 elementos).

FIGURA 13

**Correlación entre SDG y ODC. Se han incluido los 84 elementos**

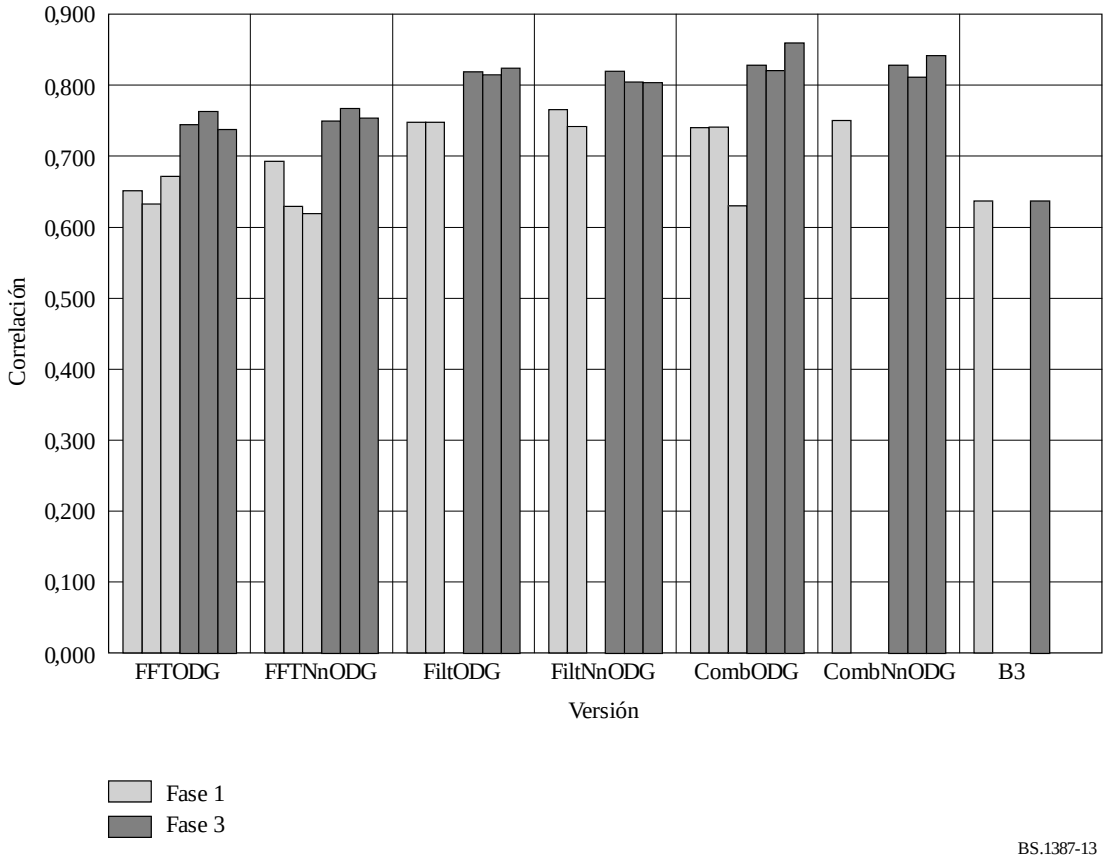
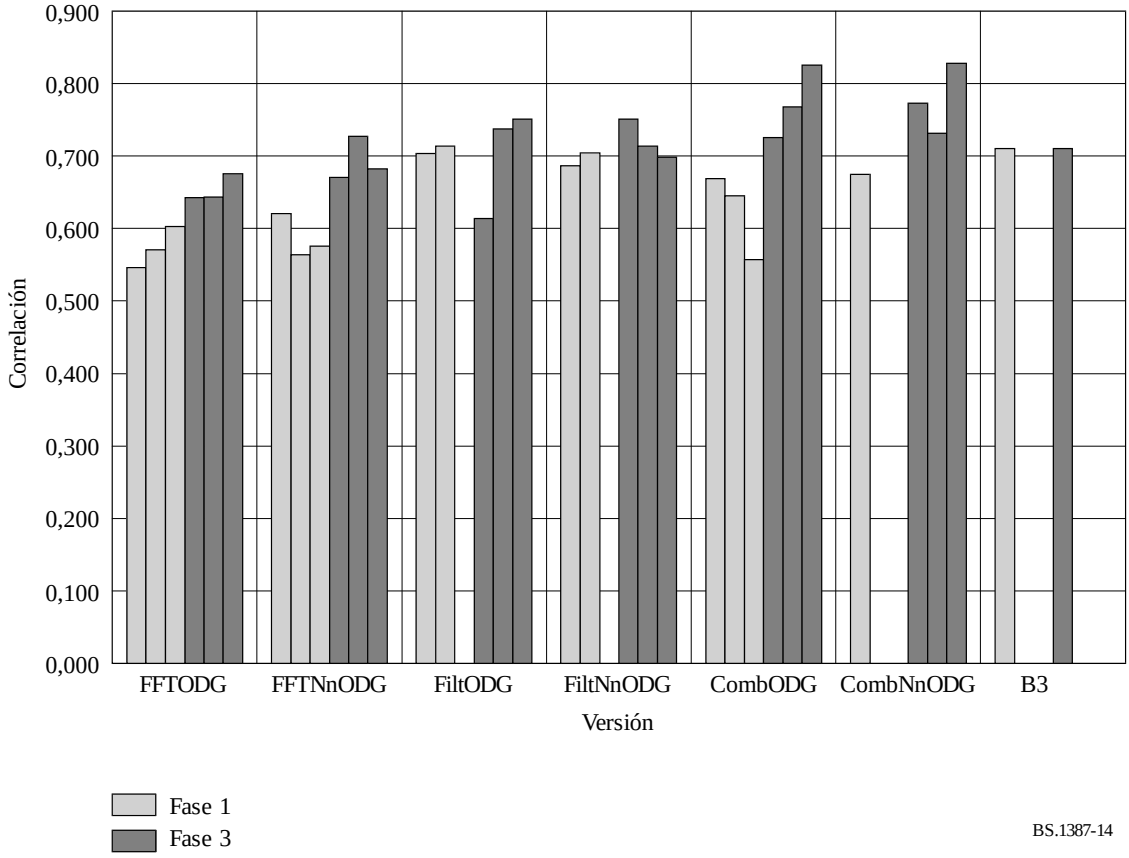


FIGURA 14

**Correlación entre SDG y ODC. Se han incluido los 32 elementos no liberados**




## 4.3 Puntuación de error absoluto (AES)

Un modelo que proporcione, en promedio, valores ODG dentro del intervalo de confianza SDG, dará un valor AES próximo a 2. Las Figs. 15 y 16 presentan una visión panorámica de los valores AES.

FIGURA 15

**AES para diferentes versiones. Se han incluido los 84 elementos**

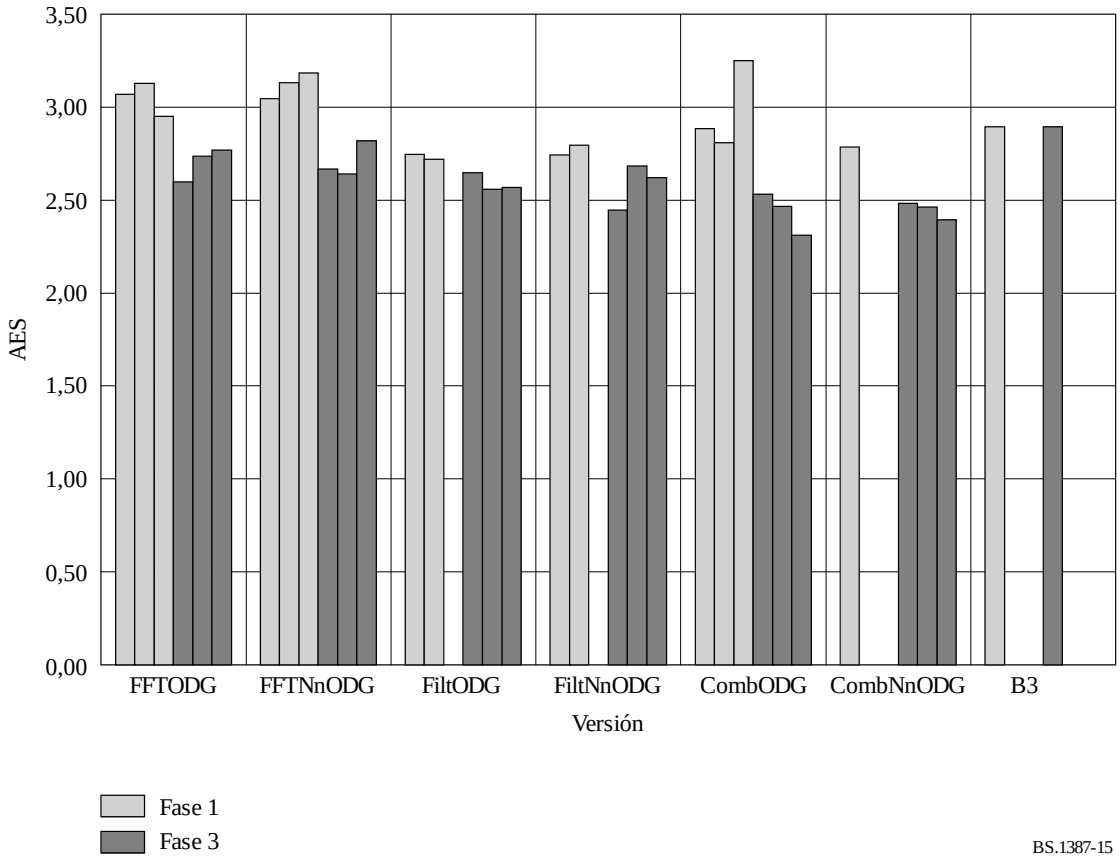
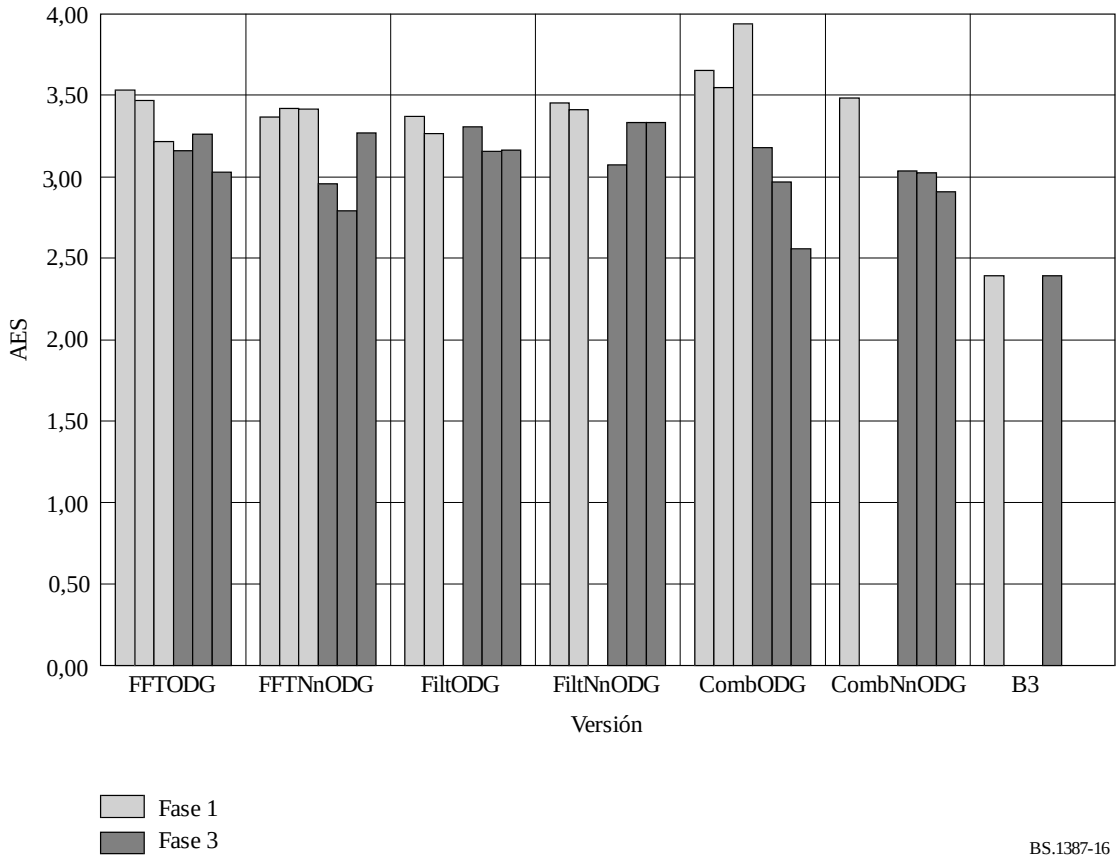


FIGURA 16

**AES para diferentes versiones. Se han incluido los 32 elementos no liberados**



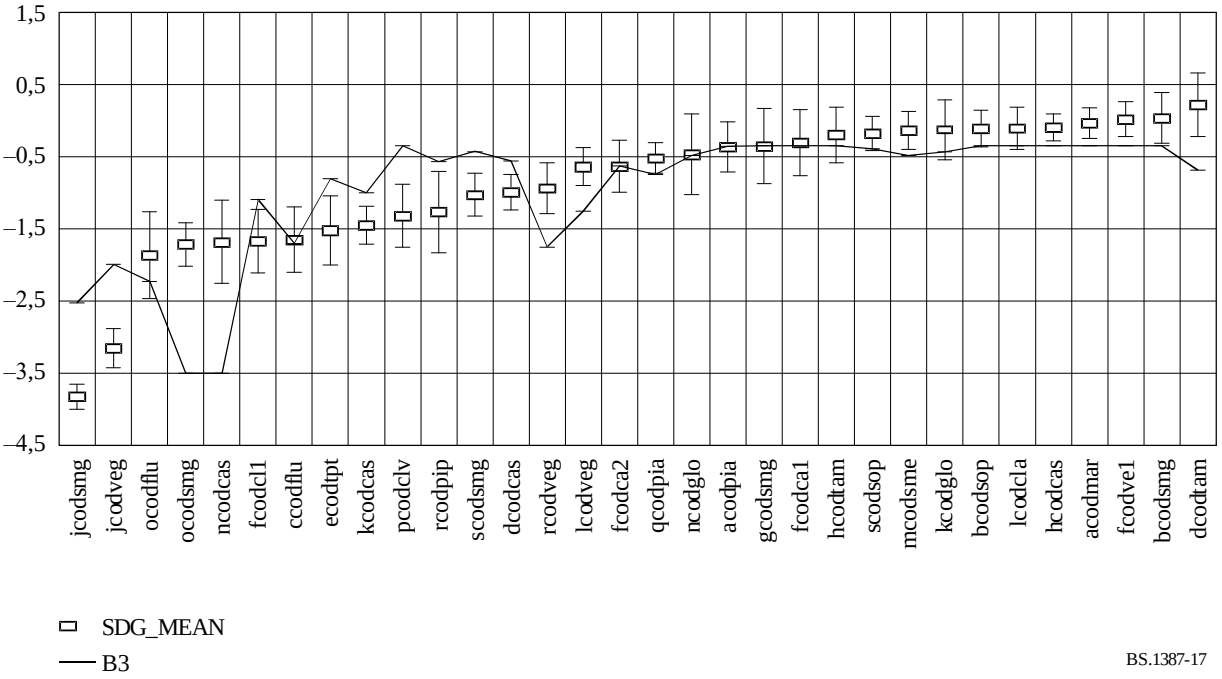

## 4.4 Comparación de ODG con respecto al intervalo de confianza

La mayor parte de las versiones se comportaron de manera similar, como era de esperar. En la reunión se presentó un gran número de gráficos, pero en esta Recomendación se ha utilizado un número limitado de ellos. Para una información mucho más detallada, véase el informe completo de las pruebas de verificación.

En las Figs. 17 a 22 se presentan los gráficos de los SDG medios, intervalo de confianza y ODG para los 32 elementos no liberados, con respecto al modelo B3 y a las versiones del modelo FFTNnODG1 y CombNnODG3.

FIGURA 17

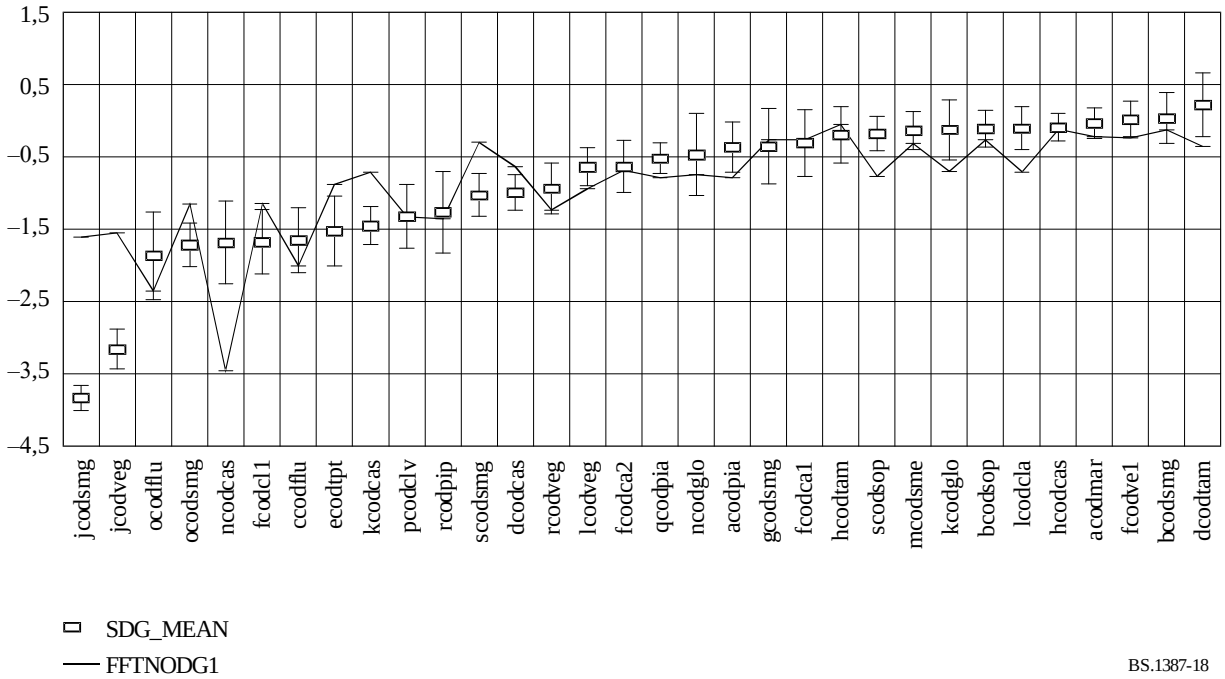
**Modelo B3: Gráfico de SDG medio, intervalo de confianza y ODG para los 32 elementos no liberados**




BS.1387-17

FIGURA 18

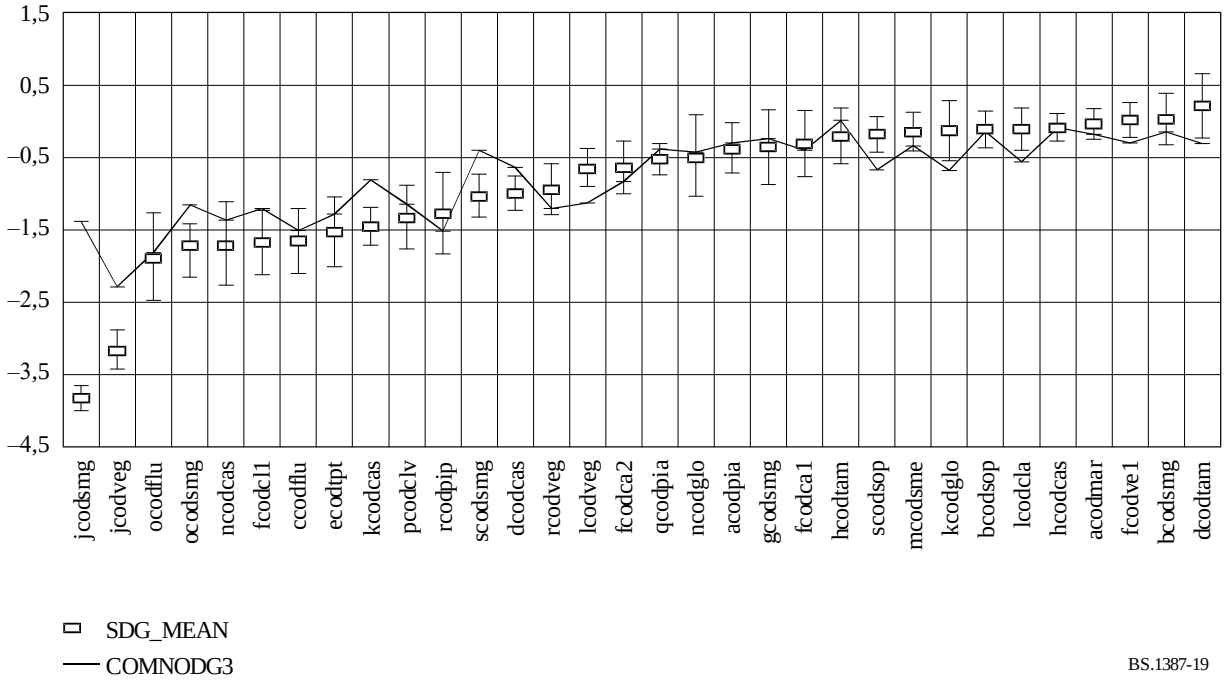
**FFTNnODG1: Gráfico de SDG medio, intervalo de confianza y ODG después de la tercera fase para los 32 elementos no liberados**




BS.1387-18

FIGURA 19

**CombNnODG3: Gráfico de SDG medio, intervalo de confianza y ODG después de la tercera fase para los 32 elementos no liberados**




En las Figs. 20 y 21 se presentan gráficos similares, pero en este caso para los 84 elementos durante la fase 3. Además, la Fig. 22 muestra el comportamiento de la versión CombNnODG3.

FIGURA 20

**Modelo B3: Gráfico de SDG medio, intervalo de confianza y ODG para los 84 elementos**

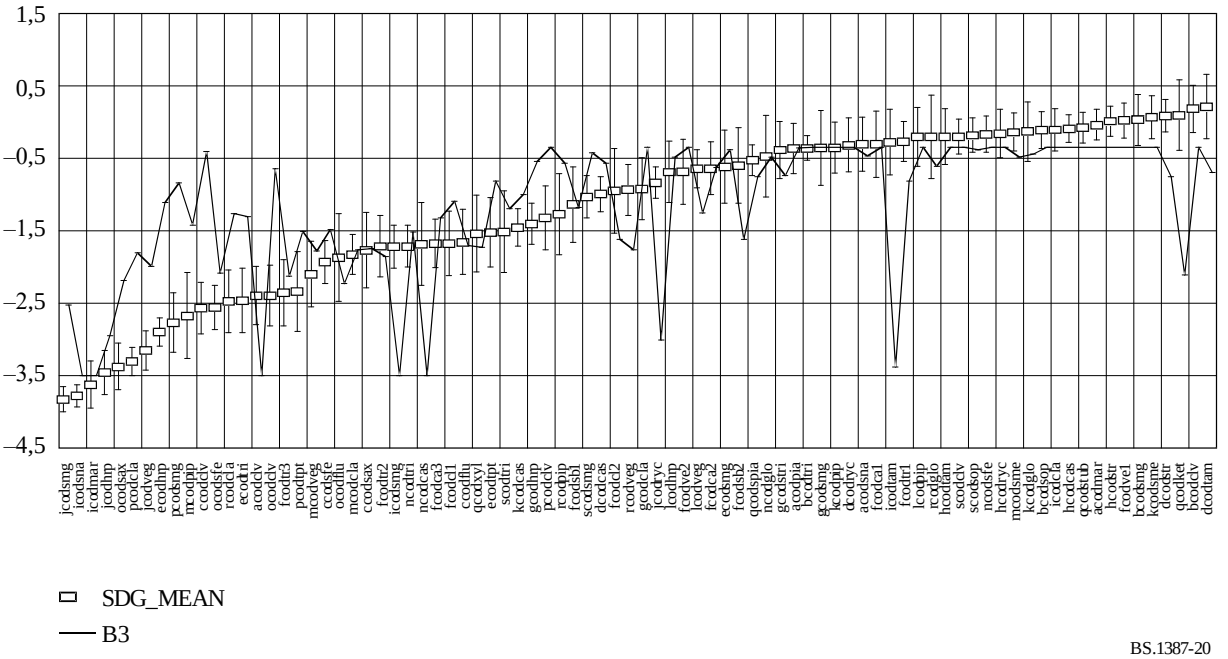


FIGURA 21

**FFTNnODG1: Gráfico de SDG medio, intervalo de confianza y ODG durante la fase 3 para los 84 elementos**

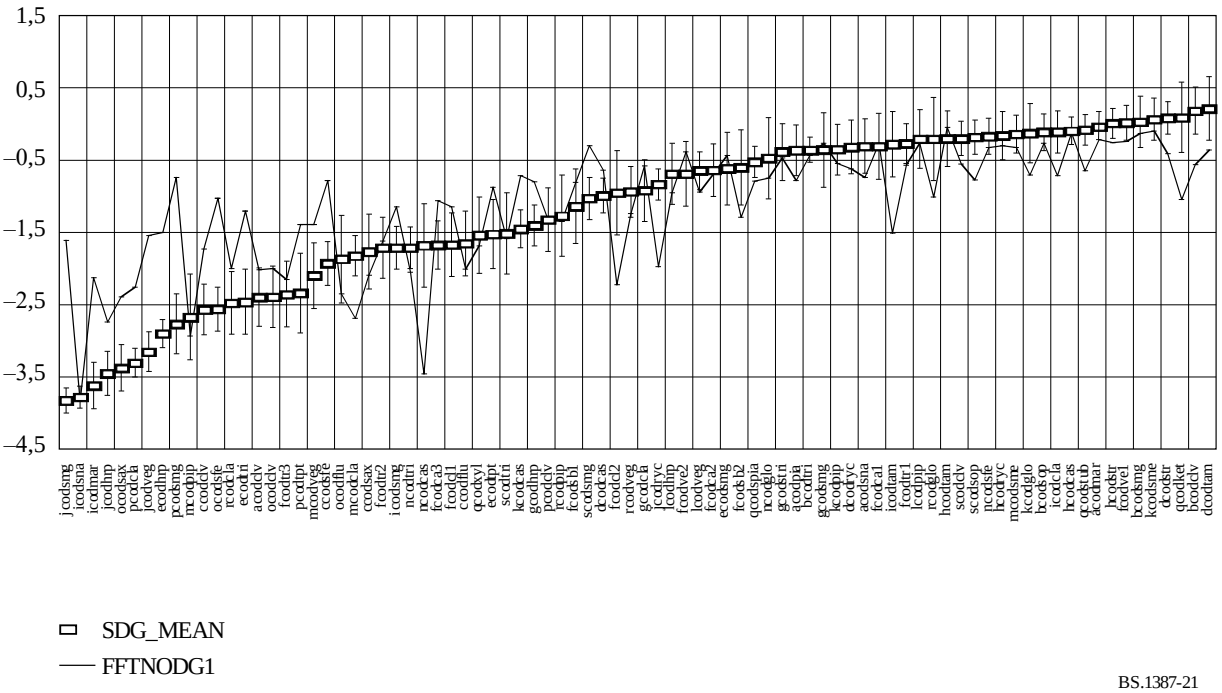
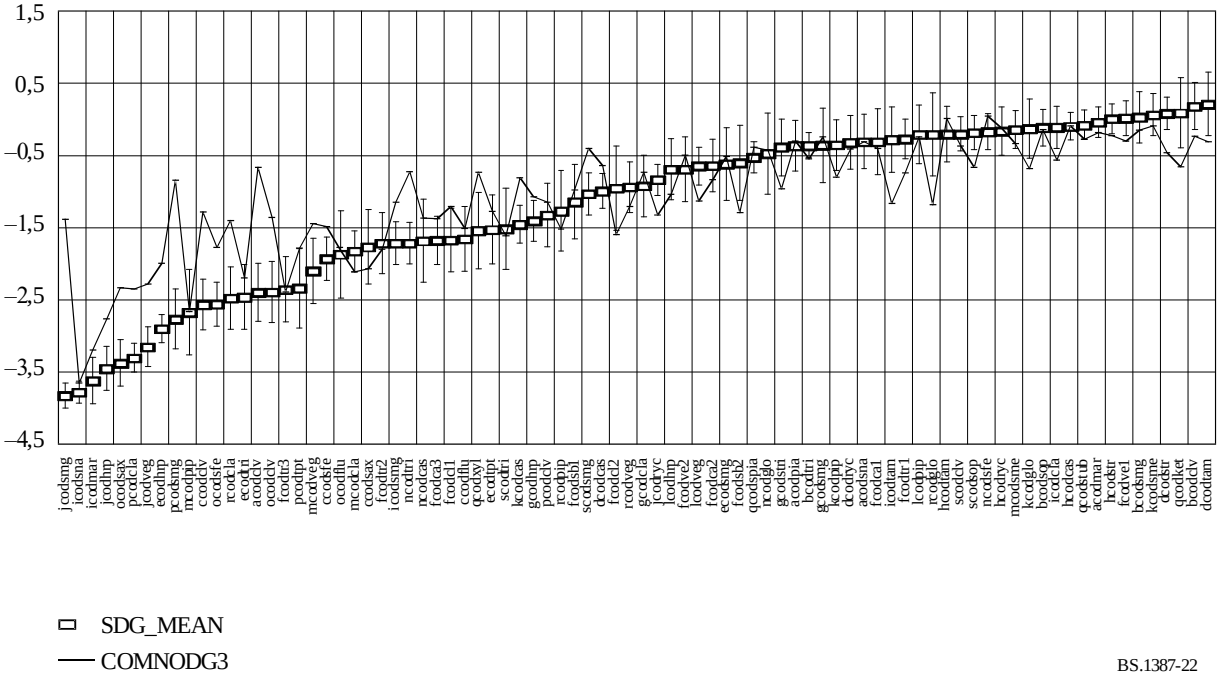



FIGURA 22

**CombNnODG3: Gráfico de SDG medio, intervalo de confianza y ODG durante la fase 3 para los 84 elementos**



## 4.5 Comparación de ODG con el intervalo de tolerancia

El UIT-R ha definido un requisito de usuario a alcanzar que puede hacerse corresponder a un intervalo de tolerancia. Los requisitos de usuario a alcanzar son más estrictos para los niveles más altos de la calidad de audio y menos estrictos para una calidad de audio más baja. Las Figuras siguientes ilustran el comportamiento, en esta dimensión, del modelo B3 y de las versiones FFTNnODG1 y CombNnODG3 para los 84 elementos durante la fase 3.

FIGURA 23

**B3: Gráfico de SDG medio, intervalo de tolerancia y ODG durante la fase 3 para los 84 elementos**

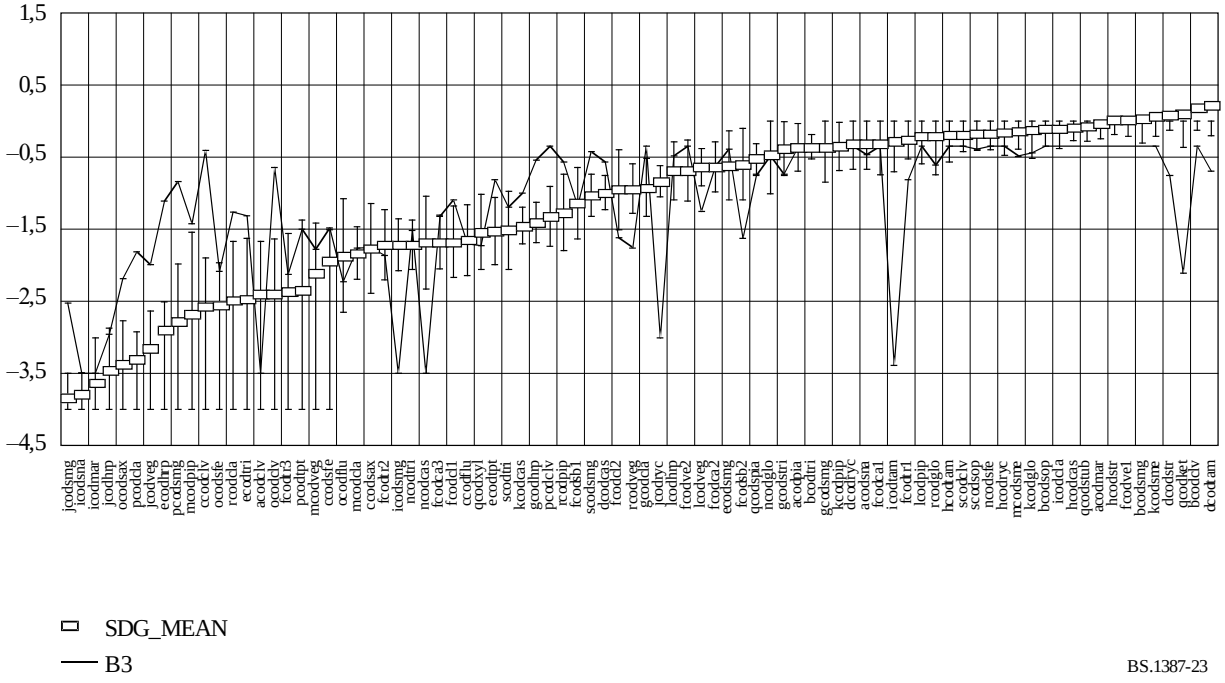


FIGURA 24

**FFTNnODG1: Gráfico de SDG medio, intervalo de tolerancia y ODG durante la fase 3 para los 84 elementos**

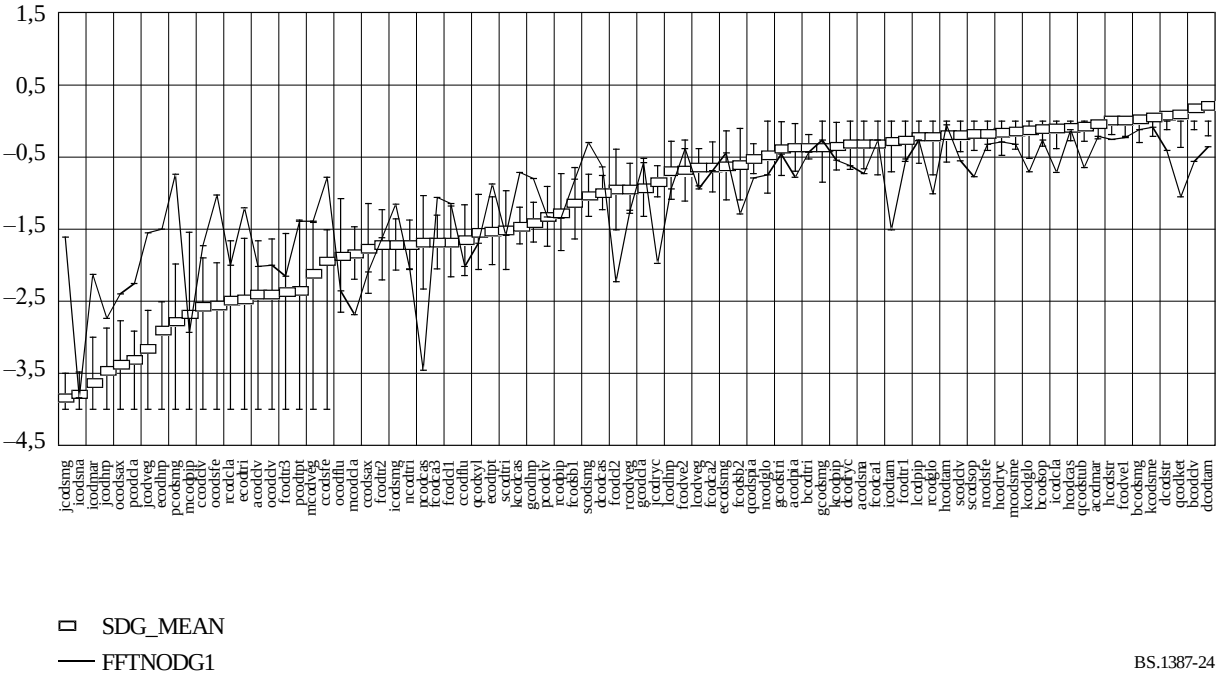
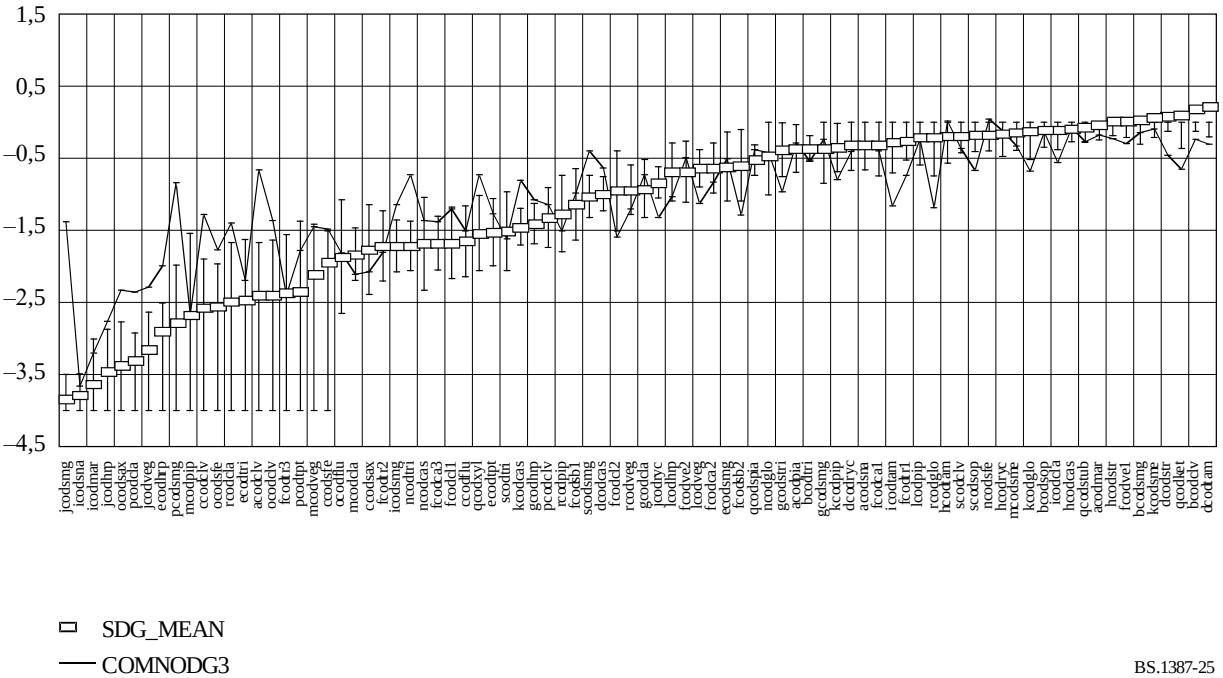



FIGURA 25

**CombNnODG3: Gráfico de SDG medio, intervalo de tolerancia y ODG durante la fase 3 para los 84 elementos**



## 5 Selección de las versiones de modelo óptimas

En el lugar en que se efectuó la prueba objetiva se presentaron 18 versiones de modelo diferentes: 6 versiones se basan en la TRF, seis versiones utilizan un banco de filtros adaptados perceptualmente y seis versiones utilizan una combinación de TRF y banco de filtros. El grupo TRF está previsto para ser utilizado en un esquema de medición que puede realizarse en tiempo real, mientras que los otros dos grupos requieren una mayor complejidad computacional y se espera que logren una mayor exactitud. Las seis versiones de modelo de cada grupo pueden dividirse en dos subgrupos: uno que utiliza redes neurales y otro que **no** utiliza redes neurales. Se incluye también, con fines de comparación, el comportamiento del modelo de referencia de abril de 1996 (B3) sobre la base de datos 3.

### 5.1 Criterios de preselección basados en correlación

– El comportamiento de las 18 versiones diferentes de modelo y del modelo de referencia B3 fue evaluado mediante el empleo de dos conjuntos de datos que, es su mayoría, no fueron utilizados para el acondicionamiento de los modelos (fase 1, fase 3 segunda parte y CRC no se han utilizado para entrenamiento).

– La base de datos 3 se creó especialmente para la evaluación de modelos perceptuales. Aproximadamente la mitad de esta base de datos se utilizó para acondicionar los modelos. La correlación entre los resultados subjetivos y los resultados objetivos en los elementos restantes (DB3_2nd) se utilizó para la evaluación de los modelos objetivos.

– La base de datos CRC fue creada por la CRC para la evaluación de los esquemas de codificación de audio perceptuales. Esta base de datos no se utilizó para acondicionar los modelos perceptuales. La correlación entre los resultados subjetivos y los resultados objetivos para todos los elementos se utilizó para la evaluación de los modelos objetivos.

Los resultados de la base de datos 3 se presentan en los Cuadros 24 y 25 (correlaciones y AES).

En la fase de preselección, todas las entradas que cuantifican el comportamiento de todas las versiones de modelo se tuvieron en cuenta con un énfasis en DB3. Sobre la base de la comparación global se decidió efectuar una ulterior comparación de las dos versiones de modelo, de cada grupo, que parecían ser las mejores. El Cuadro 24 muestra la correlación de estas seis versiones (3 veces 2 versiones de modelo).

CUADRO 24

**Correlación entre SDG y ODG**

| | **FFTNnODG1** | **FFTNnODG2** | **FiltODG2** | **FiltODG3** | **CombNnODG3** | **CombODG3** | **B3** |
|---|---|---|---|---|---|---|---|
| DB3_2nd | 0,671 | 0,728 | 0,738 | 0,751 | 0,828 | 0,826 | 0,710 |
| CRC | 0,837 | 0,779 | 0,862 | 0,839 | 0,851 | 0,777 | 0,656 |

CUADRO 25

**Nota de error absoluto**

| | **FFTNnODG1** | **FFTNnODG2** | **FiltODG2** | **FiltODG3** | **CombNnODG3** | **CombODG3** | **B3** |
|---|---|---|---|---|---|---|---|
| DB3_2nd | 2,96 | 2,79 | 3,16 | 3,16 | 2,91 | 2,56 | 2,39 |
| CRC | 1,55 | 1,85 | 1,61 | 1,67 | 1,61 | 1,90 | 2,78 |

Las dos versiones de modelo basadas en TRF muestran comportamientos diferentes para las dos bases de datos. Para decidir cuál de las dos habrá de seleccionarse será necesario establecer ulteriores criterios. Teniendo en cuenta que la base de datos CRC era totalmente desconocida, el comportamiento de la versión FFTNnODG1 parece ser algo mejor.

Las versiones de modelo combinadas dan mejores resultados para ambas bases de datos en comparación con las versiones basadas en TRF. Sin embargo, para la base de datos CRC, presentan un comportamiento peor en comparación con las versiones basadas en bancos de filtros. Estas últimas versiones pueden considerarse como un caso especial de las versiones combinadas en el que la ponderación de las variables de salida de las versiones basadas en TRF es cero. Por tanto, se prefieren las versiones de modelo combinadas.

## 5.2 Análisis del número de valores lejanos

La exactitud de los datos subjetivos obtenidos en pruebas de escucha se expresa mediante el intervalo del 95% en torno al promedio de varios oyentes. El comportamiento de las versiones de modelo se evaluó también atendiendo al número de valores lejanos. Por definición, se considera que la diferencia entre un dato subjetivo y un dato objetivo es un «valor lejano» cuando es mayor que el doble del intervalo de confianza.

El Cuadro 26 muestra el número de valores lejanos para las seis versiones de modelo, para DB3. Sensible significa que una versión de modelo indica una calidad de audio más baja que la obtenida por la evaluación subjetiva, e insensible significa lo contrario.

CUADRO 26

**Valores lejanos**

| | **FFTNnODG1** | **FFTNnODG2** | **FiltODG2** | **FiltODG3** | **CombNnODG3** | **CombODG3** |
|---|---|---|---|---|---|---|
| Sensible | 10 | 4 | 4 | 4 | 3 | 5 |
| Insensible | 13 | 13 | 11 | 13 | 12 | 14 |
| Total | 23 | 17 | 15 | 17 | 15 | 19 |

Atendiendo al número de valores lejanos, la versión de modelo FFTNnODG2 parece comportarse mejor que la versión de modelo FFTNnODG1. La versión de modelo CombNnODG3 muestra el mejor comportamiento entre las versiones más exactas.

## 5.3 Análisis de la severidad de los valores lejanos

CUADRO 27

**Elementos con desviaciones de más de 1,0 grados de diferencia entre predicción y SDG**

| **Número de valores lejanos** | **14** | **4** | **4** | **12** | **12** | **9** | **6** |
|---|---|---|---|---|---|---|---|
| Versión de modelo | B3 | CombODG3 | CombNnODG3 | FFTNnODG1 | FFTNnODG2 | FiltODG2 | FiltODG3 |
| Elemento | jcodsmg | jcodsmg | jcodsmg | jcodsmg | jcodsmg | jcodsmg | jcodsmg |
| | qcodket | qcodket | pcodsmg | qcodket | qcodket | rcodpip | rcodpip |
| | pcodsmg | pcodsmg | ccodclv | pcodsmg | pcodsmg | qcodket | qcodket |
| | pcodcla | icodmar | acodclv | ocodsfe | pcodcla | pcodsmg | pcodsmg |
| | ocodclv | | | ncodcas | ocodsfe | ocodsax | ocodsax |
| | ncodcas | | | mcodcla | ncodglo | kcodpip | fcodsb2 |
| | mcodpip | | | jcodveg | kcodcas | jcodveg | |
| | jcodryc | | | jcodryc | jcodveg | fcodsb2 | |
| | icodtam | | | icodtam | jcodryc | fcodcl2 | |
| | icodsmg | | | icodmar | icodtam | | |
| | fcodsb2 | | | fcodcl2 | ecodhrp | | |
| | ecodhrp | | | ecodhrp | ccodclv | | |
| | ccodclv | | | | | | |
| | acodclv | | | | | | |

CUADRO 28

**Elementos con desviaciones de más de 1,5 grados de diferencia entre predicción y SDG**

| **Número de valores lejanos** | **8** | **2** | **2** | **3** | **2** | **2** | **2** |
|---|---|---|---|---|---|---|---|
| Versión de modelo | B3 | CombODG3 | CombNnODG3 | FFTNnODG1 | FFTNnODG2 | FiltODG2 | FiltODG3 |
| Elemento | | jcodsmg | jcodsmg | jcodsmg | jcodsmg | jcodsmg | jcodsmg |
| | qcodket | | pcodsmg | | | | |
| | pcodsmg | pcodsmg | | pcodsmg | pcodsmg | | |
| | | | | | | pcodsmg | pcodsmg |
| | | | | ncodcas | | | |
| | ncodcas | | | | | | |
| | jcodryc | | | | | | |
| | icodtam | | | | | | |
| | icodsmg | | | | | | |
| | ccodclv | | | | | | |
| | acodclv | | | | | | |

## 6 Conclusión

De acuerdo con las decisiones tomadas durante el desarrollo, de las 18 versiones de modelo propuestas se han seleccionado dos para satisfacer los requisitos de las aplicaciones definidas de métodos de medición objetiva. Se definieron dos versiones, una de poca complejidad, concebida para tener en cuenta realizaciones en tiempo real eficientes con respecto al coste, y de mayor exactitud que no funciona necesariamente en tiempo real. En el proceso de selección se aplicaron y analizaron los criterios antes enumerados.

Como versión para funcionamiento en tiempo real se ha seleccionado un modelo basado en TRF, designado por «FFTNnODG1»; las razones para ello fueron las siguientes:

Ninguna de las versiones basadas en TRF mostraron una ventaja apreciable sobre las demás en cuanto a uno de los criterios antes indicados. En lo que respecta a los coeficientes de correlación entre ODG y SDG, el número y la severidad de los valores lejanos de cada una de las versiones verificadas tienen sus ventajas e inconvenientes. Sin embargo, debe señalarse que «FFTNnODG1» logró la mejor correlación (0,837) para la base de datos CRC'97, que era completamente desconocida.

En cuanto a la versión de mayor exactitud, se prefirió la versión combinada, que comprende una función TRF y un banco de filtros, ya que este planteamiento comprende también un subconjunto de un modelo basado en banco de filtros, puro, por lo que debe ofrecer mejores posibilidades de funcionamiento. En total, para el proceso de selección estaban disponibles seis versiones diferentes de un modelo combinado. La versión seleccionada «CombNnODG3», presentó menos valores lejanos, así como una mayor correlación (r=0,851 para CRC'97) que las otras versiones. La correlación de esta versión para base de datos 3 completa tenía el mismo orden de magnitud que la de la versión «CombNnODG3», pero presentó una mayor correlación en comparación con las otras versiones.

# Adjunto 2
# al Anexo 2

# Descripciones de las bases de datos de referencia

## 1 Introducción

Durante el desarrollo del método para la medición objetiva de la calidad de audio percibida se utilizaron varias bases de datos para el entrenamiento y la validación.

Algunas de esas bases de datos contenían datos para ser oídos con casco audífono y con altavoz, y otras contenían sólo datos para ser oídos con auriculares. En el caso de bases de datos que contenían conjuntos de datos separados para altavoz y auriculares, sólo se utilizaron los datos para auriculares.

Un elemento (sinónimo: un ítem) se define como un fragmento de audio utilizado en la evaluación subjetiva. Por una condición ha de entenderse una condición individual de degradación. Todos los elementos se utilizaron para todas las condiciones experimentales, salvo en los estudios en que se utilizaron las bases de datos DB2 y DB3. DB3 se utilizó en parte para entrenamiento y en parte para validación (52 de los 84 elementos fueron utilizados para entrenamiento en la segunda fase de la validación).

**Entrenamiento**

- MPEG90
  - El SDG medio por elemento cubría uniformemente la gama de 0,0 a −4,0.
  - Referencia: ISO/IEC JTC 1/SC 2/WG 11 MPEG/Audio test report, Document MPEG90/N0030, octubre 1990.
- MPEG91
  - Por lo menos el 88 por ciento del SDG medio por elemento estaba por encima de −2,0, y la gama era de 0,1 a −3,8.
  - Referencia: ISO/IEC JTC 1/SC 2/WG 11 MPEG/Audio test report, Document MPEG91/N0010, junio 1991.
- ITU92DI
  - 80 por ciento del SDG medio por elemento estaba por encima de −2,0, y la gama era de 0,1 a −3,4.
- ITU92CO
  - Por lo menos el 96 por ciento del SDG medio por elemento estaba por encima de −2,0, y la gama era de 0,2 a −2,4.
- ITU93
  - La mayor parte del SDG medio por elemento estaba por encima de −2,0, y la gama era de −0,1 a −2,3. No había una diferencia apreciable entre los datos de los dos laboratorios.
  - Referencias: Grusec y otros [1997].
- MPEG95
  - Por lo menos el 63 por ciento del SDG medio por elemento estaba por encima de −2,0, y la gama era de −0,2 a −3,8.
  - Referencia: Meares y Kim [1995].
- EIA95
  - Por lo menos el 93 por ciento del SDG medio por elemento estaba por encima de −2,0, y la gama era de 0,1 a −3,7.
  - Referencia: Grusec y otros [1997].
- DB2
  - No todos los elementos se utilizaron para todas las condiciones.

**Validación**

- DB3
  - No todos los elementos se utilizaron para todas las condiciones.
- CRC97
  - El SDG medio por elemento cubría uniformemente la gama de 0,1 a −3,6.
  - Referencia: Soulodre y otros [1998].

Los siguientes puntos describen los elementos que se incluyeron en las diferentes bases de datos y las condiciones que se aplicaron.

## 2 Elementos (ítems) por cada base de datos

| Elem. | MPEG90 | MPEG91 | ITU92DI | ITU92CO | ITU93 | MPEG95 | EIA95 | DB2 | DB3 | CRC97 |
|---|---|---|---|---|---|---|---|---|---|---|
| Acordeón/triángulo | | * | | | | | | | | |
| Asa Jinder | | | * | * | * | | | | | |
| Gaita | | | | | | * | | * | * | |
| Gaita 2 | | | | | | | | * | | |
| Clarinete bajo | | | | | | | | * | | |
| Bajo | * | | * | * | | | | | | |
| Sint. bajo | * | | | | | | | | | |
| Carmen | | * | | | | | | | | |
| Castañuelas | * | | * | * | * | * | | * | * | |
| Clarinete | | | | | * | | * | * | * | * |
| Clarinete mno | | | | | | | | * | | |
| Clarinete 2 | | | | | | | | * | | |
| Claves | | | | | | | | | * | |
| Dalarnas Spelmansförbund "Trettondagsmarschen" | | | * | * | | | | | | |
| Dire Straits "Ride Across the River" | | | * | | | | * | | | * |
| Double Bass | | | | | | | | | | * |
| Tambor | | | | | | | | * | | |
| Fireworks | * | | | | | | | | | |
| Flauta | | | | | | | | | * | |
| George Duke | | * | | | | | | | | |
| Carrillón | * | * | | | | * | * | * | * | |
| Clavicordio | | | * | * | * | * | * | * | * | * |
| Cuemo | | | | | | | | * | | |
| Kettle drums | | | | | | | | | * | |
| Marimba | | | | | | | | * | * | |
| MPE mono = Discurso hombre inglés mono | | | | | | | | * | | |
| Música y lluvia | | | | | | | * | | | * |
| Trompeta sordina | | | | | | | * | | | |
| Ornette Coleman | * | * | * | * | | | | | | |
| Pearl Jam | | | | | | | * | | | |
| Percusión | | * | | | | | | * | | |
| Piano Schubert | | | | | | | | | * | |
| Pitch Pipe | | | | | | * | | | * | * |
| Ravel "Feria" | | | * | | | | | | | |
| Ry Cooder | | | | | | | | * | * | |
| Ry Cooder (mono) | | | | | | | | | | |
| Saxofón | | | | | | | | | * | |
| Tambor mil. | | | | | | | | | * | |
| Soprano Mozart | | | | | | | | | * | |
| Discurso mujer inglesa | | | | | | * | | | * | |
| Discurso mujer alemana | | | | | | | | | * | |
| Discurso hombre ingés | * | * | | | | | | | * | |
| Discurso hombre alem. | | | * | * | * | | | | * | |
| Stravinsky "Wind Octet" | | | * | * | | | | | | |
| Cuerdas | | | | | | | | * | * | |
| Cuerdas mono | | | | | | | | * | | |
| Suzanne Vega "Toms Diner" | * | * | * | * | | | | * | * | * |
| Suzanne Vega with breaking glass | | | | | | | * | | | |
| Pandereta | | * | | | | | | * | * | |
| Tracy Chapman | * | | | | | | | | | |
| Triángulo | | | * | * | | | | * | * | |
| Trompeta | | | | | | | | * | * | * |
| Trompeta (Haydn) | * | | | | | | | | | |
| Tuba | | | | | | | | * | * | |
| De sålde sina hemman (violin solo) | | | | | * | | | | | |
| Sonido agua | | | | | | | * | | | |
| Conjunto viento | | | | | | | | * | | |
| Xilófono | | | | | | | | | * | |

## 3 Condiciones experimentales

Para todas las velocidades binarias con la indicación kbit/s estéreo se da la velocidad binaria total, por ejemplo 256 kbit/s estéreo significa que la velocidad 256 kbit/s está atribuida en total para ambos canales de una señal estereofónica. Si no se indica nada más, estéreo hace referencia a una codificación de canal independiente.

### 3.1 MPEG90

Tres velocidades binarias: 64 kbit/s mono, 192 kbit/s y 256 kbit/s estéreo, no todo el material estaba disponible para esta base de datos.

– Musicam.
– SB-ADPCM.

### 3.2 MPEG91

Tres velocidades binarias: 64 kbit/s mono, 192 kbit/s y 256 kbit/s estéreo.

– MPEG1 Capa I.
– MPEG1 Capa II.
– MPEG1 Capa III.
– MUSICAM.
– ASPEC.
– NICAM.

### 3.3 ITU92DI

Cinco códecs de distribución: 240 kbit/s estéreo.

Cada elemento fue procesado por el mismo códec tres veces en serie, con una caída de 0,1 dB del nivel antes de cada paso.

– MPEG1 Capa II.
– MPEG1 Capa III.
– Dolby AC-2.
– Aware.
– NHK.

### 3.4 ITU92CO

Seis códecs de contribución: 360 kbit/s estéreo. Cada elemento fue procesado por el mismo códec tres veces en serie con una caída de 0,1 dB del nivel antes de cada pasaje.

– MPEG1 Capa II.
– MPEG1 Capa III.
– Dolby AC-2.
– Dolby Low-Delay.
– Aware.

### 3.5 ITU93

Configuraciones de códecs en serie capa MPEG1 Capa II:

– Códec emisión sólo, a 256 kbit/s estéreo.
– Códec emisión sólo, a 192 kbit/s estéreo (codificación estéreo conjunta).
– Ocho códecs contribución a 360 kbit/s seguidos de un códec emisión a 256 kbit/s, todos en estéreo.
– Ocho códecs contribución a 360 kbit/s seguidos de un códec emisión a 192 kbit/s, todos en estéreo.

– Cinco códecs contribución a 360 kbit/s seguidos de tres códecs distribución a 240 kbit/s y un códec emisión a 256 kbit/s, todos en estéreo.
– Cinco códecs contribución a 360 kbit/s seguidos de tres códecs distribución a 240 kbit/s y un códec emisión a 192 kbit/s, todos en estéreo.

### 3.6 MPEG95

Realizaciones de códec (64 kbit/s):

– Veintidós variantes de codificación fueron seleccionadas de un conjunto mayor de métodos de codificación disponibles, de 6 códecs que formaban un subconjunto de 4 modelos de tiempo/frecuencia de baja resolución y 17 modelos de tiempo/frecuencia de alta resolución.
– Las organizaciones participantes fueron AT&T, Fraunhofer, Sony, GCL, RAI/Alcatel, y Philips.
– Todos los elementos fueron grabaciones monoaurales presentadas binauralmente.

### 3.7 EIA95

| | | |
|---|---|---|
| – | Eureka 147/MPEG1 Capa II #1 | 224 kbit/s estéreo (codificación estéreo conjunta) |
| – | Eureka 147/MPEG1 Capa II #2 | 192 kbit/s estéreo (codificación estéreo conjunta) |
| – | AT&T/Lucent | 160 kbit/s estéreo |
| – | AT&T/Lucent/Amati #1 | 128 kbit/s estéreo |
| – | AT&T/Lucent/Amati #2 | 160 kbit/s estéreo |
| – | VOA/JPL | 160 kbit/s estéreo |
| – | USADR-FM #1 | 128-256 kbit/s estéreo (velocidad binaria variable) |
| – | USADR-FM #2 | 128-256 kbit/s estéreo (velocidad binaria variable) |
| – | USADR-AM | 96 kbit/s estéreo |

### 3.8 DB2

| | | |
|---|---|---|
| – | MPEG1 Capa II | 256 kbit/s estéreo, 1, 3, 5, 7, y 9 etapas. |
| – | Dolby AC2 | 256 kbit/s estéreo, 1, 3, 5, 7, y 9 etapas. |
| – | MPEG1 Capa II | 192 kbit/s estéreo (codificación estéreo conjunta). |
| – | MPEG1 Capa II | 64 kbit/s mono. |
| – | MPEG2 Capa II | 64 kbit/s mono. |
| – | MPEG1 Capa II | 384 kbit/s estéreo. |
| – | MPEG1 Capa III | 128, 160, 192 kbit/s, todas estéreo. |
| – | APT-X | 256 y 384 kbit/s ambas estéreo. |
| – | Distorsión de cuantificación | |
| – | Grabación analógica 1, 2, 3 etapas | |
| – | Recorte | |

### 3.9 DB3

| | | |
|---|---|---|
| – | NICAM | |
| – | MiniDisc y MiniDisc + Capa II | 192 kbit/s, estéreo (codificación estéreo conjunta) |

| | | |
|---|---|---|
| – | Dolby AC2 | 256 kbit/s estéreo, 1, 3, 5, 7, y 9 etapas |
| – | MPEG1 Capa II | selección de base de datos de Swisscom, >192 kbit/s estéreo |
| – | MPEG1 Capa III | 128 y 160 kbit/s ambas estéreo (codificación estéreo conjunta) |
| – | MPEG AAC | 128 kbit/s estéreo (codificación estéreo conjunta) |
| – | MPEG Capa III | 128 + Capa II, 384 + Capa II, 224 kbit/s, todas estéreo |
| – | Dolby AC3 | 256 kbit/s estéreo |
| – | Dolby AC3 | 256 + MPEG Capa II, 224 kbit/s, ambas estéreo |
| – | Distorsión de cuantificación | |
| – | THD | |
| – | Ruido | |

### 3.10 CRC97

| | | |
|---|---|---|
| – | AT&T PAC | 64, 96, 128, y 160 kbit/s, todas estéreo |
| – | Dolby AC3 | 128, 160, y 192 kbit/s, todas estéreo |
| – | MPEG1 Capa II *software* | 128, 160, y 192 kbit/s, todas estéreo |
| – | MPEG1 Capa II *hardware* (ITIS) | 96, 128, 160, 192 kbit/s, todas estéreo |
| – | MPEG4 AAC | 96 y 128 kbit/s, ambas estéreo |
| – | MPEG1 Capa III | 128 kbit/s estéreo |

## 4 Elementos por cada condición para DB2 y DB3

### 4.1 DB2

| | | Condición Nº | Elementos |
|---|---|---|---|
| **Lugar de prueba I, NHK Japón** | | | |
| Capa II, 256 kbit/s | 1 etapa | CO13 | CLA,RYC,SB1,STR |
| | 3 etapas | CO11 | CLA,RYC,SB1,STR |
| | 5 etapas | CO19 | CLA,RYC,SB1,STR |
| | 7 etapas | CO18 | CLA,RYC,SB1,STR |
| | 9 etapas | CO15 | CLA,RYC,SB1,STR |
| NBC (Dolby AC2) | 1 etapa | CO1A | CAS,RYC,STR,WIN |
| | 3 etapas | CO12 | CAS,RYC,STR,WIN |
| | 5 etapas | CO17 | CAS,RYC,STR,WIN |
| | 7 etapas | CO16 | CAS,RYC,STR,WIN |
| | 9 etapas | CO14 | CAS,RYC,STR,WIN |

| | | Condición Nº | Elementos |
|---|---|---|---|
| **Lugar de prueba II, DR Dinamarca** | | | |
| Capa II, 256 kbit/s | 1 etapa | CO2B | CLA,RYC,SB1,STR |
| Capa II, 192 kbit/s js | | CO25 | CLA,RYC,SB1,STR |
| Capa II, 64 kbit/s mono | | CO27 | MLA,MPE,MTR,MYC |
| NBC (Dolby AC2) | 5 etapas | CO29 | CAS,RYC,STR,WIN |
| MPEG2/L2 LSF | | CO22 | MLA,MPE,MTR,MYC |
| Analógico 1 | | CO23 | PER |
| Analógico 2 | | CO2A | PER |
| Analógico 3 | | CO28 | PER |
| Errores 1 | | CO24 | GLO,HRN,TRI |
| Errores 2 | | CO21 | GLO,HRN,TRI |
| Recorte | | CO26 | BAS,CL2,TUB |
| **Lugar de prueba III, NRK Noruega** | | | |
| Capa II, 384 kbit/s | | CO34 | CLA,RYC,SB1,STR |
| Capa II, 256 kbit/s | 1 etapa | CO31 | CLA,RYC,SB1,STR |
| NBC (Dolby AC2) | 5 etapas | CO3B | CAS,RYC,STR,WIN |
| Capa III (ASPEC3), 192 kbit/s | | CO32 | CLA,STR,TAM,VEG |
| Capa III (ASPEC3), 128 kbit/s | | CO39 | CLA,STR,TAM,VEG |
| Capa III (ASPEC3), 160 kbit/s | | CO3A | CLA,STR,TAM,VEG |
| APT-X, 256 kbit/s | | CO33 | HAR,SB2,STR,TPT |
| APT-X, 384 kbit/s | | CO36 | HAR,SB2,STR,TPT |
| Distorsión de cuantificación 1 | | CO35 | DRU |
| Distorsión de cuantificación 2 | | CO37 | DRU |
| Distorsión de cuantificación 3 | | CO38 | DRU |

**Elementos de prueba (*test items*)**

| | |
|---|---|
| STR | Música folklórica sueca, grabación SR, previamente utilizado |
| SB1 | Gaitas, grabación SR |
| SB2 | Gaitas, grabación SR |
| CLA | Clarinete, SQUAM 16/2 |
| TAM | Tamboril, grabación SR, previamente utilizado |
| WIN | Stravinskij, Conjunto de viento, previamente utilizado |
| TPT | Trompeta, SQUAM 21/2 |
| HAR | Harlequin ensemble, grabación BBC G 49/17 |
| VEG | Suzanne Vega, grabación antigua, previamente utilizado |
| CAS | Castañuelas, SQUAM 27 |
| SPE | Habla alemana, SQUAM 54 |
| RYC | Ry Cooder, CD: JAZZ tr 11 (0.25 – 0.47) |
| PER | Percusión, Japanese Bass Marimba, CD: Sony/CBS 32DC 5027 |
| HRN | Cuerno, SQUAM 23/2 |

GLO Glockenspiel, SQUAM 35/1, previamente utilizado

TRI Triángulo, SQUAM 32/2

DRU Tambores, SQUAM 28

CL2 Clarinete, SQUAM 16/2

BAS Clarinete Bajo, SQUAM 17

TUB Tuba, SQUAM 24

MPE Mezcla monofónica de SPE

MTR Mezcla monofónica de STR

MLA Mezcla monofónica de CLA

MYC Mezcla monofónica de RYC

| **Test/elemento** | Clarinete | Clarinete mono | Ry Cooder | Ry Cooder mono | Gaita 1 | Cuerda | Cuerda mono | Castañuelas | Conjunto viento | MPE mono | Pandereta | Suzanne Vega | Clavicordio | Gaita 2 | Trompeta | Tambo | Percusión | Carrillón | Cuern | Triángulo | BAS | Clarinete 2 | Tuba |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MPEG1 Capa 2, 256 kbit/s, 1 etapa | X | | X | | X | X | | | | | | | | | | | | | | | | | |
| MPEG1 Capa 2, 256 kbit/s, 3 etapas | X | | X | | X | X | | | | | | | | | | | | | | | | | |
| MPEG1 Capa 2, 256 kbit/s, 5 etapas | X | | X | | X | X | | | | | | | | | | | | | | | | | |
| MPEG1 Capa 2, 256 kbit/s, 7 etapas | X | | X | | X | X | | | | | | | | | | | | | | | | | |
| MPEG1 Capa 2, 256 kbit/s, 9 etapas | X | | X | | X | X | | | | | | | | | | | | | | | | | |
| Dolby AC2, 256 kbit/s, 1 etapa | | | X | | | X | | X | X | | | | | | | | | | | | | | |
| Dolby AC2, 256 kbit/s, 3 etapas | | | X | | | X | | X | X | | | | | | | | | | | | | | |
| Dolby AC2, 256 kbit/s, 5 etapas | | | X | | | X | | X | X | | | | | | | | | | | | | | |
| Dolby AC2, 256 kbit/s, 7 etapas | | | X | | | X | | X | X | | | | | | | | | | | | | | |
| Dolby AC2, 256 kbit/s, 9 etapas | | | X | | | X | | X | X | | | | | | | | | | | | | | |
| MPEG1 Capa 2, 192 kbit/s est. común | X | | X | | X | X | | | | | | | | | | | | | | | | | |
| MPEG1 Capa 2, 64 kbit/s mono | | X | | X | | | X | | | X | | | | | | | | | | | | | |
| MPEG2 Capa 2, 64 kbit/s mono | | X | | X | | | X | | | X | | | | | | | | | | | | | |
| MPEG1 Capa 2, 384 kbit/s | X | | X | | X | X | | | | | | | | | | | | | | | | | |
| MPEG1 Capa 3 (ASPEC 3), 192 kbit/s | X | | | | | X | | | | | X | X | | | | | | | | | | | |
| MPEG1 Capa 3 (ASPEC 3), 128 kbit/s | X | | | | | X | | | | | X | X | | | | | | | | | | | |
| MPEG1 Capa 3 (ASPEC 3), 160 kbit/s | X | | | | | X | | | | | X | X | | | | | | | | | | | |
| APT-X, 256 kbit/s | | | | | | X | | | | | | | X | X | X | | | | | | | | |
| APT-X, 384 kbit/s | | | | | | X | | | | | | | X | X | X | | | | | | | | |
| Distorsión cuantificación 1 | | | | | | | | | | | | | | | | X | | | | | | | |
| Distorsión cuantificación 2 | | | | | | | | | | | | | | | | X | | | | | | | |
| Distorsión cuantificación 3 | | | | | | | | | | | | | | | | X | | | | | | | |
| Grabación analógica 1 etapa | | | | | | | | | | | | | | | | | X | | | | | | |
| Grabación analogical 2 etapas | | | | | | | | | | | | | | | | | X | | | | | | |
| Grabación analogical 3 etapas | | | | | | | | | | | | | | | | | X | | | | | | |
| Errores bit 1 | | | | | | | | | | | | | | | | | | X | X | X | | | |
| Errores bit 2 | | | | | | | | | | | | | | | | | | X | X | X | | | |
| Recorte | | | | | | | | | | | | | | | | | | | | | X | X | X |

## 4.2 DB3

| Elemento/test | Nomb. | 1 | 2MD | 2MDL2 | 3 1Step | 3 3Step | 3 5Step | 3 7Step | 3 9Step | 4 | 5 | 6 low | 6 high | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | |
| 13(1) Flauta | flu | | | | | | | X | X | | | | | | | | | | | |
| 16(2) Clarinete | cla | | | | X | | X | | | X | | X | | X | X | | | | | |
| 20(1) Saxofón | sax | | | | | | | X | X | | | | | | | | | | | |
| 21(2) Trompeta | tpt | | | | | X | X | | | | | | | | | | | | | |
| 24(2) Tuba | tub | | | | | | | | | | | | | | | | | X | | |
| 26(1) Claves | clv | | X | X | | | X | X | X | | | | | | | | | | | X |
| 27 Castañuelas | cas | | | | | | | | | X | | | X | | | X | X | | X | |
| 28 Tambor pequeño | sna | X | | | | | | | | | | | | | | | | | | X |
| 30 Timbales | ket | | | | | | | | | | | | | | | | | X | | |
| 32(1/2) Triángulo | tri | | X | X | X | X | | | | X | | | | | | | | | X | |
| 35(1/2) Carrillón | glo | | | | | | | | | | | X | X | | | | | | X | |
| 36(1) Xilófono | xyl | | | | | | | | | | | | | | | | | X | | |
| 40(1) Clavicordio | hrp | | | | X | X | | | | | X | | | X | | | | | | |
| 49 Voz de mujer inglesa | sfe | | | | | | | X | X | | | | | | | | | | X | |
| 54 Voz de hombre alemán | smg | X | X | X | X | X | X | | | | X | | | | | | | | | |
| 60 Piano Schubert | pia | | | | | | | | | | | | | | | | | X | | X |
| 61 Soprano Mozart | sop | | X | X | | | | | | | | | | | | | | | | |
| 53 Voz de mujer alemana | sfg | | | | | | | | | | | | | | | | | | | |
| 50 Voz de hombre inglés | sme | | | | | | | | | | | | X | | X | | | | | |
| Ref_tam | tam | X | | | | | | | | | | | | | | X | X | | | |
| Ref_str | str | | | | | | | | | | | | | | | X | X | | | |
| Ref_har | har | | | | | | | | | | | | | | | | | | | |
| Ry Cooder | ryc | | | | | | | | | | X | | | | | X | X | | | |
| Susanne Vega | veg | | | | | | | | | X | X | X | | X | X | | | | | |
| Diapason | pip | | | | | | | | | | | X | X | X | X | | | | | |
| Marimba | mar | X | | | | | | | | | | | | | | | | | | X |
| Gaita | sb1 | | | | | | | | | X | | | | | | | | | | |
| Nombre | | i | b | s | g | e | p | c | o | f | j | r | k | l | m | h | d | q | n | a |

1.) NICAM
2.) MD y MD + L2 (192 kbit/s)
3.) AC2 (256 kbit/s), 1, 3, 5, 7,9
4.) L2 (ST d-b, ³192 kbit/s)
5.) Capa 2 (256 kbit/s), 8 etapas
6.) Capa 3, 128 y 160 kbit/s
7.) AAC, 128 kbit/s
8.) L3 (128) + L2 (384) + L2 (224)
9.) AC3 (256)
10.) AC3 (256) + L2 (224)
11.) Distorsión cuantificación
12.) THD
13.) Ruido

**Ejemplos nombres:**

Referencia: irefflu
Test: icodflu

## Glosario

### Nota de error absoluto (AES)

La AES se calcula mediante una fórmula desarrollada especialmente para evaluar la calidad de los resultados obtenidos utilizando un método de medición objetiva perceptual. Tiene en cuenta los intervalos de confianza de los valores medios de las pruebas de escucha subjetivas.

### Calidad de audio básica

La calidad de audio básica se define como un atributo subjetivo global que incluye alguna, algunas o todas las diferencias detectadas entre la señal de referencia y una versión procesada de la misma.

### Margen de codificación

El margen de codificación es un parámetro de calidad que mide el margen que existe entre los artefactos de codificación inaudibles y el umbral en que comienzan a ser audibles.

### Variables de salida del modelo (MOV)

Las variables de salida del modelo son valores de salida intermedios del método de medición perceptual. Estas variables se basan en resultados sicoacústicos básicos, por lo que pueden utilizarse para una ulterior caracterización de artefactos de codificación.

### Grado de diferencia objetiva (ODG)

El grado de diferencia objetiva es el principal parámetro del método de medición perceptual. Corresponde al grado de diferencia subjetiva (SDG) y es el parámetro de medición que da la calidad de audio básica global. La gama del ODG está comprendida entre 0 y −4.

**Medición fuera de línea**

Procedimiento de medición en que no hay interacción con la transmisión de programa en curso.

**Medición en línea**

Procedimiento de medición que se basa en la transmisión de programa en curso, o en partes de la misma.

**Grado de diferencia subjetiva (SDG)**

En una prueba de audición de acuerdo con la Recomendación UIT-R BS.1116, la calidad de audio básica de la referencia oculta y la de la versión procesada de la referencia son juzgadas utilizando una escala de degradación de cinco grados. La diferencia de grado se define como la diferencia obtenida cuando al grado atribuido a la señal sometida a prueba se le resta el grado atribuido a la señal de referencia. En un orden ideal, la gama del SDG está comprendida entre 0 y −4. Si la referencia no ha sido debidamente identificada, el SDG es positivo.

**Abreviaturas**

| | |
|---|---|
| ADB | Bloque distorsionado medio (*average distorted block*) |
| AES | Nota de error absoluto (*absolute error score*) |
| ASD | Diferencia espectral auditiva (*auditory spectral difference*) |
| Avg | Promedio (lineal) (*average* (*linear*)) |
| BAQ | Calidad de audio básica (*basic audio quality*) |
| Bw | Anchura de banda (*bandwidth*) |
| CI | Intervalo de confianza (*confidence interval*) |
| CM | Margen de codificación (*coding margin*) |
| DBn | Base de datos *n* (1, 2 ó 3) (*database n (1, 2 or 3)*) |
| DC | Corriente continua (*direct current*) |
| DFT | Transformada discreta de Fourier (*discrete Fourier transform*) |
| DIX | Índice de perturbación (*disturbance index*) |
| DUT | Dispositivo sometido a prueba (*device under test*) |
| EHS | Estructura armónica del error (*error harmonic structure*) |
| ERB | Anchura de banda rectangular equivalente (*equivalent rectangular bandwidth*) |
| fac | Factor |
| FIR | Respuesta a impulso finito (*finite impulse response*) |
| IIR | Respuesta a impulso infinito (*infinite impulse response*) |
| ISO | Organización Internacional de Normalización |
| JNLD | Diferencia de nivel justamente perceptible (*just noticeable level difference*) |
| MFPD | Probabilidad filtrada máxima de detección (*maximum filtered probability of detection*) |
| MOV | Variable de salida del modelo (*model output variable*) |
| MPEG | Grupo de expertos en imágenes animadas (*Moving Picture Expert Group*) |
| NL | Sonoridad del ruido (*noise loudness*) |

| | |
|---|---|
| NMR | Relación Ruido-Enmascaramiento (*noise-to-mask ratio*) |
| OASE | Evaluación objetiva de señal de audio (*objective audio signal evaluation*) |
| OCM | Margen de codificación objetivo (*objective coding margin*) |
| ODG | Grado de diferencia subjetiva (*objective difference grade*) |
| PAQM | Medida de la calidad de audio perceptual (*perceptual audio quality measure*) |
| PEAQ | Mediciones objetivas de la calidad de audio percibida (*objective measurements of perceived audio quality*) |
| PERCEVAL | Evaluación objetiva perceptual (*perceptual evaluation*) |
| POM | Medida perceptual objetiva (*perceptual objective measure*) |
| REF | Señal de referencia (*reference signal*) |
| res | Resolución |
| RMS | Media cuadrática, sinónimo: valor medio cuadrático (*root mean squared*) |
| r.m.s. | Media cuadrática, sinónimo: valor medio cuadrático (*root mean squared*) |
| ROEX | Exponencial redondeado (*rounded exponential*) |
| ROV | Tasa de valores de salida (*rate of output values*) |
| SCM | Margen de codificación subjetivo (*subjective coding margin*) |
| SDG | Grado de diferencia subjetiva (*subjective difference grade*) |
| *S*/*N* | Relación señal/ruido |
| SPL | Nivel de presión sonora, sinónimo: nivel de presión acústica (*sound pressure level*) |
| SUT | Señal sometida a prueba (*signal under test*) |
| THD | Distorsión armónica total (*total harmonic distortion*) |
| TRF | Transformada rápida de Fourier (*FFT – fast Fourier transform*) |
| Win | Promedio ventanizado (*windowed average*) |

## Referencias bibliográficas

AURAS, W. [septiembre de 1984] *Berechnungsverfahren für den Wohlklang beliebiger Schallsignale, ein Beitrag zur gehörbezogenen Schallanalyse*. Dissertation an der Fakultät für Elektrotechnik der Technischen Universität München, República Federal de Alemania.

BEERENDS, J. G. y STEMERDINK, J. A. [diciembre de 1992] A perceptual audio quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.*, Vol. 40, p. 963-978.

BEERENDS, J. G. y STEMERDINK, J. A. [febrero de 1994] Modeling a cognitive aspect in the measurement of the quality of music codecs. Contribution to the 96th AES Convention, preprint 3800. Ámsterdam (Países Bajos).

BEERENDS, J. G. y STEMERDINK, J. A. [marzo de 1994] A perceptual speech quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.*, Vol. 42, p. 115-123.

BEERENDS, J. G., van den BRINK, W. A. C. y RODGER, B. [mayo de 1996] The role of informational masking and perceptual streaming in the measurement of music codec quality. Contribution to the 100th AES Convention, preprint 4176. Copenhague (Dinamarca).

BRANDENBURG, K. [1987] Evaluation of quality for audio encoding at low bit rates. Contribution to the 82nd AES Convention, preprint 2433. Londres (Reino Unido).

BREGMAN, A. S. [1990] *Auditory Scene Analysis: The Perceptual Organisation of Sound*, MIT Press, Cambridge MA, Estados Unidos de América.

COHEN, E. A. y FIELDER, L. D. [mayo de 1992] Determining noise criteria for recording environments. *J. Audio Eng. Soc.*, Vol. 40, p. 384-402.

COLOMES, C., LEVER, M., RAULT, J. B. y DEHERY, Y. F. [abril de 1995] A perceptual model applied to audio bit-rate reduction. *J. Audio Eng. Soc.*, Vol. 43, p. 233-240.

FEITEN, B. [marzo de 1997] Measuring the Coding Margin of Perceptual Codecs with the Difference Signal. 102nd AES-Convention, preprint 4417. Múnich (República Federal de Alemania).

GRUSEC, T., THIBAULT, L. y SOULODRE, G. [septiembre de 1997] EIA/NRSC DAR systems subjective tests. Part 1: Audio codec quality. *IEEE Trans. on Broadcasting*, Vol. 43, 3.

KARJALAINEN, J. [marzo de 1985] A new auditory model for the evaluation of sound quality of audio system. Proceedings of the ICASSP, p. 608-611. Tampa (Florida, Estados Unidos de América).

LEEK, M. R. y WATSON, C. S. [1984] Learning to detect auditory pattern components. *J. Acoust. Soc. Am.*, Vol. 76, p. 1037-1044.

MEARES, D. J. y KIM, S. W. [julio de 1995] «NBC time/frequency module subjective tests: overall results», ISO/IEC JTC 1/SC 29/WG 11 N0973 MPEG95/208.

MOORE, B. C. [1986] *Frequency Selectivity in Hearing.* Academic Press, Londres (Reino Unido).

MOORE, B. C. [1989] *An introduction to the psychology of hearing.* Academic Press, Londres (Reino Unido).

PAILLARD, B., MABILLEAU, P., MORISETTE, S. y SOUMAGNE, J. [1992] Perceval: Perceptual evaluation of the quality of audio signals. *J. Audio Eng. Soc.*, Vol. 40, p. 21-31.

SCHROEDER, M. R., ATAL, B. S. y HALL, J. L. [diciembre de 1979] Optimizing digital speech coders by exploiting masking properties of the human ear. J. Acoust. Soc. Am., Vol. 66, p. 1647-1652.

SOULODRE, G., GRUSEC, T., LAVOIE, M. y THIBAULT, L. [marzo de 1998] Subjective evaluation of state-of-the-art 2-channel audio codecs. J. Audio Eng. Society.

SPORER, T. [octubre de 1997] Objective audio signal evaluation – applied psychoacoustics for modeling the perceived quality of digital audio. 103rd AES-Convention, preprint 4512. Nueva York (Estados Unidos de América).

TERHARDT, E. [1979] Calculating Virtual Pitch, Hearing Research. Vol. 1, p. 155-182.

THIEDE, T. y KABOT, E. [1996] A New Perceptual Quality Measure for Bit Rate Reduced Audio. Contribution to the 100th AES Convention, preprint 4280. Copenhague (Dinamarca).

TREURNIET, W. C. [1996] Simulation of individual listeners with an auditory model. Proceedings of the Audio Engineering Society, Reprint Number 4154. Copenhague (Dinamarca).

von BISMARCK, G. [1974] Sharpness as an attribute of the timbre of steady sounds. *Acústica*, 30, p. 159-172.

ZWICKER, E. y FASTL, H. [1990] *Psycho-acoustics, Facts and Models*. Berlin; Heidelberg: Springer Verlag, República Federal de Alemania.

ZWICKER, E. y FELDTKELLER, R. [1967] *Das Ohr als Nachrichtenempfänger*. Stuttgart: Hirzel Verlag, República Federal de Alemania.

## Bibliografía

GRUSEC, T., THIBAULT, L. y SOULODRE, G. [1995] Subjective evaluation of high quality audio coding systems: methods and results in the two-channel case. Preprint 4065 (F-5), Proceedings of the AES. Nueva York (Estados Unidos de América).